

Generated Jacobian Equations in Freeform Optical Design

Citation for published version (APA):

Romijn, L. B. (2021). *Generated Jacobian Equations in Freeform Optical Design: Mathematical Theory and Numerics*. Eindhoven University of Technology.

Document status and date:

Published: 19/10/2021

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

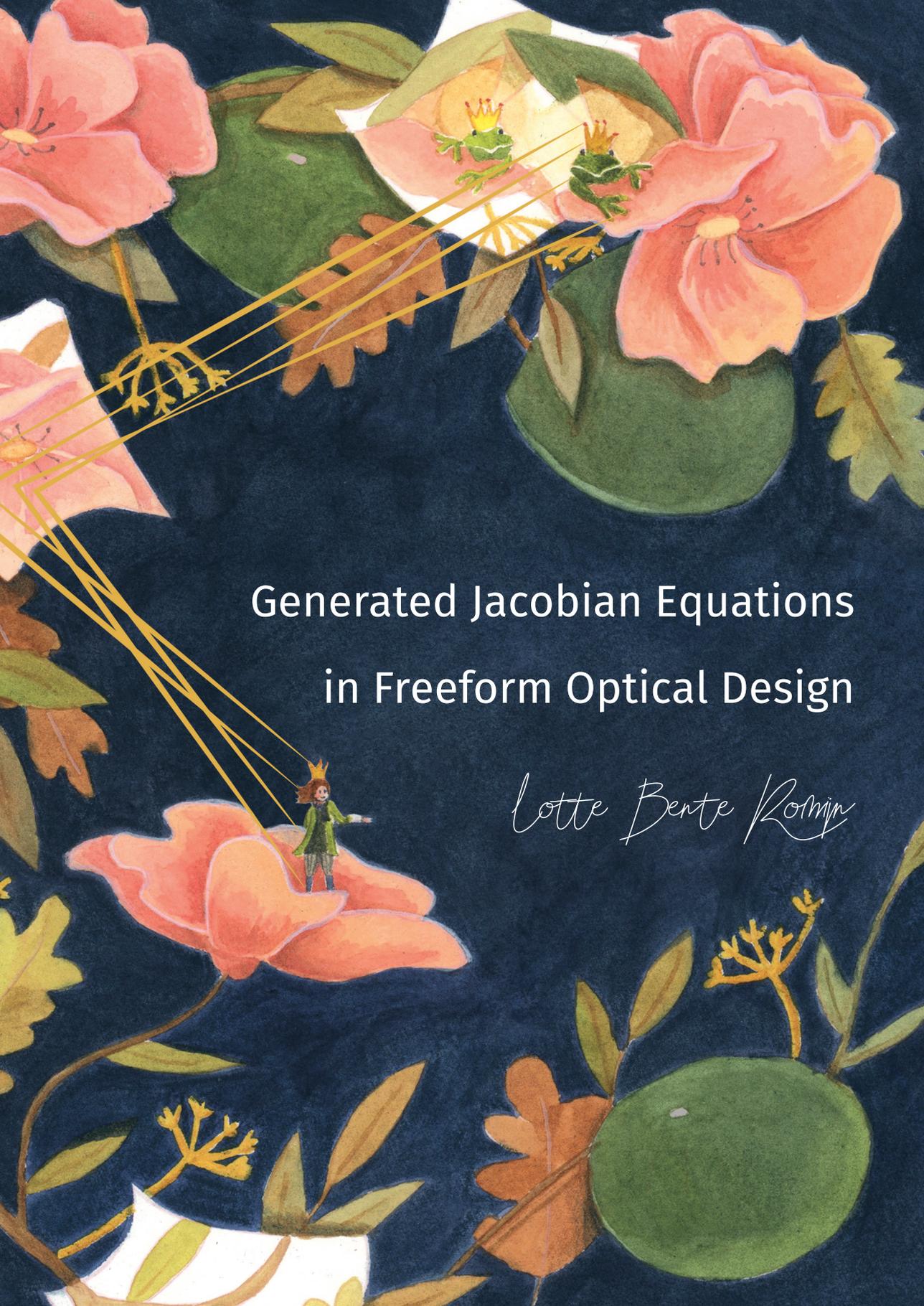
www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.



Generated Jacobian Equations
in Freeform Optical Design

Lotte Bente Romijn

Generated Jacobian Equations in Freeform Optical Design

Lotte Bente Romijn

Romijn, Lotte Bente

Generated Jacobian Equations in Freeform Optical Design

Eindhoven University of Technology, 2021

The research described in this thesis was performed at the Centre for Analysis, Scientific Computing and Applications (CASA) within the Department of Mathematics and Computer Science at Eindhoven University of Technology, the Netherlands, and at Signify at the High Tech Campus in Eindhoven, the Netherlands.

This work is part of the research program NWO-TTW Perspectief with project number P15-36, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO).

Program website: www.freeformscatteringoptics.com.

A catalogue record is available from the Eindhoven University of Technology Library.

ISBN: 978-94-6416-694-1

Cover design: © evelienjagtman.com.

Printing: Ridderprint | www.ridderprint.nl.

Copyright © 2021 by L. B. Romijn, The Netherlands.
All rights are reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of the author.

Generated Jacobian Equations in Freeform Optical Design: Mathematical Theory and Numerics

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof. dr. ir. F. P. T. Baaijens, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op dinsdag 19 oktober 2021 om 13:30 uur

door

Lotte Bente Romijn

geboren te Eindhoven

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof. dr. J. J. Lukkien
1 ^e promotor:	prof. dr. ir. W. L. IJzerman
2 ^e promotor:	dr. ir. J. H. M. ten Thije Boonkkamp
copromotor:	dr. ir. M. J. H. Anthonissen
leden:	prof. dr. ir. C. Vuik (TU Delft)
	prof. dr. V. I. Olikier (Emory University, USA)
	prof. dr. K. Veroy-Grepl
	prof. dr. I. D. Setija

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Foreword

I grew up in Eindhoven, the city of light also known as ‘Lampegat’. My grandfather worked on Philips CRT TVs, my sister and I took dancing classes in old Philips factories, and we rode the ‘Lichtjesroute’ every year. After studying in Amsterdam and Melbourne for six years, it was time for me to go home. With no sense of direction I arrived at Eindhoven University of Technology to see if I could apply my numerical mathematics background close to home. Only one month later I started my PhD in illumination optics, immediately caught by the enthusiasm of my new supervisors Jan and Wilbert, later joined by Martijn.

The subsequent years have brought me lots of experiences. I completed an optical-design course for optical engineers, joined the departmental PhD council as secretary and treasurer, worked at Signify once a week, enjoyed my numerical linear algebra teaching duties, but most of all, I learned to be more patient with myself. Most hours were spent finding a solution to a mathematical question, debugging code, and wondering if I was asking the right mathematical questions at all. After each stint of hard work, I personally loved the process of writing a research article. I also very much enjoyed writing this thesis.

My publications are given under the list of publications section at the end of this thesis. For one of those articles [142], my supervisor Jan asked me to compute a mirror or lens that transforms a parallel beam of light rays from a nonuniform light source onto a nonuniform target intensity on a screen. I chose a grayscale picture of a frog as my source light distribution and a grayscale picture of a prince as my target light distribution. In the remaining years of my PhD research I was mocked for this choice, especially since I found my own prince very shortly after publication. The computation for a mirror is included in this thesis. The cover of this thesis shows the frog, prince, a few mirrors and light rays, in a bed of flowers which resembles the floral printed clothes my colleague Vi and I like to wear.

Contents

List of Figures	xi
List of Tables	xv
Acronyms and Nomenclature	xvii
1 Introduction	1
1.1 Motivation	1
1.2 Main results of this thesis	7
1.3 Outline of this thesis	8
2 The Principles of Geometrical Optics	11
2.1 From electromagnetic theory to geometrical optics	12
2.2 Short wavelength approximation	16
2.3 The ray equation and Fermat's principle	19
2.4 Law of reflection	22
2.5 Snell's law	25
2.6 Hamiltonian optics	29
2.7 Hamilton's characteristic functions	35
2.7.1 The point characteristic	37
2.7.2 The mixed characteristic of the first kind	39
2.7.3 The mixed characteristic of the second kind	41
2.7.4 The angular characteristic	42
2.7.5 Interpretation of the characteristic functions	44
2.7.6 Hamilton's characteristics and parallel/point sources and targets	47
2.8 Radiometric and photometric units	49
2.9 Summary	51

3	Reflector and Lens Equations	53
3.1	Geometric conventions	58
3.1.1	Far-field approximation	59
3.1.2	Stereographic coordinates	60
3.1.3	Source and target distributions	63
3.1.4	Target on a projection screen in the far field	64
3.2	Parallel-to-far-field reflector	67
3.3	Parallel-to-near-field reflector	73
3.4	Point-to-far-field reflector	77
3.5	Point-to-far-field lens	83
3.6	Point-to-parallel reflector	87
3.7	Summary	92
4	Generated Jacobian Equations	99
4.1	Measure-preserving mappings	100
4.2	Convex analysis	102
4.3	C-convex analysis	106
4.3.1	An optimal-transport mapping	112
4.4	G-convex analysis	114
4.4.1	Generated Jacobian equations	120
4.4.2	A generated-Jacobian mapping	124
4.5	The transport boundary condition and edge-ray principle	127
4.6	Summary	131
5	A Literature Review on Numerical Methods	133
5.1	Direct Monge-Ampère solvers	134
5.1.1	Direct standard Monge-Ampère solvers	134
5.1.2	Direct solvers	135
5.1.3	Direct solvers for double freeform systems	136
5.2	Optimization strategies for the Monge-Kantorovich problem	137
5.2.1	Optimization strategies for single freeform systems	137
5.2.2	Optimization strategies for double freeform systems	139
5.3	Ray-mapping methods	139
5.3.1	Ray-mapping methods for single freeform systems	140
5.3.2	Ray-mapping methods for double freeform systems	140
5.4	Generated Jacobian equations	142
5.5	Summary	144

6	The Least-Squares Numerical Algorithm	145
6.1	The GLS algorithm	147
6.1.1	Minimization procedure for \mathbf{b}	151
6.1.2	Minimization procedure for \mathbf{P}	153
6.1.2.1	Case A: Regular minimizers	156
6.1.2.2	Case B: Regular minimizers with $a_4 = 0$	159
6.1.2.3	Case C: Regular minimizers with $a_4 = 0$ and $a_2 = 0$	160
6.1.2.4	Case D: Minimizers if $q_{11} = q_{22}$ and $\tilde{q}_{12} = 0$	161
6.1.2.5	Case E: Minimizers if $q_{11} = -q_{22}$	163
6.1.3	Minimization procedure for \mathbf{m}	164
6.1.4	Computation of u_1	166
6.2	Extension to polar source coordinates	168
6.2.1	Minimization procedure for \mathbf{m}	171
6.2.2	Computation of u_1	172
6.3	The GJLS algorithm	173
6.3.1	Minimization procedure for u	175
6.4	Summary	178
7	Numerical Results – Part I	179
7.1	Point-to-far-field reflector	180
7.1.1	Exact solution: tilted flat surface	180
7.1.2	Square-to-circle problem	183
7.2	Parallel-to-far-field reflector: frog to prince	188
7.3	Point-to-far-field lens	191
7.3.1	Peanut lens for road-lighting	191
7.3.2	An ellipsoidal lens comparison	193
7.3.3	Reduction in surface calculations	197
7.4	Parallel-to-near-field reflector	202
7.5	Summary	204
8	A Double Freeform Lens	205
8.1	Mathematical formulation	206
8.1.1	The first freeform surface	209
8.1.2	The second freeform surface	212
8.2	The GJLS algorithm for a double freeform lens	218
8.3	Summary	218

9 Numerical Results – Part II	221
9.1 Exact double freeform lens	221
9.2 Van Gogh double freeform lens	226
9.3 Summary	232
10 Conclusions and Recommendations	233
10.1 Summary and conclusions	233
10.2 Future research	234
A G-convex and G-concave Functions	239
A.1 G-convex and G-concave functions	239
A.1.1 G-convex and H-concave pair	240
A.1.2 G-concave and H-convex pair	242
B The Finite Volume Method	243
B.1 The finite volume method in Cartesian coordinates	243
B.1.1 Incorporating boundary conditions	247
B.2 The finite volume method in polar coordinates	251
B.2.1 Incorporating the outer boundary	255
B.2.2 Incorporating the inner boundary	256
B.3 Solving the Neumann problem for u	257
B.3.1 Incorporating boundary conditions	260
Bibliography	265
Index	278
Summary	281
List of Publications	283
Journal articles	283
Conference contributions	283
Oral presentations at scientific conferences	284
Other publications	285
Curriculum Vitae	287
Acknowledgments	289

List of Figures

Figure 1.1	An LED spotlight	2
Figure 1.2	Target light distributions for car headlights	3
Figure 1.3	Freeform scattering optics	6
Figure 2.1	An electromagnetic wave	17
Figure 2.2	A plane wave	19
Figure 2.3	Stationary optical path length	22
Figure 2.4	The law of reflection	23
Figure 2.5	The vectorial law of reflection	25
Figure 2.6	The law of refraction	26
Figure 2.7	The ray $x(z)$ with momentum $p^*(z)$ and the optical direction cosines	32
Figure 2.8	The ray $x(z)$ on the plane spanned by $x(z)$ and the z -axis	33
Figure 2.9	An alternative derivation of Snell's law	35
Figure 2.10	Hamilton's characteristic functions V, W, W^* and T	45
Figure 2.11	Hamilton's characteristic functions on the plane spanned by q_t and p_t^*	46
Figure 2.12	A virtual target on the source plane	48
Figure 3.1	The 8 base-case reflector systems	56
Figure 3.2	The 8 base-case lens systems	57
Figure 3.3	The far-field approximation for a lens	60
Figure 3.4	Stereographic projections from the unit sphere	62
Figure 3.5	A target screen P for a point-to-far-field reflector	67
Figure 3.6	Hamilton's mixed characteristic W for a parallel source, far-field target and a reflector	68
Figure 3.7	Hamilton's point characteristic V for a parallel source, near-field target and a reflector	74
Figure 3.8	Hamilton's angular characteristic T for a point source, far-field target and a reflector	77

Figure 3.9	Hamilton’s angular characteristic T for a point source, far-field target and a lens	84
Figure 3.10	Hamilton’s mixed characteristic W^* for a point source, parallel target and two reflectors	88
Figure 4.1	Cross-section of a strictly convex function $u(x)$, whose graph can be supported from below by the tangent functions	105
Figure 4.2	$u_1(x)$ can be supported from below by the functions G	108
Figure 4.3	The vectorial law of reflection is a bijection	130
Figure 4.4	A convex parabolic reflector surface and a strictly concave parabolic reflector	131
Figure 6.1	Categorization of all base-case optical systems	146
Figure 6.2	Flow chart of the GLS algorithm	150
Figure 6.3	Minimizing \mathbf{b} procedure using skew projections	152
Figure 6.4	Flow chart of the GJLS algorithm	176
Figure 7.1	“Tilded-flat-surface” problem: the mapping, reflector surface and convergence history	182
Figure 7.2	“Square-to-circle” problem: the convergence history for several grid sizes	185
Figure 7.3	“Square-to-circle” problem: the values of J_I and J_B as function of the iteration number for different values of N_b	186
Figure 7.4	“Square-to-circle” problem: the average calculation times	186
Figure 7.5	“Square-to-circle” problem: the values of $J = (1 - \alpha) J_I + \alpha J_B$ as function of the iteration number for several grid sizes and different values of α	187
Figure 7.6	“Frog-to-prince” problem: the source and target distributions	189
Figure 7.7	“Frog-to-prince” problem: the mapping, reflector surface and convergence history	190
Figure 7.8	“Frog-to-prince” problem: the ray-trace results	191
Figure 7.9	“Peanut-lens” problem: schematic representation	192
Figure 7.10	“Peanut-lens” problem: the target intensity	193
Figure 7.11	“Peanut-lens” problem: the source distribution, mapping, lens surface and convergence history	194
Figure 7.12	“Peanut-lens” problem: the ray-trace results	195
Figure 7.13	“Ellipsoidal-lens” problem: the mapping, lens surfaces and convergence histories to compute both a c-concave and G-convex solution	198

Figure 7.14	“Ellipsoidal-lens” problem: the maximum absolute errors of the final mappings and optical surfaces	199
Figure 7.15	“Ellipsoidal-lens” problem: the average calculation times	200
Figure 7.16	“Ellipsoidal-lens” problem: the effect of increasing T_u	201
Figure 7.17	“Jan” problem: the mapping, reflector surface, convergence history and ray-trace results	203
Figure 8.1	Double freeform lens with a point source and far-field target	205
Figure 8.2	Double freeform lens converting the intensity $f(\phi, \theta)$ of a point source into a far-field target intensity $g(\psi_2, \chi_2)$	207
Figure 8.3	Hamilton’s angular characteristic T for a double freeform lens	209
Figure 8.4	Freeform lens with k freeform surfaces for a point source and a far-field target	219
Figure 9.1	“Exact-lens” problem: the first mapping, first surface, maximum absolute error and convergence history	224
Figure 9.2	“Exact-lens” problem: the second mapping, second surface, maximum absolute error and convergence history	225
Figure 9.3	“Exact-lens” problem: the surfaces and absolute error of the second surface for $N = 100$	226
Figure 9.4	Regular interpolation vs. interpolation via a mapping	227
Figure 9.5	“Van-Gogh-lens” problem: the intermediate target intensities, mappings and optical surfaces	230
Figure 9.6	“Van-Gogh-lens” problem: ray-trace results	231
Figure B.1	Control volume for a cell-centered finite volume method	245
Figure B.2	Control volume for the left boundary of a cell-centered finite volume method	248
Figure B.3	Control volume for a cell-centered finite volume method on a polar grid	253
Figure B.4	Control volume for the outer boundary of a cell-centered finite volume method on a polar grid	255
Figure B.5	Control volume for the inner boundary of a cell-centered finite volume method on a polar grid	256

List of Tables

Table 3.1	Overview of the 16 base-case optical systems with generating functions and cost functions	94
Table 5.1	An overview of numerical methods for the 16 base cases . . .	143
Table 6.1	All possible minimizers for P	155
Table 7.1	“Square-to-circle” problem: number of iterations, total computation time and residuals in the GLS algorithm	184
Table 7.2	“Ellipsoidal-lens” problem: number of iterations, total computation time and residuals in the GLS and GJLS algorithms	197
Table 7.3	“Ellipsoidal-lens” problem: number of iterations, total computation time and residuals in the GJLS algorithm, varying T_u	201
Table 9.1	“Exact-lens” problem: number of iterations, total computation time and residuals in the GJLS algorithm	223
Table 9.2	“Van-Gogh-lens” problem: number of iterations, total computation time and residuals in the GJLS algorithm	231

Acronyms and Nomenclature

Acronyms

GJLS	Generated Jacobian least-squares
GLS	Generalized least-squares
LED	Light-emitting diode
ODE	Ordinary differential equation
PDE	Partial differential equation
SMS	Simultaneous multiple surfaces
SND	Symmetric negative definite
SPD	Symmetric positive definite
TIR	Total internal reflection

Nomenclature

\mathbf{b}	Function from the source boundary to the target boundary $\mathbf{b} : \partial\mathcal{X} \rightarrow \partial\mathcal{Y}$
$\mathbf{B}(x, t)$	Magnetic induction in T (tesla)
$c(x, \mathbf{y})$	Cost function in optimal transport theory in Cartesian and/or stereographic coordinates
$\mathbf{C} = D_{xy}c$	Cost function matrix of the mixed second-order partial derivatives of $c(x, \mathbf{y})$
$\mathbf{C} = D_{xy}\tilde{H}$	Mixed Hessian matrix of the mixed second-order partial derivatives of $\tilde{H}(x, \mathbf{y})$
$\mathbf{C}(\omega, \mathbf{y})$	Cost function matrix with polar source coordinates
$d(P, Q)$	Distance between point P and point Q

$D(\mathbf{x}, t)$	Displacement field in C/m ²
ϵ, ϵ_0	Permittivity of a dielectric, of vacuum in F/m (farad/meter)
$E(\mathbf{x}, t)$	Electric field vector in V/m
$f(\phi, \theta)$	Source intensity in spherical coordinates in lm/sr
$f(\mathbf{x})$	Source intensity in Cartesian coordinates in lm/m ²
$\tilde{f}(\mathbf{x})$	Source intensity in stereographic coordinates in lm/m ²
$F(\mathbf{x}, \mathbf{m}(\mathbf{x}))$ $F(\mathbf{x}, \mathbf{m}(\mathbf{x}), u(\mathbf{x}))$	Right-hand side of the generalized Monge-Ampère equation or generated Jacobian equation (Chapter 6 – 9)
$g(\psi, \chi)$	Target intensity in spherical coordinates in lm/sr
$g(\psi_2, \chi_2)$	Target intensity in spherical coordinates for a double freeform system in lm/sr
$\tilde{g}(\mathbf{y}_2)$	Target intensity in stereographic coordinates for a double freeform system in lm/m ²
$g(\mathbf{y})$	Target intensity in Cartesian coordinates in lm/m ²
$\tilde{g}(\mathbf{y})$	Target intensity in stereographic coordinates in lm/m ²
$G(\mathbf{x}, \mathbf{y}, w)$	Generating function
$h(\psi_1, \chi_1)$	Intermediate target intensity in spherical coordinates in lm/sr for a double freeform system
$\tilde{h}(\mathbf{y}_1)$	Intermediate target intensity in stereographic coordinates in lm/m ² for a double freeform system
H	Hamiltonian (Chapter 2) / Inverse of generating function G (Chapter 3 – 10)
$H(\mathbf{x}, \mathbf{y}, w)$	Inverse of generating function G
$\tilde{H}(\mathbf{x}, \mathbf{y})$	Inverse of generating function G
$\mathbf{H}(\mathbf{x}, t)$	Magnetic field in A/m
$\hat{\mathbf{i}}$	Direction of the intermediate ray
μ, μ_0	Permeability of a dielectric, of vacuum in H/m (henry/meter)
$\mathbf{m}(\mathbf{x}), \mathbf{m}(\omega)$	Mapping from source coordinate $\mathbf{x} \in \mathcal{X}$, or polar coordinate $\omega \in \mathcal{X}$ to target coordinate $\mathbf{y} \in \mathcal{Y}$
n	Refractive index
n_s	Refractive index at the source plane $z = z_s$

n_t	Refractive index at the target plane $z = z_t$
O	Origin of the coordinate system
O_s	Origin of the source plane $z = z_s$
O_t	Origin of the target plane $z = z_t$
\mathbf{p}, \mathbf{p}^*	Momentum (2D and 3D)
\mathbf{p}_s	Momentum (2D) at the source plane $z = z_s$
\mathbf{p}_t	Momentum (2D) at the target plane $z = z_t$
\mathbf{P}	SPD / SND matrix used in the numerical procedures
\mathbf{q}	Position coordinate (2D)
\mathbf{q}_s	Position coordinate (2D) at the source plane $z = z_s$
\mathbf{q}_t	Position coordinate (2D) at the target plane $z = z_t$
$\hat{\mathbf{s}}$	Direction of the incoming ray
S^2	Unit sphere
\mathbf{S}	Poynting vector in W/m^2
$\hat{\mathbf{t}}$	Direction of the outgoing ray
T	Hamilton's angular characteristic
$u(\mathbf{x})$	Location of the optical surface in Cartesian or stereographic coordinates
$u(\hat{\mathbf{s}}) = u(\phi, \theta)$	Location of the optical surface in spherical coordinates
$u_1(\mathbf{x})$	Geometrical variable related to the optical surface $u(\mathbf{x})$
$u_2(\mathbf{y})$	Geometric variable related to one of Hamilton's characteristic functions or to a second optical surface
U	Energy density of an electromagnetic wave in J/m^3
$v(\hat{\mathbf{s}}) = v(\phi, \theta)$	Location of the second optical surface for a double freeform lens in spherical coordinates
V	Hamilton's point characteristic
$\varphi(\mathbf{x})$	Phase of a wave
$w(\mathbf{y})$	Conjugate of $u(\mathbf{x})$ in a conjugate pair / c-convex or c-concave pair / G-convex or G-concave pair
$w^*(\mathbf{x})$	Legendre-Fenchel transform / c-transform / G-transform of $w(\mathbf{y})$

W	Hamilton's mixed characteristic
W^*	Hamilton's mixed characteristic of the second kind
(ξ, η)	Local Cartesian coordinates on a projection screen
x	Spatial coordinate in \mathbb{R}^3 or parametrization of a light ray (Chapter 2) / Cartesian or stereographic source coordinates (Chapter 3 – 9)
\mathcal{X}	Source domain in Cartesian or stereographic coordinate space
y	Cartesian or stereographic target coordinates
\mathcal{Y}	Target domain in Cartesian or stereographic coordinate space

Chapter 1

Introduction

1.1 Motivation

In modern optical design, the central challenge is to facilitate the efficient transport of light from A to B. Guiding light transport is important for the design of lamps, cameras and projectors, but also for the construction of earth-observing detectors in satellites [109] and high-precision metrology equipment for Internet of Things (IoT) applications and smartphones. The transport of light in these systems can be facilitated by components such as lenses and curved mirrors that transfer a given distribution of light from the light source to a desired distribution at a target. The interaction of light with lenses and mirrors generates a desired light output profile for numerous devices in daily life [44, 155].

One such device is a modern light bulb containing a light-emitting diode (LED). The Nobel-prize winning technology of the LED was developed in the early 1990s [139]. An LED is a semiconductor light source that can be inserted easily into an electrical circuit. When a voltage is applied to the circuit, electrons within the semiconductor recombine with holes and release energy in the form of light and heat. The color of the light is determined by the material of the semiconductor and resultant energy strength. The first LEDs emitted blue light, but the addition of a phosphor layer on top of the LED allowed for the development of LEDs emitting white light [139].

Components of an LED lamp

An LED lighting system is composed of electronic components, thermal components, mechanical components and optical components; see Figure 1.1. The illumination is generated by an electrical component which contains an LED,

also called an *LED driver*. Phosphor particles inside a diffuse layer on top of the LED scatter, absorb and re-emit incident light to create a white outgoing beam. The thermal component or *heat sink* is needed to dissipate heat. The white light emanating from the LED is redirected by optical components, consisting of reflectors (mirrors), lenses, diffusers and/or absorbers. Lastly, the mechanical components refer to all remaining material needed for the construction of the lamp.

Efficiency of an LED lamp

LED light bulbs are highly energy efficient and durable. Compared to incandescent bulbs, they do not have filaments that burn out, use less electricity and do not heat up as much. These characteristics make LEDs more energy efficient and durable, with lifetimes of 25,000 to 60,000 hours that easily surpass fluorescent tubes with a lifetime of approximately 15,000 hours, and incandescent light bulbs with only around 1,000 hours [121]. Their price is also decreasing since their development in the 1990s, making them cheaper in the long run.

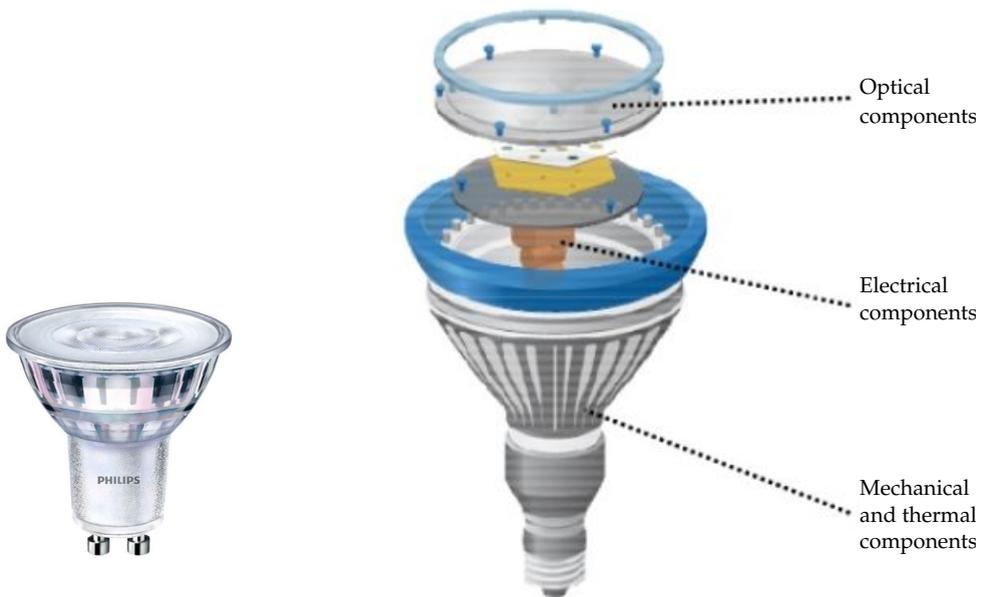


Figure 1.1: An LED spotlight with a schematic drawing, designed to replace the popular halogen spotlight. Source: Signify.

Imaging and nonimaging illumination optics

The optical components of an LED lighting system are responsible for the light output of the system, which is a specific lighting profile in many applications. Lenses and reflectors are macroscopic surfaces that direct light from point A to B through a transparent medium, following a path that obeys the famous principle of Fermat (1658), i.e., the optical length of the path is at a stationary point. The field of *illumination optics*, concerned with optical-system design, is an emerging field in recent years. It can be further subdivided into *imaging* and *nonimaging* illumination optics. Imaging optics involves techniques to form a perfect image of an object onto a target. For example, my DSLR (digital single-lens reflex) camera does a pretty good job capturing my memorable moments, as long as I carefully adjust the aperture, shutter speed and ISO sensitivity. The human eye is another example of an imaging optical system, with the cornea, pupil and lens working together to form images on the retina. Nonimaging optics, on the other hand, is concerned with the optimal transfer of light from A to B in terms of energy transport, used in lots of illumination devices such as luminaires, optical fibers, LCD backlights, car lights, etc. For example, Figure 1.2 shows a few target illumination patterns for the design of different types of car lights, e.g., a normal low beam ‘dips to the right’ with a downward/rightward bias to show the driver the road and signs ahead without blinding oncoming traffic.

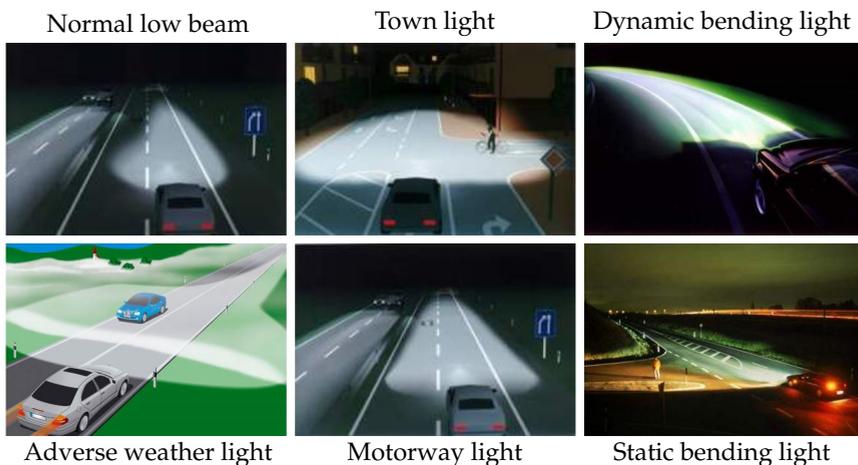


Figure 1.2: Target light distributions for car headlights. Different traffic situations require different types of lighting to allow the driver to see and be seen without blinding other road users. Source: Signify.

Freeform optics

We will be concerned with the field of nonimaging illumination optics, and in particular the field of *freeform nonimaging illumination optics*. Both imaging and nonimaging optical systems can be designed using freeform illumination optics, a technique for the design of surfaces without any symmetries that is used in the development of high-quality optical systems. Conventional rotationally symmetric lenses and mirrors have simple shapes, either convex or concave, which have limitations. These shapes cannot produce certain light-beam paths, so lenses and mirrors with a more complex aspherical or freeform surface are needed. A nonimaging example is a lens with a peanut-shape for street-lighting purposes; see Chapter 7.

An LED light source emits light with roughly a *Lambertian* light intensity, which means that the luminous intensity emitted from the LED surface is directly proportional to the cosine of the angle between the direction of the light and the normal to the LED surface. A freeform reflector or lens can redirect the light to change the intensity profile of the LED lamp to a desired non-symmetric light output.

In this thesis, the goal is to compute such freeform optical surfaces. An important assumption we make is to ignore the finite dimensions of the LED. We either assume the LED emits a parallel outgoing beam or approximate the LED as an ideal point source. These sources are examples of *zero-étendue* sources, i.e., a parallel source emits all light in the same direction and a point source emits a bundle of rays from one and the same position.

Manufacturing freeform optical components

In the last decade, the use of freeform shapes in LED lamps has become viable because of mechanical advances and thermal advantages of using LEDs. An LED light source operates at lower temperatures, with a maximum operating temperature of approximately 25°C [98], which allows for the use of freeform optics composed of plastic materials. Optical diamond turning techniques are capable of producing molds in arbitrary shapes at high precision. To make the freeform product, these molds are injected with molten plastic material which will cool and harden. Optical diamond turning techniques have pushed the field of illumination optics to develop sophisticated and highly precise methods to compute freeform shapes that convert the energy of LED light sources to a desired energy (intensity) distribution.

Computing freeform optical components

Broadly, the methods for optical system design can be categorized as either *forward* or *inverse* methods. Forward methods compute the target distribution from a known source distribution and optical system, most commonly using Monte-Carlo ray-tracing techniques [66]. A large number of rays are simulated to randomly emit from a light source using the LED's source light distribution. Subsequently, the illuminance, intensity or other photometric variables are computed at a target receiver. The design of the optical system can be improved by making modifications to the optical elements and subsequently re-evaluating the output target distribution via ray tracing. Drawbacks of such forward methods are that ray tracing can be slow if high precision is required and that the approach to create an improved design is often based on trial and error [66]. Filosa et al. [55, 56] recently developed a new ray-tracing method based on the phase space representation of the source and target domains, which improves the accuracy and reduces computation time of the classical approach.

Inverse methods directly compute the optical system converting the light from the source into the specified output. There are different design strategies, depending on the type of system. For rotationally symmetric surfaces that are used in, e.g., the reflector of an illumination spot or in a camera lens, the freeform shape is determined by an integral, given the input and output intensities [101]. For freeform surfaces the design of optical elements is more difficult. Such optical surfaces are used, e.g., in satellites [109] or outdoor lighting [150].

One approach for the inverse design of freeform optical surfaces uses the principles of geometrical optics and conservation of energy to derive a partial differential equation (PDE) for the location of the optical surface. With the laws of reflection and/or refraction of geometrical optics it is possible to construct an optical mapping that connects coordinates of the source and target domains. Substituting the mapping into the relation for energy conservation leads to a fully nonlinear second-order elliptic partial differential equation, which can be written as a so-called *generated Jacobian equation*, a term coined by the Australian mathematician *Neil Trudinger* [145].

The second-order nonlinear PDE is different for each optical system and cannot be solved analytically. If we consider the simplest optical systems, we think of the source to emit either a parallel beam of rays or a cone of rays emitted from a single point. The target lies in the near or far field, e.g., it is specified on a plane relatively close or far from the optical surface(s), or the target is required to be reached by a parallel beam or coincides with a

single point. Using a minimum number of either reflector or lens surfaces, we can already think of 16 different optical systems; see Figure 3.1 and 3.2 in Chapter 3. Actually, I should say 14, because we can eliminate two systems by symmetry (parallel-to-point reflector and lens are similar to point-to-parallel reflector and lens, respectively, by reverting the direction of the light rays). The formulation of the problem for each system is quite complicated, since the design depends on the light distributions in the source and target domains each presented as a function of the local spatial and angular coordinates.

Freeform design problems become even more intricate if we add scattering and absorbing media. White light emanating from LEDs is frequently scattered by a freeform diffuser. Presently, diffuser design is based on trial and error and only a few research studies describe light scattering coatings, diffusers, or suspensions with freeform optics [113, 114]. Because of large-scale differences it is difficult to apply conventional optical models to microscopically structured materials which are present in macroscopic freeforms. Figure 1.3 is an artistic illustration of the interplay between microscopic scattering and macroscopic refraction or reflection by freeform surfaces.

In this thesis, I am not attempting to solve freeform design problems with scattering and/or absorbing media just yet. Instead, the focus is on making freeform illumination optics readily accessible by providing a framework to solve any generated Jacobian equation.

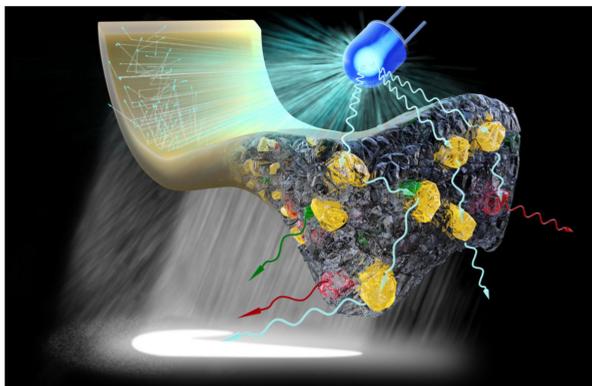


Figure 1.3: Light distribution from a blue LED source (top, blue) is transferred as efficiently as possible to a desired distribution on a target plane (bottom) in a device by an illumination system. It illustrates the combination of freeform optics (macroscopic) and of nanophotonic media (light scattering microscopic particles). See www.freeformscatteringoptics.com.

1.2 Main results of this thesis

One main result of this thesis is the development of a generic framework to derive generated Jacobian equations for the 16 base-case optical systems.

Some of these optical systems can be described using a cost function in optimal transport theory [148]. For these systems the generated Jacobian equation is also called a *generalized Monge-Ampère equation* [148, p. 282]. The cost functions can be derived using Hamilton’s characteristic functions of optical path length. We published the derivations of a variety of non-quadratic cost functions in [131–134, 141, 142, 162]. Among these articles are two journal articles I published on logarithmic cost functions for a point-to-far-field reflector and point-to-far-field lens:

- L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Inverse reflector design for a point source and far-field target. *J. Comput. Phys.*, 408:109283, 2020.
- L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Free-form lens design for a point source and far-field target. *J. Opt. Soc. Am. A*, 36(11):1926–1939, 2019.

However, many optical systems cannot be cast in the optimal-transport framework, e.g., systems involving near-field targets. For these systems we generalize the concept of a cost function to a *generating function*. We developed a method to derive the generating functions and corresponding generated Jacobian equations using Hamilton’s characteristic functions. In the following article, we found the generating functions and the generated Jacobian equations for 2 base cases, a point-to-far-field lens and a parallel-to-near-field reflector:

- L. B. Romijn, J. H. M. ten Thije Boonkkamp, M. J. H. Anthonissen, and W. L. IJzerman. An iterative least-squares method for generated Jacobian equations in freeform optical design. *SIAM J. Sci. Comput.*, 43(2):B298–B322, 2021.

In this thesis, we present the mathematical formulation of the generating functions for 5 out of the 16 base cases (and their cost functions if they exist). We also give a complete overview of the 16 base cases with all generating functions and cost functions at the end of Chapter 3.

The second main result of this thesis is our numerical algorithm to solve generated Jacobian equations. This least-squares algorithm can be used to compute freeform optical surfaces and optical mappings from source to target.

We have two main versions of the algorithm, one taking the optimal-transport cost function as input, if it exists, and the other taking the generating function as input. We call the first version of the algorithm the *generalized least-squares* (GLS) algorithm, since it solves generalized Monge-Ampère equations. It is used in the first two publications mentioned above. Similarly, we call the second version of the algorithm the *generated Jacobian least-squares* (GJLS) algorithm, since it solves generated Jacobian equations. It is used in the third publication mentioned above. The GJLS algorithm is one of the first numerical algorithms capable of solving generated Jacobian equations. It can solve all 16 base cases, while the GLS algorithm can solve 9 out of the 16 base cases.

The GJLS algorithm has opened up new possibilities for system design. We can think of optical systems outside of the 16 base cases, which brings us to the third main result of this thesis. The design of a double freeform lens with a point source and far-field target requires two freeform surfaces and adds another level of complexity compared to the base-case systems. We can formulate two generating functions for this system, one for each optical surface. To compute the freeform surfaces we run the numerical algorithm twice. In this thesis, we will show that by doing this we can formulate a tuning parameter to distribute the refractive power of the lens over both freeform surfaces. By varying this parameter we can compute multiple solutions to the same problem which differ in design. The main results of this design method are published in one of my journal articles:

- L. B. Romijn, J. H. M. ten Thije Boonkkamp, M. J. H. Anthonissen, and W. L. IJzerman. Generating-function approach for double freeform lens design. *J. Opt. Soc. Am. A*, 38(3):356–368, 2021.

This work is a first step towards handling complicated optical systems with more design freedom.

1.3 Outline of this thesis

This thesis is organized as follows. In Chapter 2, we present the fundamentals of geometrical optics, which describes light in terms of rays. We start with Maxwell's equations of electromagnetism and use the short-wavelength approximation to derive the well-known *eikonal equation*. We use the eikonal equation to derive a system of ordinary differential equations (ODEs) for a light ray and present Fermat's principle and the laws of reflection and refraction. Subsequently, we move to Hamiltonian optics, which describes the propagation of rays through an optical system by a set of *canonical equations*.

We introduce Hamilton's *characteristic functions*, which are measures of optical path length.

In Chapter 3, we present the mathematical formulation for 5 out of the 16 base cases. We consider a parallel-to-far-field reflector, a parallel-to-near-field reflector, a point-to-far-field reflector, a point-to-far-field lens, and a point-to-parallel reflector system. For each system we will use Hamilton's characteristic functions to derive the generating function. For some systems we will show that we can also find an optimal-transport cost function. Using energy conservation, i.e., all the light from the source should end up at the target, we can derive a Jacobian equation for the mapping. For some systems we will also find the mapping explicitly. At the end of Chapter 3, we will present an overview of all 16 base cases with all generating functions and cost functions.

The theoretical background on generated Jacobian equations will be discussed in Chapter 4. We will start with the *Legendre-Fenchel transform* and show that convex analysis can be used to derive the standard Monge-Ampère equation. Subsequently, we generalize this theory to *c-convex and c-concave functions* in optimal transport theory and demonstrate that we can derive generalized Monge-Ampère equations for optical systems with a cost function. We can use these cost functions to derive the optical mapping implicitly. We further generalize this theory to *G-convex functions* and generated Jacobian equations. For all optical systems described by a generating function in Chapter 3 we can derive this equation and find an implicit expression for the optical mapping.

In Chapter 5, we give an overview of the current literature on freeform illumination optics. In particular, we present the literature on numerical methods that are used to solve optical-design problems.

In Chapter 6, we give a full description of our numerical procedure used to solve generated Jacobian equations. First, we present the least-squares algorithm in an optimal-transport framework, the generalized least-squares (GLS) algorithm. Second, we extend this algorithm to polar coordinates for the source domain. Third, we present a new version of the least-squares algorithm which takes the generating function of an optical system as input, which is the generated Jacobian least-squares (GJLS) algorithm.

Numerical experiments are included in Chapter 7 for our subset of 5 out of the 16 base cases. Several experiments are done to assess the performance of the algorithm. Other examples are related to applications. We compute a peanut lens for roadlighting purposes. We also compare the performance of the generalized algorithm with the generated Jacobian algorithm for a point-to-far-field lens, for which we can formulate a cost function and a generating function.

In Chapter 8, we present the mathematical formulation of a lens with two freeform surfaces using a point source and far-field target. This system is not a base case but adds another level of complexity. For this system we can derive two generating functions, one for each optical surface, and we have an extra degree of freedom in the design by formulating an *intermediate target intensity*.

In Chapter 9, we present two numerical experiments for the double freeform lens. The first example tests the accuracy and efficiency of the numerical algorithm, and the second example illustrates how we can distribute the refractive power over both surfaces of the lens. By introducing a tuning parameter we can compute multiple solutions that generate the same light output but differ in design.

In Chapter 10, we conclude this thesis with a summary and give recommendations for future research.

Chapter 2

The Principles of Geometrical Optics

In a given transparent uniform medium, light appears to travel in straight-line paths or *rays*, already observed approximately 300 B.C. by Euclid [79]. When you walk along the street on a sunny day, your body obstructs part of the sun's light hitting you and you see a shadow shaped like you. This phenomenon was also observed by *Isaac Newton* in his *Treatise on Opticks* (1704), who thought that 'corpuscular light-particles' contained mass and were directed by his famous laws of motion [108].

Christiaan Huygens (1678) had a different proposition. He suggested that light moves in a wave-like motion. Every point reached by a luminous disturbance becomes a source of a spherical wave; the sum of these secondary waves determines the form of the wave at any subsequent time. He was able to provide a qualitative explanation of linear and spherical wave propagation, and to derive the laws of reflection and refraction using this principle [80]. This was in contradiction with the corpuscular theory of light, although both had given the correct expression for the refraction formula (Snell's law) [149].

Only later it would become clear that Newton's corpuscular theory was wrong. It was not until 1865, that *James Clerk Maxwell* posited that all that was hitherto known about light could be explained in terms of electromagnetic energy, usually described as electromagnetic waves [103]. Later in the twentieth century, Planck, Einstein and Bohr exonerated Newton somewhat by noting that electromagnetic energy is indeed quantized by tiny, albeit massless, particles called 'photons' [73].

The dimension of lenses and reflectors designed for common devices is typically in the order of millimeters to centimeters while the wavelength of light is measured in nanometers. Light appears to propagate in rays, only be-

cause the wavelength is short compared to the distances traveled. Geometrical optics, which is a branch of optics that describes light propagation in terms of rays, is a macroscopic approximation of the solutions to Maxwell's equations of electromagnetism. In this chapter, we will derive this approximation, and present the fundamentals of this branch of optics.

2.1 From electromagnetic theory to geometrical optics

Electric and magnetic fields, also known as electromagnetic fields, are composed of waves of electric and magnetic fields moving together. The fundamental equations of electrodynamics are known as Maxwell's equations. They are commonly used as a mathematical model for technologies with electric, optical or radio components, such as power generation, electric motors, wireless communication and light.

We consider an electric field $\mathbf{E} = \mathbf{E}(x, t)$, a magnetic induction $\mathbf{B} = \mathbf{B}(x, t)$, a displacement field $\mathbf{D} = \mathbf{D}(x, t)$, and magnetic field $\mathbf{H} = \mathbf{H}(x, t)$, all in \mathbb{R}^3 , dependent on the spatial coordinate \mathbf{x} and time t . In their most general form, for time-varying fields in the presence of a dielectric, i.e., nonconducting, medium, the equations can be written in integral form as follows [78, Ch. 3]:

$$\oiint_{\partial V} \mathbf{D} \cdot d\mathbf{S} = \iiint_V \rho \, dV, \quad (\text{Gauss's law}) \quad (2.1a)$$

$$\oiint_{\partial V} \mathbf{B} \cdot d\mathbf{S} = 0, \quad (\text{Gauss's law for magnetism}) \quad (2.1b)$$

$$\oint_{\partial A} \mathbf{E} \cdot d\mathbf{s} = -\frac{d}{dt} \iint_A \mathbf{B} \cdot d\mathbf{S}, \quad (\text{Faraday's law of induction}) \quad (2.1c)$$

$$\oint_{\partial A} \mathbf{H} \cdot d\mathbf{s} = \iint_A \mathbf{J} \cdot d\mathbf{S} + \frac{d}{dt} \iint_A \mathbf{D} \cdot d\mathbf{S}, \quad (\text{Ampère's law}) \quad (2.1d)$$

where ρ is the free electric charge density, \mathbf{J} is the electric current density, V is a volume enclosed by the boundary ∂V , A is a surface bounded by the closed contour ∂A , $d\mathbf{s}$ is an infinitesimal piece of the contour ∂A , and $d\mathbf{S}$ is an infinitesimal vector element of A or ∂V with magnitude equal to the area of an infinitesimal patch of the surface and direction normal to that patch. The orientation of ∂A is induced by the orientation of A , such that if you walk around the boundary ∂A in the direction of the tangential vector $\hat{\mathbf{T}}$ such that $d\mathbf{s} = \hat{\mathbf{T}} \, ds$ with your head in direction of the normal $\hat{\mathbf{n}}$ to A with $d\mathbf{A} = \hat{\mathbf{n}} \, dA$, then the boundary ∂A is oriented counterclockwise, i.e., the interior of A should be to the left of ∂A . We denote unit vectors using hats.

The electric displacement field \mathbf{D} is related to the electric field by the equation $\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P}$, accounting for free and bound charges with \mathbf{P} the

density of the net permanent and induced electric dipole moments. Here, ϵ_0 is the permittivity of *free space*, a measure of electric polarizability of vacuum in response to an electric field. The field \mathbf{P} can be expressed as $\mathbf{P} = \epsilon_0\chi_e\mathbf{E}$, where χ_e is the electric susceptibility, a dimensionless proportionality constant that indicates the degree of polarization of a dielectric material in response to an applied electric field.

The magnetizing field \mathbf{H} is related to the magnetic induction \mathbf{B} by the equation $\mathbf{B} = \mu_0\mathbf{H} + \mathbf{M}$, where \mathbf{M} is the density of the net permanent and induced magnetic dipole moments. Here, μ_0 is the permeability of free space, a measure of the degree of induced magnetism in vacuum in response to a magnetic induction. The field \mathbf{M} can be written as $\mathbf{M} = \mu_0\chi_m\mathbf{H}$, where χ_m is the magnetic susceptibility, a dimensionless proportionality constant that indicates the degree of magnetization of a material in response to an applied magnetic field.

In short, we write $\mathbf{D} = \epsilon_0\mathbf{E} + \mathbf{P} = \epsilon_0(1 + \chi_e)\mathbf{E} = \epsilon\mathbf{E}$ and similarly, $\mathbf{B} = \mu_0\mathbf{H} + \mathbf{M} = \mu_0(1 + \chi_m)\mathbf{H} = \mu\mathbf{H}$. The parameters $\epsilon = \epsilon_0(1 + \chi_e)$ and $\mu = \mu_0(1 + \chi_m)$ are the permittivity and permeability of a dielectric, respectively. We assume that ϵ and μ are piecewise constant such that $\mathbf{D} = \epsilon\mathbf{E}$ and $\mathbf{B} = \mu\mathbf{H}$ are linear constitutive relations.

In words, we can summarize Maxwell's equations as

- (2.1a): The net flux of displacement field \mathbf{D} through any closed surface ∂V is equal to the total charge enclosed by V .
- (2.1b): The net flux of a magnetic induction \mathbf{B} through any closed surface ∂V is zero.
- (2.1c): A time-varying magnetic induction \mathbf{B} will have an electric field \mathbf{E} associated with it, i.e., the total work done to move a unit of charge around a fixed closed curve ∂A equals minus the time variation of the flux of the magnetic induction through the surface A . The negative sign comes from Lenz's law, stating that the direction of an induced current is always such as to oppose the change in the circuit or the magnetic induction that produces it.
- (2.1d): A magnetizing field \mathbf{H} will be accompanied by a time-varying displacement field, i.e., the line integral of \mathbf{H} tangent to a closed curve ∂A (the circulation) is equal to the total current through the open surface A and the time variation of the flux of the displacement field \mathbf{D} through A . Maxwell himself later added the latter integral to Ampère's law in 1862, which he called the *displacement current*, by realizing that moving electric charges are not the only source of magnetic fields. A time-varying

electric field can create a magnetic field, complementary to Faraday's law, in which a time-varying magnetic field can produce an electric field. What Maxwell called the 'mutual embrace' of electric and magnetic fields can produce propagating electromagnetic waves. This would not be possible without the displacement current.

Without electric charges and currents, i.e., $\rho = 0$ and $\mathbf{J} = 0$, we can write the differential form of Maxwell's equations in terms of \mathbf{E} and \mathbf{H} as

$$\nabla \cdot \mathbf{E} = 0, \quad (2.2a)$$

$$\nabla \cdot \mathbf{H} = 0, \quad (2.2b)$$

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t}, \quad (2.2c)$$

$$\nabla \times \mathbf{H} = \epsilon \frac{\partial \mathbf{E}}{\partial t}, \quad (2.2d)$$

by applying Gauss's theorem to (2.1a) and (2.1b), and Stokes' theorem to (2.1c) and (2.1d). Note that we also used Leibniz's rule to take the differentiation inside the integral for a fixed surface A .

Taking the curl of (2.2c) and (2.2d) results in

$$\nabla \times (\nabla \times \mathbf{E}) = -\mu \nabla \times \frac{\partial \mathbf{H}}{\partial t} = -\mu \frac{\partial}{\partial t} \nabla \times \mathbf{H} = -\mu \epsilon \frac{\partial^2 \mathbf{E}}{\partial t^2}, \quad (2.3a)$$

$$\nabla \times (\nabla \times \mathbf{H}) = \epsilon \nabla \times \frac{\partial \mathbf{E}}{\partial t} = \epsilon \frac{\partial}{\partial t} \nabla \times \mathbf{E} = -\mu \epsilon \frac{\partial^2 \mathbf{H}}{\partial t^2}. \quad (2.3b)$$

Using the vector identity $\nabla \times (\nabla \times \mathbf{v}) = \nabla(\nabla \cdot \mathbf{v}) - \nabla^2 \mathbf{v}$ combined with (2.2a) and (2.2b) gives the wave equations for \mathbf{E} and \mathbf{H} as

$$\frac{\partial^2 \mathbf{E}}{\partial t^2} = \frac{1}{\mu \epsilon} \nabla^2 \mathbf{E}, \quad (2.4a)$$

$$\frac{\partial^2 \mathbf{H}}{\partial t^2} = \frac{1}{\mu \epsilon} \nabla^2 \mathbf{H}. \quad (2.4b)$$

The coefficient $\frac{1}{\mu \epsilon}$ is the speed of the light in the medium squared, i.e., the electric and magnetic waves move at speed $v = 1/\sqrt{\mu \epsilon}$. The speed of light in vacuum is $c = 1/\sqrt{\mu_0 \epsilon_0} \approx 3.00 \cdot 10^8$ m/s. The refractive index of a medium is the ratio of the speed of light c in vacuum and the speed of light v in the medium, i.e.,

$$n = \frac{c}{v} = c \sqrt{\mu \epsilon}. \quad (2.5)$$

Most commonly, $n = n(\mathbf{x}) \geq 1$ depends on the spatial coordinate \mathbf{x} and is piecewise constant, and clearly $n = 1$ in vacuum. In general, the more optically

dense a material is, the slower the speed v of a wave in the material. For most materials the refractive index is wavelength-dependent, since the electric permittivity is wavelength-dependent (we should actually write $\epsilon = \epsilon(\lambda)$ with λ the wavelength). Mostly, a single value for n is reported for a material, typically measured at $\lambda = 633$ nm (the wavelength of a helium-neon laser).

Maxwell's equations cover wave phenomena of any wavelength and frequency, although in vacuum electromagnetic waves all travel at the same speed. Maxwell's equations are applicable in a broad range of fields such as in electrical engineering, power generation, wireless communication and radar technology. Electromagnetic radiation includes radio waves and microwaves, as well as infrared, ultraviolet, gamma, and X-rays. The typical wavelength of visible light is in the range of 384 to 769 nm, and forms just a small subset of all electric and magnetic phenomena described by Maxwell's equations [78].

Electromagnetic waves carry energy. The energy density U of an electromagnetic wave is defined as

$$U = \frac{1}{2} (\epsilon |\mathbf{E}|^2 + \mu |\mathbf{H}|^2). \quad (2.6)$$

The Poynting vector \mathbf{S} , named after *John Henry Poynting* (1852–1914), is

$$\mathbf{S} = \mathbf{E} \times \mathbf{H}. \quad (2.7)$$

Maxwell's equations give a conservation law for the energy density. Taking the inner product of (2.2c) with \mathbf{H} and (2.2d) with \mathbf{E} and subtracting gives

$$(\nabla \times \mathbf{H}) \cdot \mathbf{E} - (\nabla \times \mathbf{E}) \cdot \mathbf{H} - \left(\epsilon \mathbf{E} \cdot \frac{\partial \mathbf{E}}{\partial t} + \mu \mathbf{H} \cdot \frac{\partial \mathbf{H}}{\partial t} \right) = 0.$$

Using the identity $\nabla \cdot (\mathbf{u} \times \mathbf{v}) = (\nabla \times \mathbf{u}) \cdot \mathbf{v} - \mathbf{u} \cdot (\nabla \times \mathbf{v})$ gives

$$\nabla \cdot (\mathbf{E} \times \mathbf{H}) + \frac{1}{2} \frac{\partial}{\partial t} (\epsilon |\mathbf{E}|^2 + \mu |\mathbf{H}|^2) = 0,$$

i.e.,

$$\frac{\partial U}{\partial t} + \nabla \cdot \mathbf{S} = 0, \quad (2.8)$$

which means that, indeed, the energy of the electromagnetic wave is carried in the direction of the Poynting vector and we could call \mathbf{S} the energy flux.

You might wonder how electromagnetic fields are generated in the first place. Since we are dealing with waves in the electromagnetic field, the source of electromagnetic radiation is always the result of nonuniformly moving charges. For instance, the sun is made of plasma, which is a gas of bare ions and electrons. The energy released from nuclear fusion results in heat and electromagnetic energy.

2.2 Short wavelength approximation

Geometrical optics can be formally defined as the high-frequency limit of Maxwell's equations. In order to take this limit, let us look at the solutions of Maxwell's equations (2.2) as a time harmonic field. The spatial dependence of the phase of the waves is denoted by φ [12]. We consider

$$\mathbf{E}(\mathbf{x}, t) = \mathbf{e}(\mathbf{x})e^{i(\kappa\varphi(\mathbf{x})-\omega t)}, \quad (2.9a)$$

$$\mathbf{H}(\mathbf{x}, t) = \mathbf{h}(\mathbf{x})e^{i(\kappa\varphi(\mathbf{x})-\omega t)}, \quad (2.9b)$$

where $\kappa = \omega/c$ is the free-space wave number, ω the angular frequency, and \mathbf{e} and \mathbf{h} the yet unknown spatial field amplitude vectors. The free-space wave number and wavelength λ are related as $\kappa = \frac{2\pi}{\lambda}$. Consequently, the speed of light in a medium v in (2.5) and the angular frequency are related as $v = \frac{c}{n} = \lambda \frac{\omega}{2\pi n}$. Substituting (2.9) into Maxwell's equations (2.2), in a medium without free electric charges and currents, and applying the product rules of curl and divergence, gives

$$\nabla \cdot \mathbf{e} + i\kappa \nabla \varphi \cdot \mathbf{e} = 0, \quad (2.10a)$$

$$\nabla \cdot \mathbf{h} + i\kappa \nabla \varphi \cdot \mathbf{h} = 0, \quad (2.10b)$$

$$\nabla \times \mathbf{e} + i\kappa \nabla \varphi \times \mathbf{e} = i\omega\mu\mathbf{h}, \quad (2.10c)$$

$$\nabla \times \mathbf{h} + i\kappa \nabla \varphi \times \mathbf{h} = -i\omega\epsilon\mathbf{e}. \quad (2.10d)$$

For infinitely short wavelengths we have that $\lambda \rightarrow 0$ and $\kappa \rightarrow \infty$ since $\lambda = \frac{2\pi}{\kappa}$. Dividing (2.10) by κ using $\omega = \kappa c$ and neglecting the terms that tend to 0 as $\kappa \rightarrow \infty$ gives

$$\nabla \varphi \cdot \mathbf{e} = 0, \quad (2.11a)$$

$$\nabla \varphi \cdot \mathbf{h} = 0, \quad (2.11b)$$

$$\nabla \varphi \times \mathbf{e} = c\mu\mathbf{h}, \quad (2.11c)$$

$$\nabla \varphi \times \mathbf{h} = -c\epsilon\mathbf{e}. \quad (2.11d)$$

The first and second equation indicate that electromagnetic waves are transverse waves with $\mathbf{e} \perp \nabla \varphi$ and $\mathbf{h} \perp \nabla \varphi$. The latter two equations imply that $\mathbf{e} \perp \mathbf{h}$. An illustration of a transverse electromagnetic wave is drawn in Figure 2.1.

A wave-phenomenon called *polarization* specifies the geometrical orientation of the oscillations of the transverse electromagnetic waves. Usually, the polarization of electromagnetic waves refers to the direction of the electric field [78]. Examples are linear polarization, where the fields oscillate in a

single direction, and circular or elliptical polarization, where the fields rotate at a constant rate in a plane perpendicular to the direction of the wave.

An equation involving only φ can be found by eliminating \mathbf{e} or \mathbf{h} from (2.11c) and (2.11d). For example, substituting \mathbf{h} from (2.11c) into (2.11d) gives

$$\nabla\varphi \times (\nabla\varphi \times \mathbf{e}) + c^2\mu\epsilon\mathbf{e} = \mathbf{0}. \quad (2.12)$$

Using the identity $\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) = (\mathbf{u} \cdot \mathbf{w})\mathbf{v} - (\mathbf{u} \cdot \mathbf{v})\mathbf{w}$, reduces the equation to

$$(\nabla\varphi \cdot \mathbf{e})\nabla\varphi - (\nabla\varphi \cdot \nabla\varphi)\mathbf{e} + c^2\mu\epsilon\mathbf{e} = \mathbf{0}. \quad (2.13)$$

Since $\nabla\varphi \cdot \mathbf{e} = 0$ by (2.11a) and realizing that the relation must hold for all field amplitudes \mathbf{e} , we obtain

$$|\nabla\varphi|^2 - c^2\mu\epsilon = 0, \quad (2.14)$$

which can be further simplified using (2.5) to the well-known *eikonal equation*

$$|\nabla\varphi| = n, \quad (2.15)$$

or simply $|\nabla\varphi|^2 = n^2$. Substituting \mathbf{e} from (2.11d) into (2.11c) gives the same result. Equation (2.15) is also known as the equation of *wavefronts*, which are surfaces of constant phase, with $\varphi(\mathbf{x}) = \text{constant}$ as the level sets of φ . It is considered *the* fundamental equation of geometrical optics, from which other properties can readily be derived.

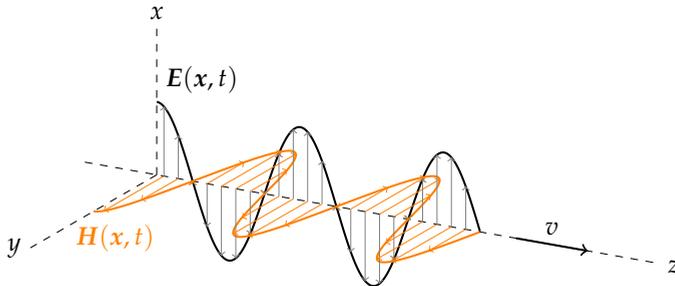


Figure 2.1: Illustration of an electromagnetic wave with electric field vector \mathbf{E} and magnetic field vector \mathbf{H} in Euclidean space propagating in the positive z -direction at speed v .

Another method to derive the eikonal equation from Maxwell's equations considers surfaces where the fields \mathbf{E} and \mathbf{H} are discontinuous [161, p. 11]. These discontinuities occur at the wavefronts and at points where μ and ϵ are discontinuous, for instance at an optical surface forming the interface between two dielectrics.

It can be shown that the Poynting vector \mathbf{S} in (2.7) is directed along $\nabla\varphi$. Introducing $\mathbf{S} = \mathbf{s}(\mathbf{x}) e^{2i(\kappa\varphi(\mathbf{x}) - \omega t)}$ and using (2.11c) and (2.11d) gives

$$\begin{aligned} \mathbf{s} &= \mathbf{e} \times \mathbf{h} \\ &= -\frac{1}{c^2\mu\epsilon} (\nabla\varphi \times \mathbf{h}) \times (\nabla\varphi \times \mathbf{e}) \\ &= -\frac{1}{n^2} [(\nabla\varphi \cdot (\mathbf{h} \times \mathbf{e}))\nabla\varphi - (\nabla\varphi \cdot (\nabla\varphi \times \mathbf{h}))\mathbf{e}], \end{aligned} \quad (2.16)$$

where we used $\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) = (\mathbf{u} \cdot \mathbf{w})\mathbf{v} - (\mathbf{u} \cdot \mathbf{v})\mathbf{w}$, and cyclic permutation of the scalar triple product. The second term in the square brackets vanishes since $(\nabla\varphi \times \mathbf{h})$ and $\nabla\varphi$ are orthogonal and the inner product vanishes. The expression can be rewritten as

$$\mathbf{s} = \frac{1}{n^2} (\nabla\varphi \cdot \mathbf{s}) \nabla\varphi. \quad (2.17)$$

From this equation we can infer that \mathbf{s} and $\nabla\varphi$ are parallel. Hence, the transport of energy in the direction of the Poynting vector \mathbf{S} is along the normal of the wavefront, i.e., along the direction of $\nabla\varphi$. This motivates the definition of light rays as orthogonal trajectories of the wavefronts.

Light rays are orthogonal trajectories of the wavefronts. Energy flows in the direction of propagation of the ray.

In the next section, we will use the eikonal equation to derive an ODE system for a light ray. Figure 2.2 shows an illustration of a *plane wave* traveling in the z -direction. We see three electromagnetic waves for which only one field component is shown. A plane wave is the wavefront of a collection of electromagnetic waves with a parallel direction of propagation. This parallel collection is also called a *collimated* light beam. The level sets where $\varphi(\mathbf{x})$ is constant have the same z -coordinate. Hence, the combined wavefronts of the waves form a plane propagating at the wave speed v and the light rays are the orthogonal trajectories to the planes, drawn as arrows in the direction $\nabla\varphi$. For non-collimated collections of rays the wavefronts are non-planar. The simplest example is a collection of spherical waves originating from a point source (starting with the same phase), which results in spherical wavefronts. This is because all points which are equidistant from the point source have the same phase.

Plane wavefronts: $\varphi = \text{constant}$

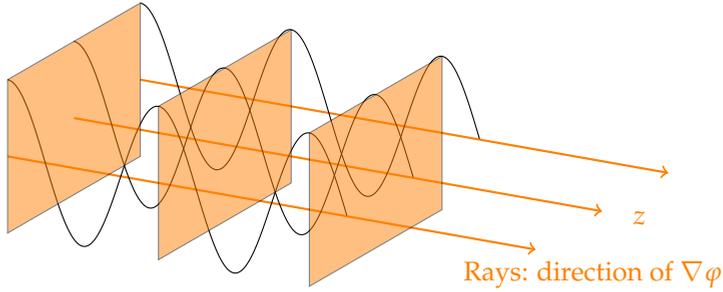


Figure 2.2: Illustration of a plane wave.

2.3 The ray equation and Fermat's principle

The eikonal equation (2.15) is a first-order nonlinear PDE, which can be used to find a set of ODEs for a light ray. We consider a generic nonlinear PDE for an unknown $u = u(\mathbf{x})$, written as

$$F(\mathbf{x}, u, \mathbf{p}^*) = 0, \quad \mathbf{p}^* = \nabla u. \quad (2.18)$$

The notation \mathbf{p}^* will become clear in Section 2.6. Parametrizing $\mathbf{x} = \mathbf{x}(s)$, $u = u(\mathbf{x}(s))$ and $\mathbf{p}^* = \mathbf{p}^*(\mathbf{x}(s))$, the system of ODEs provided by the method of characteristics [43] is given by

$$\frac{d\mathbf{x}}{ds} = \frac{\partial F}{\partial \mathbf{p}^*}, \quad (2.19a)$$

$$\frac{du}{ds} = \mathbf{p}^* \cdot \frac{\partial F}{\partial \mathbf{p}^*}, \quad (2.19b)$$

$$\frac{d\mathbf{p}^*}{ds} = -\mathbf{p}^* \frac{\partial F}{\partial u} - \frac{\partial F}{\partial \mathbf{x}}, \quad (2.19c)$$

where $\mathbf{x} = \mathbf{x}(s)$ is called the characteristic curve. Applied to the eikonal equation (2.15), i.e.,

$$F(\mathbf{x}, \varphi, \mathbf{p}^*) = |\mathbf{p}^*| - n(\mathbf{x}) = 0, \quad \mathbf{p}^* = \nabla \varphi, \quad (2.20)$$

we obtain

$$\frac{d\mathbf{x}}{ds} = \frac{1}{n} \mathbf{p}^*, \quad (2.21a)$$

$$\frac{d\varphi}{ds} = n, \quad (2.21b)$$

$$\frac{d\mathbf{p}^*}{ds} = \nabla n. \quad (2.21c)$$

From (2.21a) we observe that $\frac{dx}{ds}$ is parallel to $\nabla\varphi$ and we conclude that the characteristic curve x represents a light ray. It follows that $|\frac{dx}{ds}| = 1$ using the eikonal equation (2.15) and $p^* = \nabla\varphi$. Hence, s is the arc length and (2.21a) becomes

$$n \frac{dx}{ds} = \nabla\varphi. \quad (2.22)$$

Differentiating this equation and using (2.20) and (2.21) results in the *ray equation*

$$\frac{d}{ds} \left(n \frac{dx}{ds} \right) = \frac{dp^*}{ds} = \nabla n. \quad (2.23)$$

From this equation we can derive straightaway that rays are straight lines when the refractive index n is constant, since p^* is the direction of a ray. Dielectrics for which n is constant in all polarization directions, i.e., μ and ϵ are scalar quantities independent of the position x are called *isotropic*.

The ray equation is a system of ODEs that can be used to compute the trajectory of individual light rays without first solving the eikonal equation. We can also derive that these differential equations are the Euler-Lagrange equations [64] for the optical path length.

Definition 2.3.1. Let C be a continuous curve with endpoints P and Q , then its optical path length is

$$\int_C n(x(s)) ds, \quad (2.24)$$

where s is the arc length.

To derive the ray equation, we use the parametrization $x = x(\tau)$, with $\tau_1 \leq \tau \leq \tau_2$, and $x(\tau_1) = P$ and $x(\tau_2) = Q$ as endpoints of the curve C . The optical path length integral can be written as

$$\int_{\tau_1}^{\tau_2} n(x(\tau)) |x'(\tau)| d\tau =: \int_{\tau_1}^{\tau_2} L(x, x') d\tau, \quad (2.25)$$

with

$$L(x, x') = n(x) |x'|. \quad (2.26)$$

We can derive the Euler-Lagrange equations of (2.25) as

$$\mathbf{0} = \frac{d}{d\tau} \frac{\partial L}{\partial x'} - \frac{\partial L}{\partial x} = \frac{d}{d\tau} \left(n \frac{x'}{|x'|} \right) - \nabla n |x'|, \quad (2.27)$$

which is equal to the ray equation (2.23) if we choose $\tau = s$ since it holds that $|x'| = |dx/ds| = 1$.

With φ a solution to the eikonal equation and the directional derivative $\frac{dx}{ds}$ proportional to $\nabla\varphi$, the curve \mathcal{C} measures a stationary optical path length. *Stationary* in calculus of variations means that slight variations in the path do not affect the optical path length and the optical path length is a maximum, minimum or inflection point, which could be local or global.

Since the speed of the wave is $v = c/n$, the electromagnetic wave propagates at velocity v and we can rewrite the curve integral to

$$\int_{\mathcal{C}} \frac{c}{v(x(s))} ds = c \int_{t_1}^{t_2} dt, \quad (2.28)$$

where dt is the time needed for the ray to travel a distance ds with speed v . The optical path length along the curve \mathcal{C} from a point P to a point Q is equal to the product of the speed of light c in vacuum and the time needed for light to travel from P to Q with speed v .

Pierre de Fermat (1657) proposed a principle to understand both reflection and refraction at interfaces of optical media. A beam of light traversing an optical surface does not travel in a straight line or with a minimum spatial path between a point in the incident medium and one in the transmitting medium [78]. Fermat proposed the *principle of least time*, stating that a light ray is a curve on which the travel time is minimal. The more modern formulation of the principle uses the optical path length.

The principle of Fermat, known also as the principle of shortest optical path or the principle of least time, asserts that the optical path length

$$\int_P^Q n ds \quad (2.29)$$

of an actual ray between any two points P and Q is stationary with respect to variations of the path.

Formulated this way, this principle has its roots in calculus of variations. Stationary points may be found using the ray equation (2.23). Figure 2.3 is an illustration of variational paths. The orange path x from P to Q has a stationary optical path length with respect to small changes in the path δx . Note that in a medium with a variable index field $n = n(x)$, the path with a stationary optical path length is curved. Such a medium is called *anisotropic*.

Other related theorems concerning rays and wavefronts are the *Lagrange integral invariant*, the *Theorem of Malus and Dupin* and the *Huygens-Fresnel principle*; see [12, p. 130].

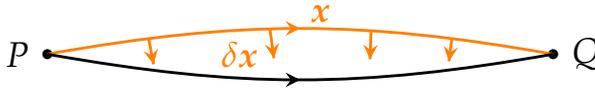


Figure 2.3: Illustration of variations to a path from P to Q with stationary optical path length (orange).

Here, we have shown that the eikonal equation (2.15) implies the ray equation (2.23). Fermat's principle also implies the ray equation, i.e., light rays that are solutions to (2.23) give a maximum or minimum of the optical path length. In fact, Fermat's principle is the formal solution to the eikonal equation. Fermat's principle is usually proven rigorously from Lagrange's integral invariant, but can also be derived by writing the system of equations (2.21) as a *Hamiltonian system* [93, p. 13].

2.4 Law of reflection

At an optical interface separating two media, such as air and glass, discontinuities occur in ϵ and μ , and thus also in n . When a beam of light hits an optical surface that acts as an interface, some light is always scattered backwards, which we call *reflection*, and some light is transmitted through the interface in the new medium, which we call *refraction*. Here we consider specular reflection and refraction only, which means that the light rays reflect or refract and remain concentrated in a bundle. On the other hand, if a surface, e.g., a free-form diffuser, is rough, the light rays will reflect and diffuse in many different directions, which is called diffuse reflection or refraction, also simply referred to as *scattering*. The relationships between the direction of the incoming beam, specularly reflected beam and transmitted beam are captured by the *law of reflection* and *law of refraction*. These two laws can be directly derived from Fermat's principle. In this section, we will derive the law of reflection. The law of refraction is treated in Section 2.5.

At an optical interface, light is always both reflected and refracted (except at *Brewster's angle* of perfect transmission) although the energy is split between the two in various ratios. *Augustin-Jean Fresnel* (1823) wrote down physically accurate equations for the fraction of energy that is reflected and transmitted as a function of the incident angle [78]. For simplicity, in this thesis we assume that reflectors act as perfect mirrors and reflect all the incoming light. Examples of highly reflective surfaces are electrically conductive metals, such

as a silver mirror. Similarly, we assume that lenses act as perfect transmitters and we ignore Fresnel reflections. A glass lens transmits most incoming light. For an air-glass interface, Fresnel's equations state that a light beam at normal incidence reflects 4% of the incoming light and transmits 96%.

In Figure 2.4 a light ray originating from a point P with direction \hat{s} hits a horizontal reflective surface and changes direction to \hat{t} arriving at point Q . Unit vectors are denoted by hats. In geometrical optics, we can derive that *the incident ray, reflected ray, and normal to the reflective surface lie in the same plane*, called the *plane of incidence*, which can be explained from describing the system as a Hamiltonian system; we will derive this result in Section 2.6.

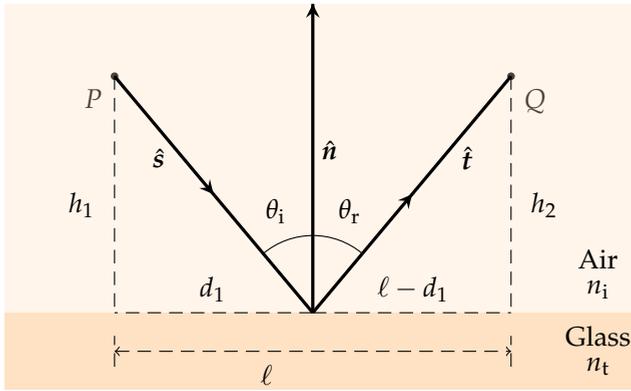


Figure 2.4: Illustration of the law of reflection at an optical interface.

The horizontal distance between P and Q is given by ℓ . The two rays have a common point being the incident point on the surface, here separating air and glass. The normal \hat{n} is perpendicular to the surface and directed towards the light source, i.e., $\hat{s} \cdot \hat{n} < 0$, and the angles of incidence and reflection satisfy $0 \leq \theta_i, \theta_r \leq \pi/2$, respectively. Using Figure 2.4 we can calculate the time required for the light to travel from P to Q , parametrized by d_1 , by calculating the length and dividing by the speed, i.e.,

$$t(d_1) = \frac{\sqrt{d_1^2 + h_1^2}}{v} + \frac{\sqrt{(\ell - d_1)^2 + h_2^2}}{v}, \quad (2.30)$$

where $v = c/n_i$. Using the principle of least time we set

$$\frac{dt}{dd_1} = \frac{d_1}{v\sqrt{d_1^2 + h_1^2}} - \frac{\ell - d_1}{v\sqrt{(\ell - d_1)^2 + h_2^2}} = 0. \quad (2.31)$$

Note that we can verify that indeed $\frac{d^2t}{dd_1^2}(d_1^*) > 0$, where d_1^* is the solution of (2.31), so that we compute a minimum. Equation (2.31) can be rearranged to

$$\frac{d_1}{\sqrt{d_1^2 + h_1^2}} = \frac{(\ell - d_1)}{\sqrt{(\ell - d_1)^2 + h_2^2}}, \quad (2.32)$$

i.e.,

$$\sin \theta_i = \sin \theta_r. \quad (2.33)$$

Since $0 \leq \theta_i, \theta_r \leq \pi/2$ we get

$$\theta_i = \theta_r, \quad (2.34)$$

i.e., *the angle of incidence is equal to the angle of reflection*. This is the scalar version of the law of reflection.

Next, we derive the vectorial version. We can decompose the vector \hat{s} into two orthogonal vectors as

$$\hat{s} = \mathbf{s}_1 + \mathbf{s}_2, \quad (2.35)$$

as drawn in Figure 2.5. Assuming the incident ray, reflected ray and normal to the surface all lie in the same plane, and since $\theta_i = \theta_r$ we can write

$$\hat{t} = -\mathbf{s}_1 + \mathbf{s}_2, \quad (2.36)$$

and

$$\mathbf{s}_1 = (\hat{s} \cdot \hat{n}) \hat{n}. \quad (2.37)$$

We also have that

$$\mathbf{s}_2 = \hat{s} - \mathbf{s}_1 = \hat{s} - (\hat{s} \cdot \hat{n}) \hat{n}. \quad (2.38)$$

Hence, substituting (2.37) and (2.38) into (2.36) gives

$$\hat{t} = \hat{s} - 2(\hat{s} \cdot \hat{n}) \hat{n}. \quad (2.39)$$

Note that this also accounts for the case $\theta_i = 0$. As a final note, we remark that the law of reflection is only dependent on the local normal to the surface at the point where the incident ray hits the surface, and independent of the curvature of the surface.

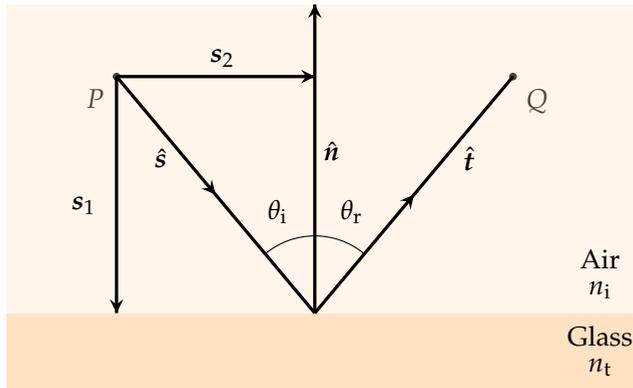


Figure 2.5: Illustration for the derivation of the vectorial law of reflection.

2.5 Snell's law

From the relation $n = c/v$ with $n > 1$ for most media other than vacuum, such as air, glass or water, we know that the speed v of light rays in the medium is smaller than the speed of light in vacuum. From a microscopic point of view, an electromagnetic wave passing through a medium causes other particles such as electrons to oscillate. The oscillating electrons emit their own electromagnetic waves that interact with the original wave. The resulting 'combined' wave travels at a slower speed than the speed of light. This 'slowing down' becomes visible as the abrupt bending of a light ray when it hits an interface between media of different refractive indices. When the refractive index changes continuously throughout an anisotropic medium, this slowing down becomes a smooth bending of the rays.

Actually, the index of refraction is found to vary with the frequency or wavelength of radiation. This phenomenon is called *dispersion* and is true for all transparent optical media, such as glass with $n \approx 1.5$. This dependency is a result of the interaction of electrons in the optical medium with the incoming light. A prime example of this is the separation of colors when light passes through a prism. When a white beam of light rays with wavelength in the visible spectrum of approximately 384 to 769 nm hits the prism, each wavelength has a different refractive index and is deflected in a different direction, showing a separation of colors in straight lines.

The bending at an optical interface separating media with different refractive index is captured by *Snell's law*. The Dutch astronomer *Willebrord Snel van Royen* (1621) is given most of the credit in high school and optics

textbooks, although he did not publish his result [78]. Early discoveries of the law date back to approximately 100 A.D. by *Ptolemy of Alexandria* [78]. It was *René Descartes* who published the law in 1637 and he used a proof based on the conservation of momentum using a derivation resembling Newton's corpuscular theory of light which followed in 1704 [137, 149]. Pierre de Fermat rejected Descartes' solution, because Descartes assumed that the denser the medium, the greater the speed of the light waves [149]. Fermat himself derived the exact opposite result using his principle of least time. Christiaan Huygens (1678) later showed how Snell's law could also be derived from the wave nature of light, which we now call the *Huygens–Fresnel principle* [78, 81].

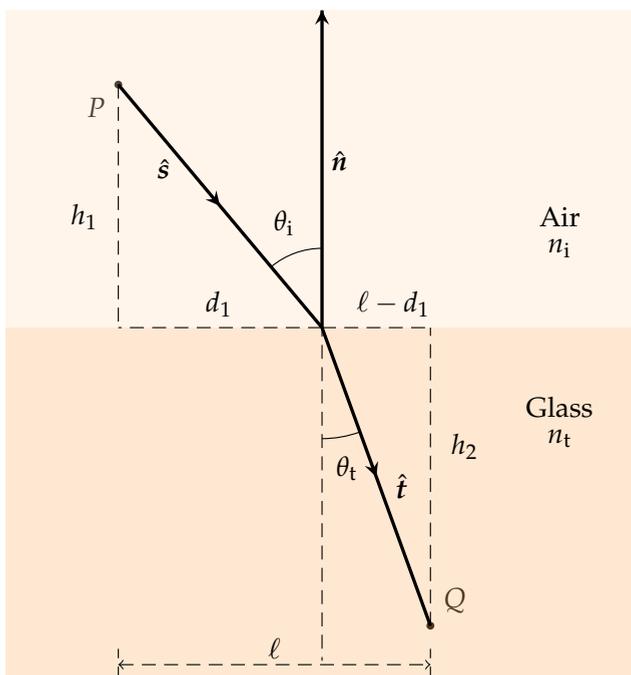


Figure 2.6: Illustration of the law of refraction at an optical interface separating air and glass. Refer back to Figure 2.4 to see the similarities and differences between the laws of reflection and refraction.

In Figure 2.6 a light ray originating from a point P with direction \hat{s} hits a horizontal refractive surface and changes direction to \hat{t} arriving at point Q . In geometrical optics, *the incident ray, refracted ray, and normal to the refractive surface lie in the same plane, called the plane of incidence, which can be explained from describing the system as a Hamiltonian system; we will derive this result in Section 2.6.*

The distance between P and Q is given by ℓ . The two rays have a common point being the incident point on the surface, here separating air and glass. The normal \hat{n} is perpendicular to the surface and directed towards the light source, i.e., $\hat{s} \cdot \hat{n} < 0$, and the angles of incidence and refraction satisfy $0 \leq \theta_i, \theta_t \leq \pi/2$, respectively. Using Figure 2.6 we can calculate the time required for the light to travel from P to Q , parametrized by d_1 , by calculating the length and dividing by the speed, i.e.,

$$t(d_1) = \frac{\sqrt{d_1^2 + h_1^2}}{c/n_i} + \frac{\sqrt{(\ell - d_1)^2 + h_2^2}}{c/n_t}. \quad (2.40)$$

Using the principle of least time we set

$$\frac{dt}{dd_1} = \frac{n_i d_1}{c\sqrt{d_1^2 + h_1^2}} - \frac{n_t(\ell - d_1)}{c\sqrt{(\ell - d_1)^2 + h_2^2}} = 0. \quad (2.41)$$

Note that we can verify that indeed $\frac{d^2 t}{dd_1^2}(d_1^*) > 0$, where d_1^* is the solution to (2.41), so that we compute a minimum. Equation (2.41) can be rewritten to

$$\frac{n_i d_1}{\sqrt{d_1^2 + h_1^2}} = \frac{n_t(\ell - d_1)}{\sqrt{(\ell - d_1)^2 + h_2^2}},$$

resulting in

$$n_i \sin \theta_i = n_t \sin \theta_t, \quad (2.42)$$

or

$$\frac{\sin \theta_i}{\sin \theta_t} = \frac{v_i}{v_t}, \quad (2.43)$$

with $v_i = \frac{c}{n_i}$ and $v_t = \frac{c}{n_t}$, i.e., *the ratios of the sine of the angle of incidence and the sine of angle of refraction is the ratio of the speeds of light in those media*. This is the scalar version of the law of refraction.

To derive the vectorial version of the law of refraction, we write the refracted ray as a linear combination of the incident ray and the normal, assuming they all lie in the same plane, i.e.,

$$\hat{t} = a \hat{s} + b \hat{n}, \quad (2.44)$$

for some real constants a, b . With $0 \leq \theta_i, \theta_t \leq \pi/2$ we have

$$\cos(\theta_i) = -\hat{s} \cdot \hat{n}, \quad \cos(\theta_t) = -\hat{t} \cdot \hat{n}. \quad (2.45)$$

Taking the inner product of (2.44) with \hat{n} gives

$$\hat{t} \cdot \hat{n} = a (\hat{s} \cdot \hat{n}) + b,$$

i.e.,

$$b = -a (\hat{s} \cdot \hat{n}) + \hat{t} \cdot \hat{n}. \quad (2.46)$$

Requiring \hat{t} to be a unit vector results in

$$1 = \hat{t} \cdot \hat{t} = a^2 + 2 a b (\hat{s} \cdot \hat{n}) + b^2. \quad (2.47)$$

Using (2.45), Snell's law and $\sin(\theta_i) = \sqrt{1 - (\hat{s} \cdot \hat{n})^2}$, $\sin(\theta_t) = \sqrt{1 - (\hat{t} \cdot \hat{n})^2}$, gives

$$n_i^2(1 - (\hat{s} \cdot \hat{n})^2) = n_t^2(1 - (\hat{t} \cdot \hat{n})^2). \quad (2.48)$$

Moreover, substituting (2.46) into (2.47) gives

$$1 - (\hat{t} \cdot \hat{n})^2 = a^2(1 - (\hat{s} \cdot \hat{n})^2). \quad (2.49)$$

From (2.46) and (2.49) we get

$$b = -a \hat{s} \cdot \hat{n} \pm \sqrt{1 - a^2 (1 - (\hat{s} \cdot \hat{n})^2)}. \quad (2.50)$$

From (2.48) and (2.49) we know that $a = \pm n_i/n_t$. With $a = \pm n_i/n_t$ there are now four possibilities for a and b . The only correct solution [66, p. 139] is

$$a = \frac{n_i}{n_t}, \quad b = -\frac{n_i}{n_t} \hat{s} \cdot \hat{n} - \sqrt{1 - \frac{n_i^2}{n_t^2} (1 - (\hat{s} \cdot \hat{n})^2)}, \quad (2.51)$$

since the other expressions each represent a reflection or refraction into one of the wrong quadrants bounded by the normal and surface tangent. The expression for \hat{t} from (2.44) becomes

$$\hat{t} = \frac{n_i}{n_t} \hat{s} - \left(\frac{n_i}{n_t} (\hat{s} \cdot \hat{n}) + \sqrt{1 - \frac{n_i^2}{n_t^2} (1 - (\hat{s} \cdot \hat{n})^2)} \right) \hat{n}. \quad (2.52)$$

The expression under the square root can be rewritten as $1 - \frac{n_i^2}{n_t^2} \sin^2(\theta_i)$ using (2.45). When this expression becomes negative, the refracted ray \hat{t} becomes complex, which is physically impossible. With $0 \leq \theta_i \leq \pi/2$ which results in $\sin(\theta_i) \geq 0$, we can say that refraction does not occur if

$$\sin(\theta_i) > \frac{n_t}{n_i}. \quad (2.53)$$

In this situation, the light ray is not refracted but reflected in the direction given by the law of reflection. This is called total internal reflection (TIR). Note that since $\sin(\theta_i) \leq 1$ always holds, the condition (2.53) is never satisfied if $n_i < n_t$. This means that TIR never occurs when a light ray moves to an optically denser medium, for instance, from air ($n \approx 1$) to glass ($n \approx 1.5$). If $\sin(\theta_i) = n_t/n_i$ for an incoming ray, we can verify that $\theta_t = \frac{1}{2}\pi$, $\hat{\mathbf{t}} \cdot \hat{\mathbf{n}} = 0$, and the outgoing ray lies in the tangent plane to the surface at the incident point. The angle $\theta_c = \theta_i$ at which this happens is called the *critical angle*.

2.6 Hamiltonian optics

The Irish mathematician *William Rowan Hamilton*, whose formulations in statistical and quantum mechanics I spent hours studying during my postgraduate mathematical physics courses, introduced the concept of a *Hamiltonian system*. A Hamiltonian system is a mathematical formalism to describe the evolution of a physical system.

In Hamiltonian mechanics, a classic physical system is described by a set of canonical coordinates (\mathbf{q}, \mathbf{p}) . The coordinates \mathbf{q} are called *generalized coordinates*, and are chosen so as to eliminate the constraints or to take advantage of the symmetries of the problem, and \mathbf{p} are their *conjugate momenta*.

Hamilton also contributed to the field of geometrical optics. In a series of classic papers [75–77], Hamilton introduced systems of canonical equations to describe the propagation of rays through an optical system, both in general terms and for specific cases. In order to give algebraic expressions for reflection and refraction of light rays in an optical system, Hamilton introduced four *characteristic functions*: the point characteristic V , the mixed characteristic of the first kind W , the mixed characteristic of the second kind W^* and the angular characteristic T . The characteristics are measures of the optical path length between a specified source and target point [12, 99].

We consider a light ray that travels between two planes of reference in \mathbb{R}^3 . Without loss of generality, we choose these planes parallel to the xy -plane in Euclidean space. The z -axis is often called the *optical axis* in this configuration.

We parametrize a light ray in \mathbb{R}^3 as: $\mathbf{x} = \mathbf{x}(z) = (q_1(z), q_2(z), z)$. The projection on a plane $z = \text{constant}$ is written as a vector $\mathbf{q}(z) = (q_1(z), q_2(z))$ in \mathbb{R}^2 . Using this parametrization, the optical path length (2.24) from a plane $z = z_s$ to a plane $z = z_t$ can be written as a functional of \mathbf{q} only, i.e.,

$$V[\mathbf{q}] = \int_{z_s}^{z_t} n(\mathbf{q}, z) \sqrt{1 + |\mathbf{q}'|^2} dz, \quad (2.54)$$

where we denote $\mathbf{q}' = \frac{d\mathbf{q}}{dz}$. The Euler-Lagrange equations of this functional are

$$\frac{d}{dz} \left(\frac{n\mathbf{q}'}{\sqrt{1+|\mathbf{q}'|^2}} \right) - \sqrt{1+|\mathbf{q}'|^2} \frac{\partial n}{\partial \mathbf{q}} = \mathbf{0}. \quad (2.55)$$

We can reformulate the Euler-Lagrange equations as a first-order system of ODEs of a Hamiltonian system using a series of steps.

To this end we start by introducing the *direction cosines*

$$\cos(\alpha_1) = \frac{q'_1}{\sqrt{1+|\mathbf{q}'|^2}}, \quad \cos(\alpha_2) = \frac{q'_2}{\sqrt{1+|\mathbf{q}'|^2}}, \quad \cos(\beta) = \frac{\sigma}{\sqrt{1+|\mathbf{q}'|^2}}, \quad (2.56)$$

where the parameter $\sigma = -1$ represents backward propagation with respect to the z -direction, $\sigma = 0$ represents marginal propagation perpendicular to the z -axis, and $\sigma = 1$ represents forward propagation. From now on we only consider forward propagation, i.e., $\sigma = 1$. The *optical direction cosines* are defined as

$$p_1 = n \cos(\alpha_1) = \frac{n q'_1}{\sqrt{1+|\mathbf{q}'|^2}}, \quad p_2 = n \cos(\alpha_2) = \frac{n q'_2}{\sqrt{1+|\mathbf{q}'|^2}}. \quad (2.57)$$

The optical direction cosines are drawn schematically in Figure 2.7 in Euclidean space.

Straightforward evaluation of (2.57) using $\mathbf{p}(z) = (p_1(z), p_2(z))$ results in the relation between \mathbf{p} and \mathbf{q}'

$$\sqrt{1+|\mathbf{q}'|^2} = \frac{n}{\sqrt{n^2 - |\mathbf{p}|^2}}. \quad (2.58)$$

Combining this relation with (2.55) and (2.57), using $n \geq 1$, results in the coupled ODE system

$$q'_1 = \frac{p_1}{\sqrt{n^2 - |\mathbf{p}|^2}}, \quad q'_2 = \frac{p_2}{\sqrt{n^2 - |\mathbf{p}|^2}}, \quad (2.59a)$$

$$p'_1 = \frac{n}{\sqrt{n^2 - |\mathbf{p}|^2}} \frac{\partial n}{\partial q_1}, \quad p'_2 = \frac{n}{\sqrt{n^2 - |\mathbf{p}|^2}} \frac{\partial n}{\partial q_2}. \quad (2.59b)$$

The *Hamiltonian* is defined as

$$H = H(z, \mathbf{q}, \mathbf{p}) = -n \cos(\beta) = -\sqrt{n^2 - |\mathbf{p}|^2}. \quad (2.60)$$

We can interpret this function as the projection on the z -axis of the unit tangent vector along a ray, multiplied by $-n$.

Writing (2.57) as

$$\mathbf{p} = \frac{n}{\sqrt{1 + |\mathbf{q}'|^2}} \frac{d\mathbf{q}}{dz} = n \frac{d\mathbf{q}}{ds}, \quad (2.61)$$

and defining $\mathbf{p}^* = (\mathbf{p}, p_3)$, where

$$p_3 = n \cos(\beta) = \frac{n}{\sqrt{1 + |\mathbf{q}'|^2}} = n \frac{dz}{ds}, \quad (2.62)$$

gives

$$\mathbf{p}^* = (\mathbf{p}, p_3) = n \left(\frac{d\mathbf{q}}{ds}, \frac{dz}{ds} \right) = n \frac{d\mathbf{x}}{ds}, \quad (2.63)$$

cf. (2.21a). We will refer to either \mathbf{p} or \mathbf{p}^* as the *momentum*. The direction cosine vector \mathbf{p}^* is always confined to the so-called *Descartes' sphere*, meaning $|\mathbf{p}^*| = n$ [156, p. 9]. The directional derivative $\frac{dp^*}{ds}$ is equal to ∇n . The angles α_1, α_2 , and β are the angles between the unit tangent vector along the ray and the x -axis, y -axis and z -axis, respectively. (Here, the ray is simply drawn as a straight line to illustrate the angles, so the tangent is in the direction of the ray.) Figure 2.8 shows a projection of the ray on the plane spanned by the ray and the z -axis, with the z -axis now drawn as the vertical axis. The position and momentum change as a ray moves in the positive z -direction.

Using the Hamiltonian, we can write down a *Hamiltonian system*

$$\mathbf{q}' = \frac{\partial H}{\partial \mathbf{p}} = \frac{\mathbf{p}}{\sqrt{n^2 - |\mathbf{p}|^2}}, \quad \mathbf{p}' = -\frac{\partial H}{\partial \mathbf{q}} = \frac{n}{\sqrt{n^2 - |\mathbf{p}|^2}} \frac{\partial n}{\partial \mathbf{q}}, \quad (2.64)$$

for the change of \mathbf{q} and \mathbf{p} in the plane perpendicular to the optical axis as the ray $x(z)$ moves along the optical axis. We can rewrite the optical path length in (2.54) to

$$V[\mathbf{q}] = \int_{z_s}^{z_t} \frac{n^2}{\sqrt{n^2 - |\mathbf{p}|^2}} dz = \int_{z_s}^{z_t} (\mathbf{q}' \cdot \mathbf{p} - H(z, \mathbf{q}, \mathbf{p})) dz. \quad (2.65)$$

In general, we can write the optical path length as the functional

$$V[\mathbf{q}] = \int_{z_s}^{z_t} L(z, \mathbf{q}, \mathbf{q}') dz, \quad (2.66)$$

with Lagrangian $L(z, \mathbf{q}, \mathbf{q}') := \mathbf{q}' \cdot \mathbf{p} - H(z, \mathbf{q}, \mathbf{p})$ and with associated momenta

$$\mathbf{p} = \frac{\partial L}{\partial \mathbf{q}'}(z, \mathbf{q}, \mathbf{q}'). \quad (2.67)$$

For fixed z and \mathbf{q} , we can invert relation (2.67) to obtain $\mathbf{q}' = \mathbf{q}'(z, \mathbf{q}, \mathbf{p})$ provided

$$\det \left(\frac{\partial^2 L}{\partial q'_i \partial q'_j} \right) \neq 0, \quad (2.68)$$

by the implicit function theorem. With the Hamiltonian in (2.65) written as

$$H = H(z, \mathbf{q}, \mathbf{p}) = \mathbf{q}' \cdot \mathbf{p} - L(z, \mathbf{q}, \mathbf{q}'), \quad (2.69)$$

we can derive that

$$\mathbf{p} = \frac{\partial L}{\partial \mathbf{q}'} \left(z, \mathbf{q}, \frac{\partial H}{\partial \mathbf{p}} \right), \quad \mathbf{q}' = \frac{\partial H}{\partial \mathbf{p}} \left(z, \mathbf{q}, \frac{\partial L}{\partial \mathbf{q}'} \right), \quad (2.70)$$

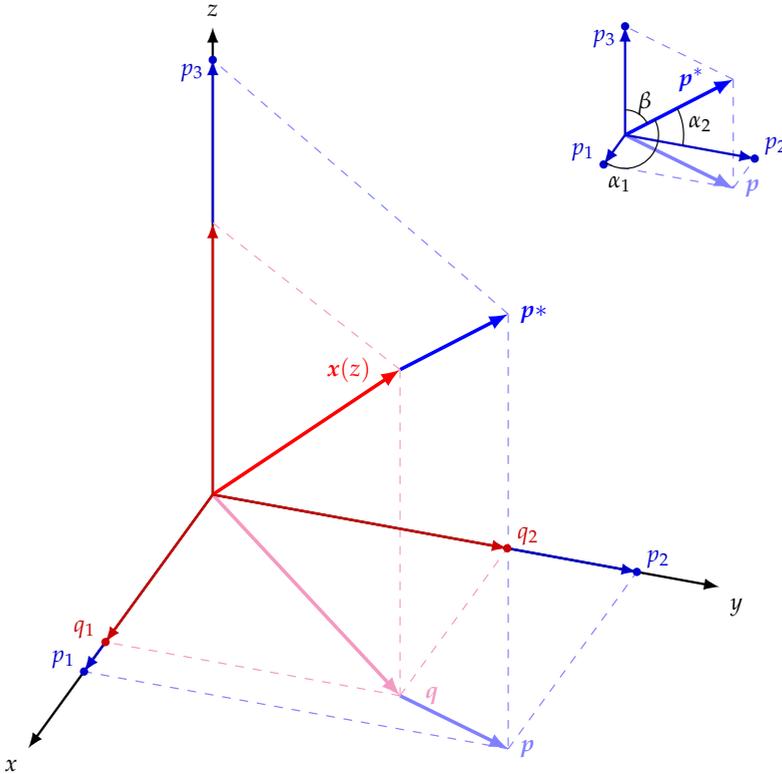


Figure 2.7: Illustration of the ray $\mathbf{x}(z) = (\mathbf{q}(z), z)$ (red) and momentum $\mathbf{p}^*(z) = (\mathbf{p}(z), p_3)$ (blue) with $\mathbf{q} = (x, y)$. The optical direction cosines α_1, α_2 and β are drawn in the upper right corner as the angles between the components of $\mathbf{p}^* = (p_1, p_2, p_3)$ and the Cartesian coordinate axes.

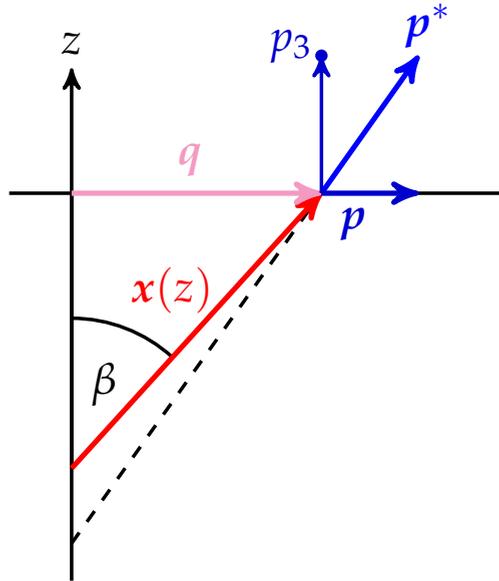


Figure 2.8: The ray $x(z) = (q(z), z)$ on the plane spanned by $x(z)$ and the z -axis. (Note that the vectors p and p^* can still be directed out of the page or into the page, since p and q do not necessarily span the same plane.)

i.e., $\frac{\partial L}{\partial q'}$ and $\frac{\partial H}{\partial p}$ are inverse functions of each other, for fixed q and z . The transformation from the function $L(z, q, q')$ and the coordinates q, q' to $H(z, q, p)$ and q, p is referred to as a Legendre transformation [32, p. 32].

The manifold of all positions q and momenta p is called the *phase space*. Hamiltonian systems conserve phase space volume, also known by the term *étendue*. Any volume element will have constant volume as q and p change according to Hamilton's equations, as shown by my fellow group member Bart van Lith [93] using arguments from fluid dynamics.

Before we define Hamilton's characteristics, we end this section by noting that the law of reflection and Snell's law can also be derived using the Hamiltonian system. We briefly discuss Snell's law below.

Suppose a ray hits an interface between two dielectric isotropic media, where the refractive index jumps from n_i to n_t . The interface has local unit normal $\hat{n} = -\hat{e}_z$. We choose the z -axis in the opposite direction of the normal, starting at the point where the ray hits the interface, so $q = 0$, as illustrated in Figure 2.9. Using the Hamiltonian system (2.64) we see that $\frac{\partial n}{\partial q} = 0$ implying

that $\frac{d\mathbf{p}}{dz} = \mathbf{0}$ with this coordinate system. Hence, the tangential momentum \mathbf{p} is conserved across the interface. We can write the incident momentum as $\mathbf{p}_i^* = \mathbf{p}_i - p_{i,3} \hat{\mathbf{n}}$ and the transmitted momentum as $\mathbf{p}_t^* = \mathbf{p}_t - p_{t,3} \hat{\mathbf{n}}$. From Figure 2.6 we can see that $\mathbf{p}_i^* = n_i \hat{\mathbf{s}}$ and $\mathbf{p}_t^* = n_t \hat{\mathbf{t}}$. A change in momentum can only occur perpendicular to the interface, which implies that \mathbf{p}_i^* must lie in the plane spanned by the incident momentum \mathbf{p}_i^* and the normal $\hat{\mathbf{n}}$. This plane is called the *plane of incidence*.

Since the change in momentum is in the same direction as the normal, we have

$$\hat{\mathbf{n}} = \pm \frac{\mathbf{p}_i^* - \mathbf{p}_t^*}{|\mathbf{p}_i^* - \mathbf{p}_t^*|}. \quad (2.71)$$

Assuming forward propagation of the ray and the normal $\hat{\mathbf{n}}$ directed towards the light source, the sign is such that $\mathbf{p}_i^* \cdot \hat{\mathbf{n}} \leq 0$. Taking the cross product of (2.71) with $\hat{\mathbf{n}}$ leads to another form of Snell's law, i.e.,

$$\mathbf{p}_i^* \times \hat{\mathbf{n}} = \mathbf{p}_t^* \times \hat{\mathbf{n}}, \quad (2.72)$$

and using a property of cross products we know that

$$|\mathbf{p}_i^*| \sin(\theta_i) = |\mathbf{p}_t^*| \sin(\theta_t), \quad (2.73)$$

with $0 \leq \theta_i, \theta_t \leq \pi/2$ the angles of incidence and refraction, respectively, and since \mathbf{p}_i^* and \mathbf{p}_t^* both lie on a Descartes' sphere with radii n_i and n_t , we obtain

$$n_i \sin(\theta_i) = n_t \sin(\theta_t). \quad (2.74)$$

Considering the critical angle and the TIR condition in (2.53) in Section 2.5, the law of reflection actually follows immediately from Snell's law. Rays with $\theta_i > \theta_c$ are all reflected back at the interface. We can think of the transmitted medium for these rays being just the incident medium and $n_t = n_i$. This leads to $\sin(\theta_i) = \sin(\theta_r)$, i.e., $\theta_i = \theta_r$. Using the same reasoning as for the transmitted momentum above, we can derive that the reflected ray in Section 2.4 with $0 \leq \theta_i, \theta_r \leq \pi/2$ also lies in the plane of incidence.

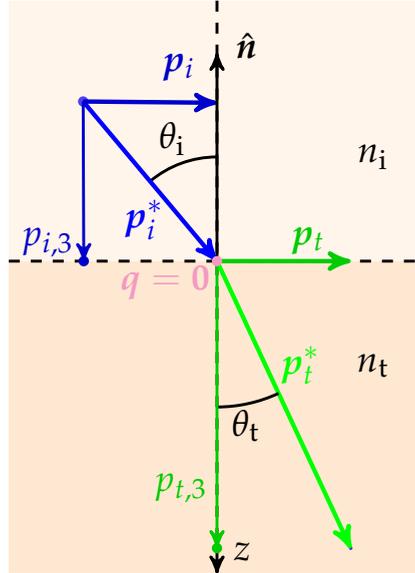


Figure 2.9: Illustration of an alternative derivation of Snell's law. The z -axis is in the opposite direction of the normal \hat{n} and starts at the point where the ray hits the interface, so $q = 0$. The tangential component p_i is conserved across the interface, i.e., $p_i = p_t$.

2.7 Hamilton's characteristic functions

In the following, we introduce the mathematical theory of Hamilton's characteristic functions V , W , W^* and T in Section 2.7.1 – 2.7.4, respectively. In Section 2.7.5, we give a geometrical interpretation to each characteristic function.

The main result of Hamilton's theory is that a light ray from a source plane $z = z_s$ to a target plane $z = z_t$ can be completely specified by choosing two out of the four local position and optical direction coordinates $\{q_s, q_t, p_s, p_t\}$.

We consider an optical system in the vicinity of the planes $z = z_s$ (source plane) and $z = z_t > z_s$ (target plane). For instance, a lens is positioned between two planes. A light source is located in the source plane and emitting a bundle of light, which propagates through the optical system. We would like to trace a typical ray from the source plane to the target plane. More precisely, we would like to determine the position q or direction coordinates p on both planes. The propagation of a light ray through the optical system, starting

from $z = z_s$ is governed by the Hamiltonian initial value problem

$$q' = \frac{\partial H}{\partial p}, \quad p' = -\frac{\partial H}{\partial q}, \quad z_s < z < z_t, \quad (2.75a)$$

$$q(z_s) = q_s, \quad p(z_s) = p_s. \quad (2.75b)$$

The solution depends on the initial conditions and we formally write

$$q = q(z; z_s, q_s, p_s), \quad p = p(z; z_s, q_s, p_s). \quad (2.76)$$

The solution at the target reads

$$q_t = q(z_t; z_s, q_s, p_s), \quad p_t = p(z_t; z_s, q_s, p_s). \quad (2.77)$$

Thus, the position and direction coordinates at the target plane are uniquely determined by the initial condition (2.75b). Assuming that

$$\det \left(\frac{\partial q_t}{\partial p_s} \right) \neq 0, \quad (2.78)$$

the implicit function theorem states that, in theory, we can compute p_s from the first equation in (2.77) as a function of the other variables and we can write $p_s = p_s(z_s, z_t, q_s, q_t)$. Substituting this into (2.76) we get

$$q = q(z; z_s, z_t, q_s, q_t), \quad p = p(z; z_s, z_t, q_s, q_t), \quad (2.79)$$

which represents a light ray that starts at the point (q_s, z_s) in the source plane and arrives at (q_t, z_t) in the target plane. Note that we have used the same notation for two different representations of the solution. Alternatively, we can define q and p as solutions to the boundary value problem

$$q' = \frac{\partial H}{\partial p}, \quad p' = -\frac{\partial H}{\partial q}, \quad z_s < z < z_t, \quad (2.80a)$$

$$q(z_s) = q_s, \quad q(z_t) = q_t, \quad (2.80b)$$

where, instead of specifying q and p at the source plane, we now specify q both at the source and target plane. However, we should note that a boundary value problems often has no solution or infinitely many, whereas an initial value problem has a unique solution under mild conditions.

Below we will introduce four characteristics, each dependent on different coordinates, one from the source and one from the target.

2.7.1 The point characteristic

The optical path length expressed in terms of \mathbf{q}_s and \mathbf{q}_t , which are assumed to be fixed, is also called the *point characteristic function*. Recall the optical path length is given by the functional

$$V[\mathbf{q}] = \int_{z_s}^{z_t} L(z, \mathbf{q}, \mathbf{q}') dz, \quad (2.81)$$

with $L(z, \mathbf{q}, \mathbf{q}') = \mathbf{q}' \cdot \mathbf{p} - H(z, \mathbf{q}, \mathbf{p})$ as in (2.66) and $\mathbf{q}' = \frac{d\mathbf{q}}{dz}$. Substituting (2.79) we can write V as the function $V = V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t)$. We can write the momenta \mathbf{p}_s and \mathbf{p}_t in terms of derivatives of V . Taking the partial derivative of $L(z, \mathbf{q}, \mathbf{q}')$ with respect to $c \in \{q_{s,1}, q_{s,2}, q_{t,1}, q_{t,2}\}$ gives

$$\frac{\partial L}{\partial c} = \frac{\partial \mathbf{q}'}{\partial c} \cdot \mathbf{p} + \mathbf{q}' \cdot \frac{\partial \mathbf{p}}{\partial c} - \frac{\partial H}{\partial \mathbf{q}} \cdot \frac{\partial \mathbf{q}}{\partial c} - \frac{\partial H}{\partial \mathbf{p}} \cdot \frac{\partial \mathbf{p}}{\partial c}, \quad (2.82a)$$

and using the Hamiltonian equations (2.64), this is equal to

$$\frac{\partial L}{\partial c} = \frac{\partial \mathbf{q}'}{\partial c} \cdot \mathbf{p} + \mathbf{p}' \cdot \frac{\partial \mathbf{q}}{\partial c} = \left(\frac{\partial \mathbf{q}}{\partial c} \cdot \mathbf{p} \right)'. \quad (2.82b)$$

Taking the partial derivative of V with respect to $c \in \{q_{s,1}, q_{s,2}, p_{t,1}, p_{t,2}\}$ results in

$$\frac{\partial V}{\partial c} = \int_{z_s}^{z_t} \frac{\partial L}{\partial c} dz = \left[\frac{\partial \mathbf{q}}{\partial c} \cdot \mathbf{p} \right]_{z_s}^{z_t}. \quad (2.83)$$

Comparing this expression with the differentiation rule

$$\frac{\partial V}{\partial c} = \frac{\partial V}{\partial \mathbf{q}_s} \cdot \frac{\partial \mathbf{q}_s}{\partial c} + \frac{\partial V}{\partial \mathbf{q}_t} \cdot \frac{\partial \mathbf{q}_t}{\partial c}, \quad (2.84)$$

and using that $\mathbf{q}(z_s) = \mathbf{q}_s$ and $\mathbf{q}(z_t) = \mathbf{q}_t$ we obtain

$$\mathbf{p}_s = -\frac{\partial V}{\partial \mathbf{q}_s}, \quad \mathbf{p}_t = \frac{\partial V}{\partial \mathbf{q}_t}, \quad (2.85)$$

i.e., the momenta at the source and target can be derived from the point characteristic connecting (\mathbf{q}_s, z_s) and (\mathbf{q}_t, z_t) .

Hamilton's point characteristic is defined as the optical path length

$$V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t) = \int_{z_s}^{z_t} L(z, \mathbf{q}, \mathbf{q}') dz = \int_{z_s}^{z_t} \mathbf{q}' \cdot \mathbf{p} - H(z, \mathbf{q}, \mathbf{p}) dz, \quad (2.86)$$

with $L(z, \mathbf{q}, \mathbf{q}') = n(\mathbf{q}, z) \sqrt{1 + |\mathbf{q}'|^2}$ and $H(z, \mathbf{q}, \mathbf{p}) = -\sqrt{n^2 - |\mathbf{p}|^2}$ between a point (\mathbf{q}_s, z_s) on a source plane $z = z_s$ and a target coordinate (\mathbf{q}_t, z_t) on a target plane $z = z_t$. The momenta on the source and target planes are

$$\mathbf{p}_s = -\frac{\partial V}{\partial \mathbf{q}_s}, \quad \mathbf{p}_t = \frac{\partial V}{\partial \mathbf{q}_t}. \quad (2.87)$$

We will now show that V satisfies the eikonal equation (2.15). That is no surprise considering that $|\nabla \phi|$ is a measure of the path length normal to the wavefront, which is in the direction of propagation of the light ray as in (2.17). We can show that V satisfies the eikonal equation by expressing the partial derivatives of V with respect to z_s and z_t in terms of H . Consider

$$\begin{aligned} \frac{\partial V}{\partial z_t} &= L(z_t, \mathbf{q}_t, \mathbf{q}'_t) + \int_{z_s}^{z_t} \frac{\partial L}{\partial z_t}(z, \mathbf{q}, \mathbf{q}') dz \\ &= \mathbf{q}'_t \cdot \mathbf{p}_t - H(z, \mathbf{q}_t, \mathbf{p}_t) + \left[\frac{\partial \mathbf{q}}{\partial z_t} \cdot \mathbf{p} \right]_{z_s}^{z_t} \\ &= -H(z, \mathbf{q}_t, \mathbf{p}_t) + \mathbf{p}_t \cdot \left(\mathbf{q}'_t + \frac{\partial \mathbf{q}_t}{\partial z_t} \right) - \mathbf{p}_s \cdot \frac{\partial \mathbf{q}_s}{\partial z_t}, \end{aligned} \quad (2.88)$$

where $\mathbf{q}'_t = \mathbf{q}'(z_t; z_s, z_t, \mathbf{q}_s, \mathbf{q}_t)$ and we use (2.83) with $c = z_t$ to derive the second equality. Since $\mathbf{q}_s = \mathbf{q}(z_s; z_s, z_t, \mathbf{q}_s, \mathbf{q}_t)$ and $\mathbf{q}_t = \mathbf{q}(z_t; z_s, z_t, \mathbf{q}_s, \mathbf{q}_t)$, cf. (2.79), are given fixed points, differentiating with respect to z_t gives

$$\frac{\partial \mathbf{q}}{\partial z_t} \Big|_{z=z_s} = \frac{\partial \mathbf{q}_s}{\partial z_t} = \mathbf{0}, \quad \left(\mathbf{q}' + \frac{\partial \mathbf{q}}{\partial z_t} \right) \Big|_{z=z_t} = \mathbf{q}'_t + \frac{\partial \mathbf{q}_t}{\partial z_t} = \mathbf{0}, \quad (2.89)$$

where the second equation follows from taking the derivative with respect to the variable $z = z_t$ and parameter z_t . Using (2.89) we see that the inner products in the last line of (2.88) vanish. Hence, we get

$$\frac{\partial V}{\partial z_t} = -H(z_t, \mathbf{q}_t, \mathbf{p}_t), \quad (2.90)$$

with the Hamiltonian (2.60). Combining this equation with (2.87) results in

$$\left(\frac{\partial V}{\partial z_t} \right)^2 + \left| \frac{\partial V}{\partial \mathbf{q}_t} \right|^2 = n(\mathbf{q}_t, z_t)^2. \quad (2.91)$$

This is a special case of the eikonal equation, since using $\mathbf{p}^* = \nabla \varphi$ as in (2.20), we have that $\mathbf{p} = \frac{\partial \varphi}{\partial \mathbf{q}}$ and the eikonal equation (2.15) can be rewritten (in quadratic form) as

$$\left| \frac{\partial \varphi}{\partial z} \right|^2 + \left| \frac{\partial \varphi}{\partial \mathbf{q}} \right|^2 = n^2. \quad (2.92)$$

V satisfies the eikonal equation formulated in terms of the target coordinates \mathbf{q}_t and z_t . From (2.90) it follows that $\frac{\partial V}{\partial z_t} > 0$, which means that the optical path length increases if the target plane is moving away from the source plane.

We can repeat the above for the partial derivative $\frac{\partial V}{\partial z_s}$ and obtain

$$\frac{\partial V}{\partial z_s} = H(z_s, \mathbf{q}_s, \mathbf{p}_s), \quad (2.93)$$

and the eikonal equation

$$\left(\frac{\partial V}{\partial z_s} \right)^2 + \left| \frac{\partial V}{\partial \mathbf{q}_s} \right|^2 = n(\mathbf{q}_s, z_s)^2, \quad (2.94)$$

but now we obtain $\frac{\partial V}{\partial z_s} < 0$, i.e., the optical path length decreases if the source plane is moving towards the target plane.

2.7.2 The mixed characteristic of the first kind

In this section, we introduce the mixed characteristic function W , which is a function of the position coordinate \mathbf{q}_s and direction coordinate \mathbf{p}_t . We start by expressing the solution of the Hamiltonian system (2.75) in terms of \mathbf{q}_s and \mathbf{p}_t , for which we have to eliminate \mathbf{p}_s from the second relation in (2.77). To use the implicit function theorem we require

$$\det \left(\frac{\partial \mathbf{p}_t}{\partial \mathbf{p}_s} \right) \neq 0, \quad (2.95)$$

and we can write $\mathbf{p}_s = \mathbf{p}_s(z_s, z_t, \mathbf{q}_s, \mathbf{p}_t)$. Substituting this relation into (2.76) gives

$$\mathbf{q} = \mathbf{q}(z; z_s, z_t, \mathbf{q}_s, \mathbf{p}_t), \quad \mathbf{p} = \mathbf{p}(z; z_s, z_t, \mathbf{q}_s, \mathbf{p}_t), \quad (2.96)$$

which represents a light ray originating from (\mathbf{q}_s, z_s) and arriving at the target plane $z = z_t$ with direction \mathbf{p}_t . This solution is also a solution to the boundary value problem

$$\mathbf{q}' = \frac{\partial H}{\partial \mathbf{p}}, \quad \mathbf{p} = -\frac{\partial H}{\partial \mathbf{q}}, \quad z_s < z < z_t, \quad (2.97a)$$

$$\mathbf{q}(z_s) = \mathbf{q}_s, \quad \mathbf{p}(z_t) = \mathbf{p}_t. \quad (2.97b)$$

Next, we introduce the *mixed characteristic function* W , defined as

$$W = W(z_s, z_t, \mathbf{q}_s, \mathbf{p}_t) = V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t) - \mathbf{q}_t \cdot \mathbf{p}_t, \quad (2.98)$$

where $\mathbf{q}_t = \mathbf{q}_t(z_s, z_t, \mathbf{q}_s, \mathbf{p}_t)$. We have that

$$\frac{\partial W}{\partial \mathbf{p}_s} = \mathbf{0}, \quad \frac{\partial W}{\partial \mathbf{q}_t} = \mathbf{0}, \quad (2.99)$$

by using the right-hand side of the definition (2.98) for the first equation and using (2.87) for the second equation, so that indeed $W = W(z_s, z_t, \mathbf{q}_s, \mathbf{p}_t)$ does not depend on \mathbf{p}_s and \mathbf{q}_t . Taking the partial derivatives of W with respect to $c \in \{q_{s,1}, q_{s,2}, p_{t,1}, p_{t,2}\}$ gives

$$\begin{aligned} \frac{\partial W}{\partial c} &= \frac{\partial V}{\partial c} - \frac{\partial \mathbf{q}_t}{\partial c} \cdot \mathbf{p}_t - \mathbf{q}_t \cdot \frac{\partial \mathbf{p}_t}{\partial c} \\ &= \left[\frac{\partial \mathbf{q}}{\partial c} \cdot \mathbf{p} \right]_{z_s}^{z_t} - \frac{\partial \mathbf{q}_t}{\partial c} \cdot \mathbf{p}_t - \mathbf{q}_t \cdot \frac{\partial \mathbf{p}_t}{\partial c} \\ &= -\frac{\partial \mathbf{q}_s}{\partial c} \cdot \mathbf{p}_s - \mathbf{q}_t \cdot \frac{\partial \mathbf{p}_t}{\partial c}, \end{aligned} \quad (2.100)$$

where we use (2.83) to derive the second equality. Comparing this with

$$\frac{\partial W}{\partial c} = \frac{\partial W}{\partial \mathbf{q}_s} \cdot \frac{\partial \mathbf{q}_s}{\partial c} + \frac{\partial W}{\partial \mathbf{p}_t} \cdot \frac{\partial \mathbf{p}_t}{\partial c}, \quad (2.101)$$

we obtain

$$\mathbf{p}_s = -\frac{\partial W}{\partial \mathbf{q}_s}, \quad \mathbf{q}_t = -\frac{\partial W}{\partial \mathbf{p}_t}. \quad (2.102)$$

Combining the latter equation with the second equation in (2.87) we can also see that

$$\mathbf{q}_t = -\frac{\partial W}{\partial \mathbf{p}_t} \left(z_s, z_t, \mathbf{q}_s, \frac{\partial V}{\partial \mathbf{q}_t} \right), \quad \mathbf{p}_t = \frac{\partial V}{\partial \mathbf{q}_t} \left(z_s, z_t, \mathbf{q}_s, -\frac{\partial W}{\partial \mathbf{p}_t} \right), \quad (2.103)$$

i.e., for a given z_s, z_t and \mathbf{q}_s we have that $\frac{\partial V}{\partial \mathbf{q}_t}$ and $-\frac{\partial W}{\partial \mathbf{p}_t}$ are each other's inverses. Hence, the definition of W in (2.98) is a Legendre transformation from $V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t)$ to $W(z_s, z_t, \mathbf{q}_s, \mathbf{p}_t)$.

Similarly to V in the previous section, we can show that W satisfies the same eikonal equations since $\frac{\partial V}{\partial z_s} = \frac{\partial W}{\partial z_s}$ and $\frac{\partial V}{\partial z_t} = \frac{\partial W}{\partial z_t}$. The mixed characteristic satisfies

$$\frac{\partial W}{\partial z_s} = H(z_s, \mathbf{q}_s, \mathbf{p}_s), \quad \frac{\partial W}{\partial z_t} = -H(z_t, \mathbf{q}_t, \mathbf{p}_t), \quad (2.104)$$

and the eikonal equations are

$$\left(\frac{\partial W}{\partial z_s}\right)^2 + \left|\frac{\partial W}{\partial \mathbf{q}_s}\right|^2 = n(\mathbf{q}_s, z_s)^2, \quad \frac{\partial W}{\partial z_s} < 0, \quad (2.105a)$$

$$\left(\frac{\partial W}{\partial z_t}\right)^2 + |\mathbf{p}_t|^2 = n(\mathbf{q}_t, z_t)^2, \quad \frac{\partial W}{\partial z_t} > 0. \quad (2.105b)$$

Hamilton's mixed characteristic W is defined as the Legendre transformation

$$W(z_s, z_t, \mathbf{q}_s, \mathbf{p}_t) = V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t) - \mathbf{q}_t \cdot \mathbf{p}_t. \quad (2.106)$$

The momentum on the source plane and position on the target plane are

$$\mathbf{p}_s = -\frac{\partial W}{\partial \mathbf{q}_s}, \quad \mathbf{q}_t = -\frac{\partial W}{\partial \mathbf{p}_t}. \quad (2.107)$$

2.7.3 The mixed characteristic of the second kind

There is a variant to Hamilton's mixed characteristic called W^* , which is a function of the direction coordinate \mathbf{p}_s and position coordinate \mathbf{q}_t . This characteristic function is exactly the same as W if we flip the source and target planes and the ray travels from $z = z_t$ to $z = z_s$ in the opposite direction. It is defined as

$$W^*(z_s, z_t, \mathbf{p}_s, \mathbf{q}_t) = V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t) + \mathbf{q}_s \cdot \mathbf{p}_s, \quad (2.108)$$

where $\mathbf{q}_s = \mathbf{q}_s(z_s, z_t, \mathbf{p}_s, \mathbf{q}_t)$. We have that

$$\frac{\partial W^*}{\partial \mathbf{q}_s} = \mathbf{0}, \quad \frac{\partial W^*}{\partial \mathbf{p}_t} = \mathbf{0}, \quad (2.109)$$

using (2.87) for the first equation, so that indeed $W^* = W^*(z_s, z_t, \mathbf{p}_s, \mathbf{q}_t)$ does not depend on \mathbf{q}_s and \mathbf{p}_t . We can again take the partial derivatives and derive eikonal equations. The derivations are completely analogous to the ones presented above for the point and mixed characteristics and we will only present the results here. For the mixed characteristic W^* we obtain

$$\mathbf{q}_s = \frac{\partial W^*}{\partial \mathbf{p}_s}, \quad \mathbf{p}_t = \frac{\partial W^*}{\partial \mathbf{q}_t}. \quad (2.110)$$

Moreover, for a given z_s, z_t and \mathbf{q}_t we have that $-\frac{\partial V}{\partial \mathbf{q}_s}$ and $\frac{\partial W^*}{\partial \mathbf{p}_s}$ are each other's inverses, since

$$\mathbf{q}_s = \frac{\partial W^*}{\partial \mathbf{p}_s} \left(z_s, z_t, -\frac{\partial V}{\partial \mathbf{q}_s}, \mathbf{q}_t \right), \quad \mathbf{p}_s = -\frac{\partial V}{\partial \mathbf{q}_s} \left(z_s, z_t, \frac{\partial W^*}{\partial \mathbf{p}_s}, \mathbf{q}_t \right), \quad (2.111)$$

i.e., the definition of W^* in (2.108) is a Legendre transformation from the function $V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t)$ to $W^*(z_s, z_t, \mathbf{p}_s, \mathbf{q}_t)$.

The mixed characteristic W^* satisfies

$$\frac{\partial W^*}{\partial z_s} = H(z_s, \mathbf{q}_s, \mathbf{p}_s), \quad \frac{\partial W^*}{\partial z_t} = -H(z_t, \mathbf{q}_t, \mathbf{p}_t), \quad (2.112)$$

and the corresponding eikonal equations are

$$\left(\frac{\partial W^*}{\partial z_s} \right)^2 + |\mathbf{p}_s|^2 = n(\mathbf{q}_s, z_s)^2, \quad \frac{\partial W^*}{\partial z_s} < 0, \quad (2.113a)$$

$$\left(\frac{\partial W^*}{\partial z_t} \right)^2 + \left| \frac{\partial W^*}{\partial \mathbf{q}_t} \right|^2 = n(\mathbf{q}_t, z_t)^2, \quad \frac{\partial W^*}{\partial z_t} > 0. \quad (2.113b)$$

Hamilton's mixed characteristic W^ is defined as the Legendre transformation*

$$W^*(z_s, z_t, \mathbf{p}_s, \mathbf{q}_t) = V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t) + \mathbf{q}_s \cdot \mathbf{p}_s. \quad (2.114)$$

The position on the source plane and momentum on the target plane can be derived as

$$\mathbf{q}_s = \frac{\partial W^*}{\partial \mathbf{p}_s}, \quad \mathbf{p}_t = \frac{\partial W^*}{\partial \mathbf{q}_t}. \quad (2.115)$$

2.7.4 The angular characteristic

Lastly, we present the *angular characteristic function* T , defined as

$$T = T(z_s, z_t, \mathbf{p}_s, \mathbf{p}_t) = V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t) + \mathbf{q}_s \cdot \mathbf{p}_s - \mathbf{q}_t \cdot \mathbf{p}_t, \quad (2.116)$$

where $\mathbf{q}_s = \mathbf{q}_s(z_s, z_t, \mathbf{p}_s, \mathbf{p}_t)$ and $\mathbf{q}_t = \mathbf{q}_t(z_s, z_t, \mathbf{p}_s, \mathbf{p}_t)$. It can also be written as

$$T = W(z_s, z_t, \mathbf{q}_s, \mathbf{p}_t) + \mathbf{q}_s \cdot \mathbf{p}_s, \quad (2.117)$$

or

$$T = W^*(z_s, z_t, \mathbf{p}_s, \mathbf{q}_t) - \mathbf{q}_t \cdot \mathbf{p}_t. \quad (2.118)$$

Using these two alternative definitions we can immediately see that

$$\frac{\partial T}{\partial \mathbf{q}_s} = \mathbf{0}, \quad \frac{\partial T}{\partial \mathbf{q}_t} = \mathbf{0}, \quad (2.119)$$

so that indeed $T = T(z_s, z_t, \mathbf{p}_s, \mathbf{p}_t)$ does not depend on \mathbf{q}_s and \mathbf{q}_t . We can again take partial derivatives and derive eikonal equations. The derivations

are again completely analogous to the ones presented above for the point and mixed characteristics and we will only present the results here. For the angular characteristic T we obtain

$$\mathbf{q}_s = \frac{\partial T}{\partial \mathbf{p}_s}, \quad \mathbf{q}_t = -\frac{\partial T}{\partial \mathbf{p}_t}. \quad (2.120)$$

For a given z_s, z_t and \mathbf{p}_t we have that

$$\mathbf{q}_s = \frac{\partial T}{\partial \mathbf{p}_s} \left(z_s, z_t, -\frac{\partial W}{\partial \mathbf{q}_s}, \mathbf{p}_t \right), \quad \mathbf{p}_s = -\frac{\partial W}{\partial \mathbf{q}_s} \left(z_s, z_t, \frac{\partial T}{\partial \mathbf{p}_s}, \mathbf{p}_t \right), \quad (2.121)$$

so that T is a Legendre transformation from $W(z_s, z_t, \mathbf{q}_s, \mathbf{p}_t)$ to $T(z_s, z_t, \mathbf{p}_s, \mathbf{p}_t)$. Moreover, for a given z_s, z_t and \mathbf{p}_s it is true that

$$\mathbf{q}_t = -\frac{\partial T}{\partial \mathbf{p}_t} \left(z_s, z_t, \mathbf{p}_s, \frac{\partial W^*}{\partial \mathbf{q}_t} \right), \quad \mathbf{p}_t = \frac{\partial W^*}{\partial \mathbf{q}_t} \left(z_s, z_t, \mathbf{p}_s, -\frac{\partial T}{\partial \mathbf{p}_t} \right), \quad (2.122)$$

so that T is a Legendre transformation from $W^*(z_s, z_t, \mathbf{p}_s, \mathbf{q}_t)$ to $T(z_s, z_t, \mathbf{p}_s, \mathbf{p}_t)$. The angular characteristic satisfies

$$\frac{\partial T}{\partial z_s} = H(z_s, \mathbf{q}_s, \mathbf{p}_s), \quad \frac{\partial T}{\partial z_t} = -H(z_t, \mathbf{q}_t, \mathbf{p}_t), \quad (2.123)$$

and the corresponding eikonal equations are

$$\left(\frac{\partial T}{\partial z_s} \right)^2 + |\mathbf{p}_s|^2 = n(\mathbf{q}_s, z_s)^2, \quad \frac{\partial T}{\partial z_s} < 0, \quad (2.124a)$$

$$\left(\frac{\partial T}{\partial z_t} \right)^2 + |\mathbf{p}_t|^2 = n(\mathbf{q}_t, z_t)^2, \quad \frac{\partial T}{\partial z_t} > 0. \quad (2.124b)$$

Hamilton's angular characteristic T is defined as the Legendre transformation

$$T(z_s, z_t, \mathbf{p}_s, \mathbf{p}_t) = V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t) + \mathbf{q}_s \cdot \mathbf{p}_s - \mathbf{q}_t \cdot \mathbf{p}_t. \quad (2.125)$$

The positions on the source plane and on the target plane can be derived as

$$\mathbf{q}_s = \frac{\partial T}{\partial \mathbf{p}_s}, \quad \mathbf{q}_t = -\frac{\partial T}{\partial \mathbf{p}_t}. \quad (2.126)$$

2.7.5 Interpretation of the characteristic functions

The characteristics V , W , W^* en T can be interpreted geometrically. The interpretation below is based on explanations given by the mathematicians and physicists *Rudolf Luneburg* in his *Mathematical Theory of Optics* (1964) [99, p. 102] and *Max Born and Emil Wolf* in the textbook *Principles of Optics* (1959) [12, p. 136]. These two books are classic science books of the twentieth century, and probably the most influential in optics published in the past sixty years. However, reading these sections together with my team led to multiple occasions of confusion about the geometrical meaning of the characteristic functions. The characteristic functions are measures of optical path length from a point on the source plane to a point on the target plane, but a different plane or planes needs to be considered if the characteristic depends on the direction coordinate \mathbf{p}_s and/or \mathbf{p}_t . With this section I am aiming to resolve the confusion once and for all!

As in the previous sections, we consider an optical system in the vicinity of the source plane $z = z_s$ with origin O_1 and target plane $z = z_t \geq z_s$ with origin O_2 . A light source is located in the source plane at $O_s(\mathbf{q}_s, z_s)$, a ray propagates through the optical system and arrives at the target plane at $O_t(\mathbf{q}_t, z_t)$. The point characteristic function V defined in (2.86) is equal to the optical path length from O_s to O_t as the ray passes through the system. The refractive index at O_s is $n_1(z_s, \mathbf{q}_s)$ and at O_2 it is $n_2(z_t, \mathbf{q}_t)$. The momentum at the source plane is given by $\mathbf{p}_s^* = (\mathbf{p}_s, p_{s,3})$ and the momentum at the target plane is given by $\mathbf{p}_t^* = (\mathbf{p}_t, p_{t,3})$.

We use Figure 2.10 as a reference. The source and target planes $z = z_s$ and $z = z_t$ are drawn in pink. We next consider the mixed characteristic W in (2.106), dependent on the initial position \mathbf{q}_s and final momentum \mathbf{p}_t . To get a feeling for the interpretation of the function, we measure the distance between the intercepts of the light ray with two planes. The first plane is just the source plane $z = z_s$, because of the dependence of W on \mathbf{q}_s . The second plane is the plane that is normal to the final momentum \mathbf{p}_t^* of the ray and that contains the origin O_2 . This plane is drawn in green in Figure 2.10. In the figure, we consider a piecewise constant n so that the final portion of the ray with direction \mathbf{p}_t^* is a straight line. The point Q_2 is found if we drop a line from the origin O_2 perpendicular onto the final portion of the ray. In Figure 2.11, a two-dimensional representation is drawn of the ray on the plane spanned by \mathbf{q}_t and \mathbf{p}_t^* . In Figure 2.11a, the point Q_2 has a z -coordinate $z < z_t$ and we can derive

$$\overrightarrow{O_2 O_t} \cdot \frac{\overrightarrow{Q_2 O_t}}{d(Q_2, O_t)} = d(O_2, R), \quad (2.127)$$

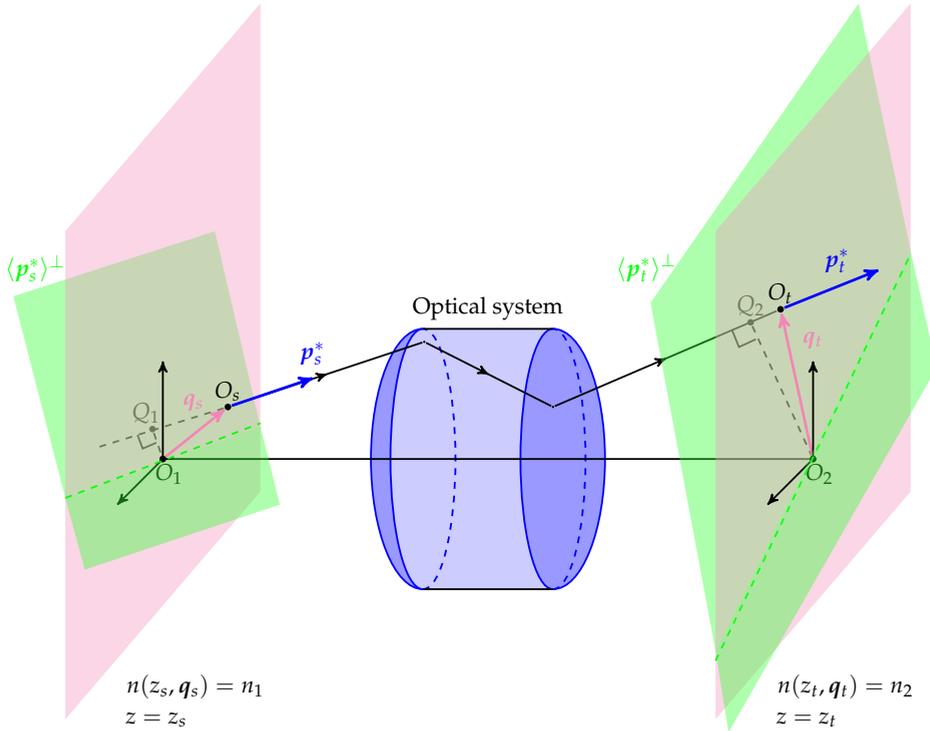


Figure 2.10: Illustration for the geometric interpretation of Hamilton's characteristic functions.

where we use the notation \overrightarrow{PQ} as the vector from point P to Q , $d(P, Q)$ denotes the distance between P and Q and R is the point as indicated in Figure 2.11a. Equation (2.127) is the same as, using $|\mathbf{p}_t^*| = n_2$,

$$(\mathbf{q}_t, 0) \cdot \frac{1}{n_2} \mathbf{p}_t^* = d(Q_2, O_t). \quad (2.128)$$

Similarly, if the point Q_2 has z -coordinate $z > z_t$ as in Figure 2.11b, we have

$$\overrightarrow{O_2 O_t} \cdot \frac{\overrightarrow{O_t Q_2}}{d(O_t, Q_2)} = -d(O_2, R), \quad (2.129)$$

which is equal to

$$(\mathbf{q}_t, 0) \cdot \frac{1}{n_2} \mathbf{p}_t^* = -d(Q_2, O_t). \quad (2.130)$$

We see that (2.128) and (2.130) reduce to

$$\mathbf{q}_t \cdot \mathbf{p}_t = \pm n_2 d(Q_2, O_t), \quad (2.131)$$

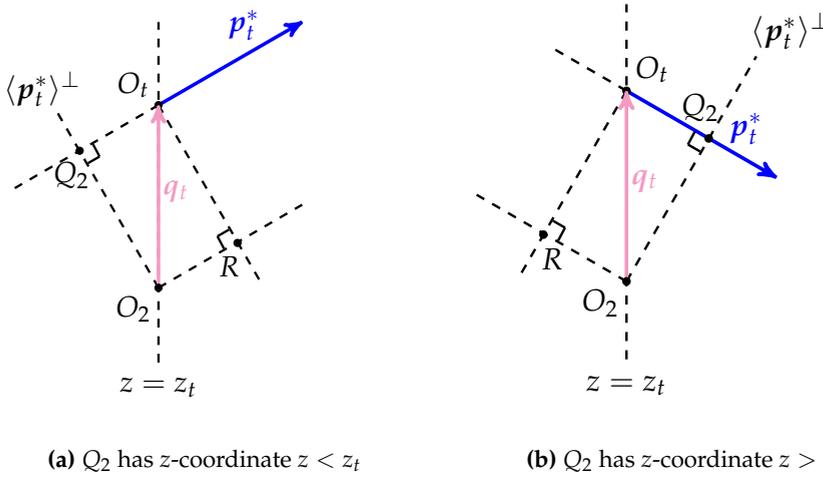


Figure 2.11: Illustration for the geometric interpretation of Hamilton's characteristic functions.

as a *signed optical distance function* from Q_2 to O_t . The sign is determined by the inner product. It is positive if the point Q_2 has a z -coordinate $z < z_t$ as in (2.128) and negative if $z > z_t$ as in (2.130) (and zero if $z = z_t$).

The mixed characteristic W can be written as

$$W(z_s, z_t, \mathbf{q}_s, \mathbf{p}_t) = V - \mathbf{q}_t \cdot \mathbf{p}_t = V \mp n_2 d(Q_2, O_t), \quad (2.132)$$

where V is the optical path length from O_s to O_t . The distance $d(Q_2, O_t)$ is subtracted if Q_2 has a z -coordinate $z < z_t$ and added if $z > z_t$. Hence, W is the distance from O_s to Q_2 .

A similar relation for W^* in (2.114) can be derived, which is essentially the opposite of the relation for W . Here the second plane is just the target plane $z = z_t$, and the first plane is the plane that is normal to the initial momentum \mathbf{p}_s^* of the ray and that contains the origin O_1 . This plane is drawn in green in Figure 2.10. We determine Q_1 by drawing a perpendicular from the origin O_1 to the initial portion of the ray. The signed optical distance function from the green plane to O_s is

$$n_1 d(Q_1, O_s) = \pm \mathbf{q}_s \cdot \mathbf{p}_s, \quad (2.133)$$

and the mixed characteristic W^* can be written as

$$W^*(z_s, z_t, \mathbf{p}_s, \mathbf{q}_t) = V + \mathbf{q}_s \cdot \mathbf{p}_s = V \pm n_1 d(Q_1, O_s), \quad (2.134)$$

where V is the optical path length from O_s to O_t . The distance $d(Q_1, O_s)$ is added if Q_1 has a z -coordinate $z < z_s$ and subtracted if $z > z_s$. Hence, W^* is the distance from Q_1 to O_t .

The angular characteristic T in (2.125) is a combination of both interpretations above. We have that

$$\begin{aligned} T(z_s, z_t, \mathbf{p}_s, \mathbf{p}_t) &= V + \mathbf{q}_s \cdot \mathbf{p}_s - \mathbf{q}_t \cdot \mathbf{p}_t \\ &= V \pm n_1 d(Q_1, O_s) \mp n_2 d(Q_2, O_t), \end{aligned} \quad (2.135)$$

where V is the optical path length from O_s to O_t .

Lastly, we note that it is very convenient to choose the target plane $z = z_t$ to coincide with the source plane $z = z_s$, so $z_t = z_s$. If the optical system is a reflector which reflects the light rays backwards along the optical axis, it is physically reasonable to choose $z_t = z_s$. If the optical system is a lens refracting the light rays but not altering the forward propagation, we can still consider $z_t = z_s$ by forming a virtual image of the final portion of the ray onto the source plane, found by extending a line tangent to the final direction \mathbf{p}_t^* of the optical system back onto the source plane; see Figure 2.12. Here the final portion of the ray is straight in an isotropic medium after it has passed through the optical system and we can draw a tangent line anywhere along the final portion of the ray. Since $z_t = z_s$, the origin O_2 coincides with O_1 and we can find Q_2 with respect to O_1 . In the next chapter of this thesis, Hamilton's characteristic functions are derived exactly this way, with $z = z_s = z_t$, for numerous examples of optical systems.

2.7.6 Hamilton's characteristics and parallel/point sources and targets

We have previously remarked that Hamiltonian systems conserve étendue, i.e., phase space volume. In this section we will make a special remark about zero-étendue light sources, where the phase space volume of a collection of rays originating at a source plane is zero. The two main examples are a parallel beam of rays, for which $\mathbf{p}_s = \mathbf{0}$, i.e., the beam is oriented such that all rays are directed parallel to the optical axis, and a point source, for which $\mathbf{q}_s = \mathbf{0}$ for all rays as they originate from the same point. Hence, the volume of the four-dimensional manifold of all positions \mathbf{q} and momenta \mathbf{p} is zero. Similarly, we will say something about parallel and point targets, where the target is required to be reached by a parallel beam or coincides with a single point. In those cases, $\mathbf{p}_t = \mathbf{0}$ or $\mathbf{q}_t = \mathbf{0}$, respectively.

For zero-étendue sources, a particular choice of Hamilton's characteristics gives additional properties. For parallel sources, the characteristics V and W have the additional property that

$$\mathbf{p}_s = -\frac{\partial V}{\partial \mathbf{q}_s} = -\frac{\partial W}{\partial \mathbf{q}_s} = \mathbf{0}, \quad (2.136)$$

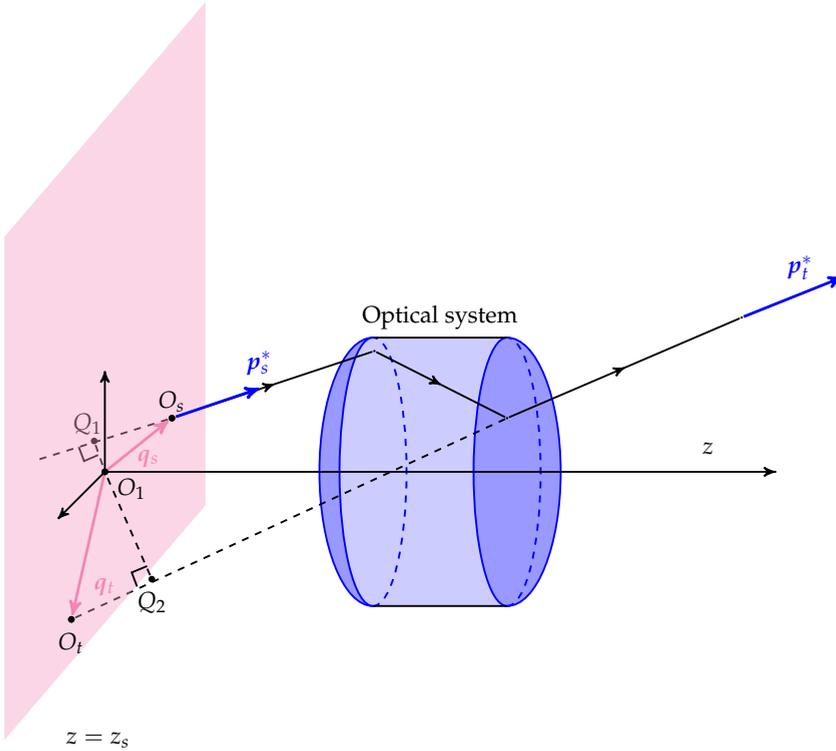


Figure 2.12: Extending p_t^* onto the source plane gives a virtual image O_t .

using (2.87) and (2.107). Hence, $V = V(z_s, z_t, q_t)$ and $W = W(z_s, z_t, p_t)$ are now independent of q_s . Similarly, for point sources, the characteristics W^* and T have the additional property that

$$q_s = \frac{\partial W^*}{\partial p_s} = \frac{\partial T}{\partial p_s} = 0, \quad (2.137)$$

using (2.115) and (2.126). Hence, $W^* = W^*(z_s, z_t, q_t)$ and $T = T(z_s, z_t, p_t)$.

We can find similar results for parallel and point targets. For parallel targets, it holds for the characteristics V and W^* that

$$p_t = \frac{\partial V}{\partial q_t} = \frac{\partial W^*}{\partial q_t} = 0, \quad (2.138)$$

using (2.87) and (2.115). Hence, $V = V(z_s, z_t, q_s)$ and $W^* = W^*(z_s, z_t, p_s)$ are now independent of q_t . For point targets, we obtain for the characteristics W

and T that

$$\mathbf{q}_t = -\frac{\partial W}{\partial \mathbf{p}_t} = -\frac{\partial T}{\partial \mathbf{p}_t} = \mathbf{0}, \quad (2.139)$$

using (2.107) and (2.126). Hence, $W = W(z_s, z_t, \mathbf{q}_s)$ and $T = T(z_s, z_t, \mathbf{p}_s)$ are now independent of \mathbf{p}_t .

For parallel light sources the characteristics V and W are independent of the source coordinates \mathbf{q}_s and momentum \mathbf{p}_s . For point light sources the characteristics W^ and T are independent of the source coordinates \mathbf{q}_s and momentum \mathbf{p}_s . For parallel targets the characteristics V and W^* are independent of the target coordinates \mathbf{q}_t and \mathbf{p}_t . For point targets the characteristics W and T are independent of the target coordinates \mathbf{q}_t and \mathbf{p}_t .*

A consequence of the above is that for systems with parallel or point sources *and* parallel or point targets one of Hamilton's characteristic functions becomes a constant.

2.8 Radiometric and photometric units

An electromagnetic wave carries energy in the direction of the light ray and Poynting vector. This energy is proportional to the amplitude of the wave squared, cf. (2.6). If multiple light rays are emitted from a source domain, we can speak of a light distribution on the source domain.

The magnitude of the Poynting vector \mathbf{S} in (2.7) is the power per unit area crossing a surface parallel to the wavefront, in W/m^2 . Multiple light rays together form a beam of light, which can be parallel, diverging or converging. When we talk about light illuminating a particular target object, surface or domain, we often talk about the *irradiance* on the target.

The energy of the light reaching a detector cannot be measured instantaneously, but the detector must integrate the energy flux over some finite time. The *radiant flux* Φ_R arriving at a surface Σ is defined as the energy flux per unit time in W and also often denoted as

$$\Phi_R = \int_{\Sigma} \langle \mathbf{S} \rangle \cdot \hat{\mathbf{n}} \, dA, \quad (2.140)$$

where $\hat{\mathbf{n}}$ is the unit normal vector to the surface, dA is the area of an infinitesimal piece of Σ , and $\langle \mathbf{S} \rangle$ is the time-average of the instantaneous Poynting vector \mathbf{S} , i.e.,

$$\langle \mathbf{S} \rangle = \frac{1}{T} \int_0^T \mathbf{S}(t) \, dt, \quad (2.141)$$

cf. (2.7) and (2.9), taken over a sufficiently large time interval T .

Moreover, a detector always has an entrance window that admits radiant flux through some fixed area. If we divide the radiant flux incident on a surface by the area of the surface, we have measured the *irradiance*. Irradiance is the power per unit area incident on or emitted from a surface and can be written as

$$P_R = \frac{d\Phi_R}{dA}, \quad (2.142)$$

where A is the area of the entrance window. When light is emitted from a surface instead of incident on a surface, we speak of the *radiant exitance*, as the emitted flux per unit area. Both the irradiance and radiant exitance can also be referred to as a *flux density*.

In the case where light is emitted from a point light source, we are often interested in the angular density of the emitted conical beam, also called the *radiant intensity*. The area of the unit sphere centered around the light source through which the rays cross, is also called a *solid angle*. We denote solid angles by Ω and the unit of solid angle is *steradian* [sr]. The radiant intensity I_R is the energy flux per unit solid angle, given by

$$I_R = \frac{d\Phi_R}{d\Omega}. \quad (2.143)$$

The radiant flux, exitance, irradiance and intensity are *radiometric* quantities, characterizing the distribution of the radiation's power. Radiometers can for example be used to determine the temperature of objects and gases, think about an infrared thermometer, and for earth remote sensing.

The output of lighting systems is often measured in lumens [lm] instead of watts, taking into account the impression the light gives on the human eye. It is a measure of the total energy of visible light emitted by a source per unit time. The sensitivity of the human eye to light in the visible spectrum (with wavelengths of approximately 384-769 nm) varies with wavelength. We perceive light of different wavelengths in the visible spectrum as different colors: 635-700 nm is red, 520-560 nm is green, 450-490 nm is blue, for instance. The photoreceptors in our eyes, *rods and cones*, are responsible for detecting *chromaticity* and luminous intensity, perceived by our brain as color and brightness, respectively. In this thesis, I will not discuss how the human perception of color is translated into measurable quantities such as *tristimulus* values (e.g., RGB values) and the *CIE Chromaticity Diagram*, but I am referring to the book 'Color Vision and Colorimetry: Theory and Applications' by *Daniel Malacara-Hernández* [102].

The *spectral radiant flux* $\Phi_{R,\lambda}$ in W/m is defined as the radiant flux per unit

wavelength, i.e.,

$$\Phi_{R,\lambda} = \frac{d\Phi_R}{d\lambda}. \quad (2.144)$$

The *spectral luminous flux* Φ_λ in lm/m is the flux as perceived by the human eye as a function of wavelength. Experiments have been done to quantify the subjective impression produced by stimulating the human visual system with radiant energy. The *luminous efficacy* $\sigma(\lambda)$, which is the ratio between the spectral luminous flux and spectral radiant flux, has been determined for each wavelength as

$$\sigma(\lambda) = \frac{\Phi_\lambda}{\Phi_{R,\lambda}}. \quad (2.145)$$

This function represents the response of a typical eye and is a bell-shaped curve with a maximum of 683 lm/W at green light of 540 nm.

The radiometric quantities described in (2.142) and (2.143) have *photometric* analogues that are metrics in terms of the perceived brightness to the human eye. In the remainder of this thesis, we will only use photometric quantities. The radiant flux has an analogue called the *luminous flux*, which is the spectral luminous flux integrated over all wavelengths that make up a light beam, i.e.,

$$\Phi = \int_0^\infty \Phi_\lambda d\lambda. \quad (2.146)$$

Monochromatic light refers to visible light with a narrow band of wavelengths. White light is a combination of multiple wavelengths of the visible spectrum. The *illuminance* is the luminous flux per unit area on a surface in lm/m² given as

$$L = \frac{d\Phi}{dA}, \quad (2.147)$$

where A is the area of the surface. When light is emitted from a surface instead of incident to it, we speak of the *emittance*. The *luminous intensity* is the luminous flux per unit solid angle in lm/sr, i.e.,

$$I = \frac{d\Phi}{d\Omega}, \quad (2.148)$$

where Ω is the subtended solid angle by a conical beam. The luminous intensity is sometimes also expressed in unit of candela cd, and 1 cd = 1 lm/sr, equal to the luminous intensity emitted by a standardized candle.

2.9 Summary

In this chapter, we presented the principles of geometrical optics and described the propagation of light in terms of rays. From Maxwell's equations

and the short-wavelength approximation we obtained the eikonal equation. Subsequently, we presented Fermat's principle and the laws of reflection and refraction.

By describing the propagation of a ray in terms of the position coordinates q and direction coordinates p we reformulated the Euler-Lagrange equations of the optical path length from a plane $z = z_s$ to a plane $z = z_t$ as a first-order system of ODEs of a Hamiltonian system. We introduced Hamilton's characteristic functions V , W , W^* and T , which are measures of optical path length.

For parallel or point light sources and targets we found some special properties of the characteristic functions. For optical systems with a parallel or point source, we can always find a characteristic function which is independent of the source coordinates. In the next chapter, we will consider 16 optical systems with parallel and point sources. Some of these systems also have a parallel or point target, in which case we can find a characteristic function that is independent of both the source and the target coordinates. We will use these properties of the characteristic functions to find useful geometric relations for each optical system.

Chapter 3

Reflector and Lens Equations

This chapter is a compendium of mathematical descriptions of optical systems. We consider either ideal parallel or ideal point sources. The targets are defined as a domain located in the *near field* or in the *far field*, an area on a screen reached by a parallel beam, or a single point. Using a minimum number of either reflector or lens surfaces, we can already think of 16 different optical systems. We will call these systems the 16 *base cases*.

For each of the base cases, we can use Hamilton's characteristic functions to find useful geometric relations. In the previous chapter, in Section 2.7.6, we saw that for parallel and point sources we can find at least one characteristic function which is independent of \mathbf{q}_s and \mathbf{p}_s . Similarly, for parallel and point targets we can find at least one independent of \mathbf{q}_t and \mathbf{p}_t . In Figure 3.1 and 3.2 our results are summarized for the 16 base cases, which consist of 8 reflector and 8 lens systems. The source and target coordinates and distributions are also given, which will be explained in detail later.

We see that for systems with both a parallel or point source *and* a parallel or point target the optical path length for all rays is equal to the same constant, i.e., one of Hamilton's characteristic functions becomes a constant and is independent of $\mathbf{q}_s, \mathbf{q}_t, \mathbf{p}_s$ and \mathbf{p}_t . For example, if we consider a parallel source and a parallel target, the characteristic $V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t)$ can be shown to be independent of \mathbf{q}_s and \mathbf{q}_t since \mathbf{p}_s and \mathbf{p}_t are zero vectors, as demonstrated in Section 2.7.6. This is a consequence of the *Theorem of Malus and Dupin* [12, p. 130], which states that the direction of each ray will remain perpendicular to the wavefront after any number of reflections or refractions. For parallel and point sources and targets we only consider parallel and spherical wavefronts, which are surfaces of constant optical path length. As a result, one of Hamilton's characteristic functions becomes constant. These optical systems are composed of two freeform optical surfaces, since one optical sur-

face cannot simultaneously make a collimated or spherical beam and change the intensity profile of the beam [164].

In this chapter, we derive Hamilton's characteristic functions for a subset of the base cases. We consider a parallel-to-far-field reflector, parallel-to-near-field reflector, point-to-far-field reflector, point-to-far-field lens and a point-to-parallel reflector. By choosing these systems we cover all characteristic functions and my early work on point-to-far-field systems [131, 133]. As we will see later, this subset includes optical systems that can be treated from an optimal-transport point of view and optical systems that cannot, such as the parallel-to-near-field system published in [130].

For all 16 base cases, there are two unknowns. One unknown is the location of the optical surface and the other unknown is either a function related to the characteristic function or the location of the second optical surface in the system if we consider a parallel or point target. The relationship between these two unknowns can take several forms. For some systems these unknowns can be linked by an associated *cost function* in optimal transport theory. In our subset this holds for all systems except for the parallel-to-near-field reflector. The unknowns can also be linked by formulating a *generating function*. We will show that we can find a generating function for all systems in our subset.

We start this chapter by giving an introduction on cost functions and generating functions. Subsequently, we introduce the concepts of the *far-field approximation* and *stereographic coordinates*. Using a target domain in the far field means that we can neglect the dimensions of the optical element compared to the distance the light travels. In the near field the size of the optical element is not negligible. Stereographic coordinates can be used when dealing with spherical coordinate systems in the case of a point light source and/or in the case of a far-field target. Next, we define the source and target intensity distributions. After these introductory steps, we discuss the optical systems in our subset in detail. We present a summary of the known cost functions and all generating functions of the 16 base cases in Section 3.7.

Unless mentioned otherwise, we consider the medium through which the light rays travel before and after the optical system to be isotropic with the refractive index of air, i.e., $n = 1$. If the optical system is a lens, the lens itself is considered isotropic as well with a refractive index $n > 1$ and assumed to be a *perfect transmittor*. Note that TIR can still occur. If the optical system is a *perfect mirror* or reflector, all light rays are reflected specularly. It should be mentioned that for lenses the index does have an effect on the polarization of the outgoing light, but we do not consider polarization effects.

We list the following conventions for the geometric derivations and notation presented in this chapter.

- Unit vectors are denoted using hats. Example: \hat{s} , \hat{t} and \hat{n} .
- We consider the medium through which the light travels before and after the optical system to be isotropic with refractive index $n = 1$.
- The unit vector $\hat{s} = (s_1, s_2, s_3)$ denotes the direction of the incoming ray, i.e., the momentum \mathbf{p}_s^* divided by the refractive index n of the medium through which the ray travels.
- The unit vector $\hat{t} = (t_1, t_2, t_3)$ is the direction of the outgoing ray after it has been reflected or refracted, i.e., the momentum \mathbf{p}_t^* divided by the refractive index n of the medium through which the ray travels.
- The source coordinates denoted by $\mathbf{x} \in \mathcal{X}$ are defined as $\mathbf{x} = \mathbf{q}_s$ for a parallel beam of light and as the stereographic projections of \hat{s} for a point light source. This is not to be confused with the parametrization of a light ray $\mathbf{x}(z)$ in the previous chapter.
- The target coordinates denoted by $\mathbf{y} \in \mathcal{Y}$ are defined as $\mathbf{y} = \mathbf{q}_t$ for a near-field target and as the stereographic projections of \hat{t} for a far-field target.
- We orient the z -axis as the vertical axis in all figures.

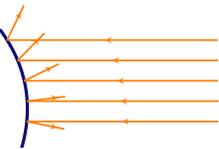
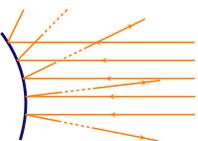
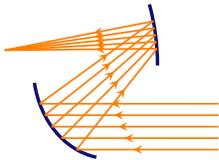
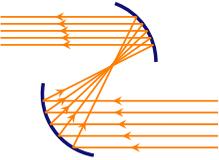
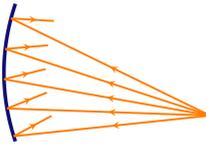
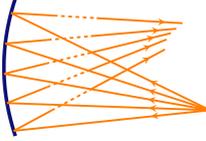
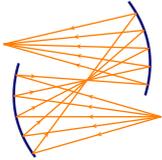
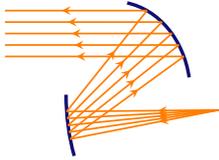
<p>Target Source</p>	<p>Near-field $g(y)$ $y = q_t$</p>	<p>Far-field $\tilde{g}(y)$ stereographic y</p>	<p>Point $\tilde{g}(y)$ stereographic y</p>	<p>Parallel $g(y)$ $y = q_t$</p>
<p>Parallel $f(x)$ $x = q_s$</p>	<p>$V(q_t)$</p> 	<p>$W(p_t)$</p> 	<p>W</p> 	<p>V</p> 
<p>Point $\tilde{f}(x)$ stereographic x</p>	<p>$W^*(q_t)$</p> 	<p>$T(p_t)$</p> 	<p>T</p> 	<p>W^*</p> 

Figure 3.1: Reflector systems: overview of different source and target domains and their respective coordinate systems.

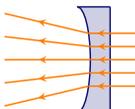
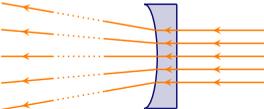
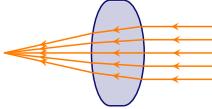
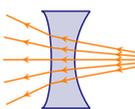
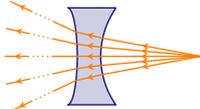
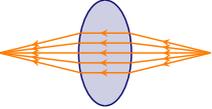
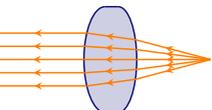
Target Source	Near-field $g(y)$ $y = q_t$	Far-field $\tilde{g}(y)$ stereographic y	Point $\tilde{g}(y)$ stereographic y	Parallel $g(y)$ $y = q_t$
	Parallel $f(x)$ $x = q_s$	 $V(q_t)$	 $W(p_t)$	 W
Point $\tilde{f}(x)$ stereographic x	 $W^*(q_t)$	 $T(p_t)$	 T	 W^*

Figure 3.2: Lens systems: overview of different source and target domains and their respective coordinate systems.

3.1 Geometric conventions

For each optical system we will derive geometric relations using Hamilton's characteristic functions. The unknown variable $u(x)$ defines the location of (one of) the optical surface(s) in a given parametrization as a function of the source coordinates $x \in \mathcal{X}$. For some systems we can formulate an *optimal-transport formulation*, i.e., an equation of the form

$$u_2(\mathbf{y}) - u_1(x) = c(x, \mathbf{y}), \quad (3.1)$$

where $u_1(x)$ is a geometric variable related to $u(x)$. The function $u_2(\mathbf{y})$ is a geometric variable related to one of Hamilton's characteristic functions or to a second optical surface if we consider a point or parallel target. Lastly, $c(x, \mathbf{y})$ is an optimal-transport cost function. The source coordinates $x \in \mathcal{X}$ are position coordinates, i.e., $x = q_s$, or stereographic coordinates, which will be introduced in Section 3.1.2. The same holds for the target coordinates $\mathbf{y} \in \mathcal{Y}$; either $\mathbf{y} = q_t$ or \mathbf{y} denotes stereographic coordinates.

For a number of systems, 7 out of the 16 base cases, the relation in (3.1) cannot be found using any parametrization of the optical surface(s). Prime examples of optical systems which cannot be put in the optimal-transport framework are systems with a near-field target, such as the parallel-to-near-field reflector in the subset of systems which we will discuss in this chapter. However, for all systems we can find a *generating function* of the form

$$u(x) = G(x, \mathbf{y}, w(\mathbf{y})), \quad (3.2)$$

where $w(\mathbf{y})$ is a function related to one of Hamilton's characteristic functions or to the second optical surface in the system for a point or parallel target. In fact, $w(\mathbf{y})$ can be written as $w(\mathbf{y}) = H(x, \mathbf{y}, G(x, \mathbf{y}, w(\mathbf{y})))$, where H is the unique inverse of G for given $x \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$. Note that H is not related to the Hamiltonian H discussed in the previous chapter. In Chapter 4 we will present the mathematical theory on cost functions in optimal transport theory and on generating functions.

For each optical system we can obtain a mapping $\mathbf{y} = \mathbf{m}(x)$ that connects coordinates from the source domain $x \in \mathcal{X}$ to coordinates from the target domain $\mathbf{y} \in \mathcal{Y}$. For some systems it turns out that we can find this mapping \mathbf{m} relatively easily. In fact, \mathbf{m} is just the law of reflection or Snell's law in some cases, written in a different coordinate system. We will deduce a compact expression for the mapping for the parallel-to-far-field reflector, parallel-to-near-field reflector and point-to-far-field reflector. The derivation of the mapping for the point-to-far-field lens and point-to-parallel reflector is more complicated and we will not present these mappings in this thesis.

However, in Chapter 4, we show that we can find an implicit expression for the mapping using the relations (3.1) or (3.2) and results from *convexity theory*.

For the optical systems considered in this thesis we can formulate a generating function

$$u(\mathbf{x}) = G(\mathbf{x}, \mathbf{y}, w(\mathbf{y})), \quad (3.3)$$

with unique inverse

$$w(\mathbf{y}) = H(\mathbf{x}, \mathbf{y}, u(\mathbf{x})). \quad (3.4)$$

Here, $w(\mathbf{y})$ is related to one of Hamilton's characteristic functions or to a second optical surface. For all base-case optical systems it is a function of the target coordinates \mathbf{q}_t or \mathbf{p}_t only, or equal to a constant. For some systems a cost function in optimal transport theory can also be formulated.

3.1.1 Far-field approximation

Let us illustrate the far-field approximation for a parallel source and a lens surface. The light source emits a parallel beam in the positive z -direction and we have $\hat{\mathbf{s}} = \hat{\mathbf{e}}_z$, $\mathbf{x} = \mathbf{q}_s$ and $\mathbf{p}_s = \mathbf{0}$. We consider a lens composed of a flat bottom surface and a freeform surface $\mathcal{L} : z = u(\mathbf{x})$. A 2D representation in Figure 3.3 is given showing the plane of incidence. The first surface does not alter the direction of the rays $\hat{\mathbf{s}} = \hat{\mathbf{e}}_z$. The rays are refracted by the second surface \mathcal{L} . The refracted light ray is denoted by $\hat{\mathbf{t}}$ and obviously $\hat{\mathbf{t}} = \hat{\mathbf{t}}(\mathbf{x})$.

The position vector $\mathbf{r} = \mathbf{r}(\mathbf{x})$ of a point on the refracted ray at a distance R from the freeform lens surface, measured along the ray from the point P to Q , is given by

$$\mathbf{r}(\mathbf{x}) = u(\mathbf{x}) \hat{\mathbf{e}}_z + R \hat{\mathbf{t}}(\mathbf{x}). \quad (3.5)$$

From the relation $|\mathbf{r}|^2 = (u \hat{\mathbf{e}}_z + R \hat{\mathbf{t}}) \cdot (u \hat{\mathbf{e}}_z + R \hat{\mathbf{t}}) = u^2 + 2uRt_3 + R^2$ with $t_3 = \hat{\mathbf{t}} \cdot \hat{\mathbf{e}}_z$, we can find the following expression for the corresponding unit vector $\hat{\mathbf{r}} \in S^2$, where S^2 denotes the unit sphere, as

$$\hat{\mathbf{r}} = \frac{u \hat{\mathbf{e}}_z + R \hat{\mathbf{t}}}{\sqrt{u^2 + 2uRt_3 + R^2}} = \frac{\frac{u}{R} \hat{\mathbf{e}}_z + \hat{\mathbf{t}}}{\sqrt{\left(\frac{u}{R}\right)^2 + 2\frac{u}{R}t_3 + 1}}. \quad (3.6)$$

We assume $R \gg u(\mathbf{x})$ and obtain in good approximation $\hat{\mathbf{r}} = \hat{\mathbf{t}}$, referred to as the far-field approximation. Thus, we can replace the scaled position vector $\hat{\mathbf{r}}$ by the direction vector $\hat{\mathbf{t}}$ of the refracted ray and approximate the lens surface as a point located in the origin O of the coordinate system.

For other systems, e.g., using a point source or a reflector, we can easily show a similar result.

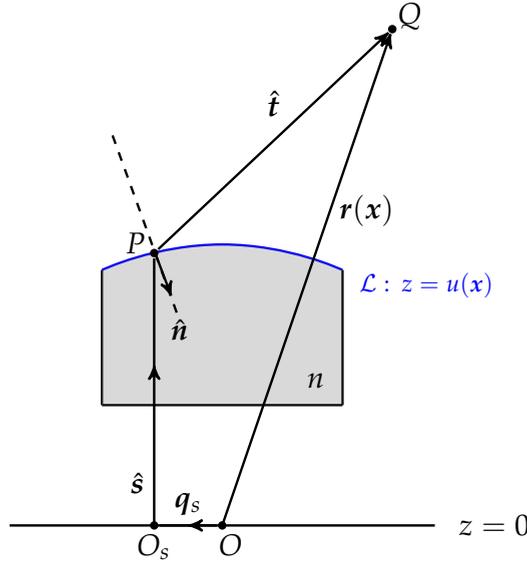


Figure 3.3: Illustration of the far-field approximation for a parallel source and lens surface.

In the far-field approximation, the optical surface is approximated as a point in space located at the origin O .

3.1.2 Stereographic coordinates

For the vectors $\hat{s}, \hat{t} \in S^2$ there are only two rotational degrees of freedom and we can choose to express the third component in terms of the first two components. For this reason it is convenient to perform coordinate transformations from spherical to stereographic. We use a 2-tuple representation for stereographic coordinate vectors and define

$$x(\hat{s}) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{1 + s_3} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \frac{1}{1 + \cos(\phi)} \begin{pmatrix} \sin(\phi) \cos(\theta) \\ \sin(\phi) \sin(\theta) \end{pmatrix}, \quad (3.7)$$

for the source direction \hat{s} with $0 \leq \phi \leq \pi$ the zenith and $0 \leq \theta < 2\pi$ the azimuth in the spherical coordinate system.

Equation (3.7) expresses the incoming rays \hat{s} as a stereographic projection from the south pole $(0, 0, -1)$ onto the plane $z = 0$, as drawn schematically in Figure 3.4a. By making this choice we assume that the point light source emits light mainly in the upward z -direction. The stereographic projection in (3.7) is

undefined at the south pole, and we consider $s_3 \neq -1$ and $0 \leq \phi < \pi$. The reason for using the south pole for the incoming rays is that if we consider the point source to emit a conical beam of rays in the upward direction, we obtain a bounded source domain in stereographic coordinates which we can easily discretize.

For the target direction $\hat{\mathbf{t}}$ we define

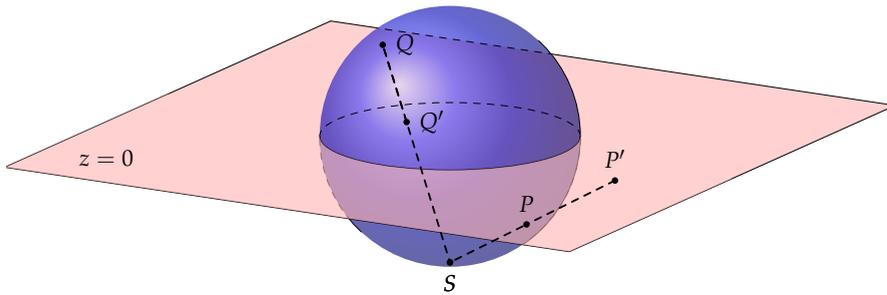
$$\mathbf{y}(\hat{\mathbf{t}}) = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \frac{1}{1 \pm t_3} \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = \frac{1}{1 \pm \cos(\psi)} \begin{pmatrix} \sin(\psi) \cos(\chi) \\ \sin(\psi) \sin(\chi) \end{pmatrix}, \quad (3.8)$$

$0 \leq \psi \leq \pi$ is the zenith and $0 \leq \chi < 2\pi$ is the azimuth in the spherical coordinate system. The origin of the coordinate system describing the target is the optical surface approximated as a point in space and $\hat{\mathbf{r}} = \hat{\mathbf{t}}$ as shown in the previous section.

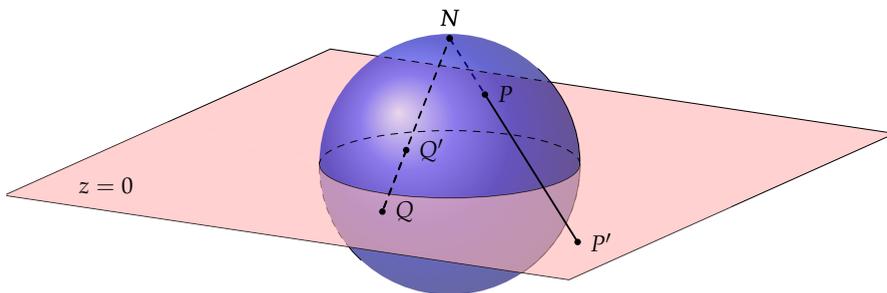
The \pm -sign in (3.8) indicates two options; we take the stereographic projection with respect to a south or north pole, respectively. We use a stereographic projection from the north pole $(0, 0, 1)$ onto the plane $z = 0$ in case the optical surface is a reflector \mathcal{R} , as shown in Figure 3.4b. This leads to a minus sign in (3.8). Hence, we assume that the outgoing rays are reflected back and not reflected upwards. The stereographic projection in (3.8) is undefined at the north pole, and we consider $t_3 \neq 1$ and $0 < \psi \leq \pi$, i.e., we assume the reflector surface does not reflect rays upward parallel to the z -axis. We choose the north pole to ensure that the stereographic projection is always defined. In case the optical surface is a lens \mathcal{L} , we use a stereographic projection from the south pole $(0, 0, -1)$ onto the plane $z = 0$ as shown in Figure 3.4a. The stereographic projection in (3.8), where we now have a plus sign, is undefined at the south pole, and we consider $t_3 \neq -1$ and $0 \leq \psi < \pi$, i.e., we assume the lens surface does not refract the rays downward parallel to the z -axis, so that the stereographic projection is always defined.

The stereographic projections have corresponding inverse projections

$$\hat{\mathbf{s}}(\mathbf{x}) = \hat{\mathbf{e}}_r = \frac{1}{1 + |\mathbf{x}|^2} \begin{pmatrix} 2x_1 \\ 2x_2 \\ 1 - |\mathbf{x}|^2 \end{pmatrix}, \quad \hat{\mathbf{t}}(\mathbf{y}) = \frac{1}{1 + |\mathbf{y}|^2} \begin{pmatrix} 2y_1 \\ 2y_2 \\ \pm(1 - |\mathbf{y}|^2) \end{pmatrix}. \quad (3.9)$$



(a) From the south pole S



(b) From the north pole N

Figure 3.4: Schematic representation of the stereographic projections of the unit sphere S^2 (with center O). The points P are projected to P' and the points Q to Q' .

3.1.3 Source and target distributions

We consider given source and target distributions, which are defined on a particular source and target domain. For each optical system the notation is a little different depending on the type of source and target domains at hand.

In this section, we give a short overview of the notation, summarized in Figure 3.1 and 3.2.

The source distribution of a parallel or point source is expressed as follows, respectively:

- For a parallel source, the emittance of the source is given by a distribution function $f(x)$ [lm/m^2] for $x \in \mathcal{X}$, where \mathcal{X} is the supporting domain of $f(x)$ and $x = \mathbf{q}_s$ are spatial coordinates.
- For a point source, the intensity of the source is given by a distribution function $f(\phi, \theta)$ [lm/sr]. We define our source domain \mathcal{X} as the supporting domain of $\tilde{f}(x) = f(\phi(x), \theta(x))$, with x the stereographic coordinates obtained using the transformation (3.7).

We consider near-field and far-field targets, but also point targets and parallel outgoing beams. Defining the target domain \mathcal{Y} as the image under the mapping m , i.e., $\mathcal{Y} = m(\mathcal{X})$, the target distribution is written as follows:

- For a near-field target, a target illuminance is given as $g(\mathbf{y})$ [lm/m^2] for $\mathbf{y} \in \mathcal{Y}$, where $\mathbf{y} = \mathbf{q}_t$ are spatial coordinates.
- For a far-field target, a target intensity is denoted as $g(\psi, \chi)$ [lm/sr], where (ψ, χ) are spherical coordinates, with zenith $0 \leq \psi \leq \pi$ and azimuth $0 \leq \chi < 2\pi$ with respect to the origin of the coordinate system, approximating the optical surface as a point in space. The target intensity can be rewritten in stereographic coordinates as $\tilde{g}(\mathbf{y}) = g(\psi(\mathbf{y}), \chi(\mathbf{y}))$, with \mathbf{y} the stereographic coordinates obtained using the transformation (3.8), either using a north or south pole.
- For a point target, a target intensity is denoted as $g(\psi, \chi)$ [lm/sr], where (ψ, χ) are spherical coordinates, with zenith $0 \leq \psi \leq \pi$ and azimuth $0 \leq \chi < 2\pi$ with respect to the point target as origin. The target intensity can be rewritten in stereographic coordinates as $\tilde{g}(\mathbf{y}) = g(\psi(\mathbf{y}), \chi(\mathbf{y}))$, with \mathbf{y} the stereographic coordinates obtained using the transformation (3.8), either using a north or south pole.
- For a parallel target, a target illuminance is given as $g(\mathbf{y})$ [lm/m^2] for $\mathbf{y} \in \mathcal{Y}$, where $\mathbf{y} = \mathbf{q}_t$ are spatial coordinates.

In the following, the source and target domains are always denoted by \mathcal{X} and \mathcal{Y} , respectively. It will be clear from the context whether these domains are in Euclidean space or stereographic coordinate space.

We noted above that for a far-field target, a target intensity is denoted as $g(\psi, \chi)$ [lm/sr] or $\tilde{g}(\mathbf{y}) = g(\psi(\mathbf{y}), \chi(\mathbf{y}))$, with \mathbf{y} the stereographic coordinates. However, for many applications with a far-field target the target distribution is given as an illuminance in [lm/m²] on a plane or screen. We will discuss how to change coordinates for such a target distribution in the next section.

3.1.4 Target on a projection screen in the far field

For many applications with far-field targets we are given a specified illuminance $L(\zeta, \eta)$ in lm/m² on a plane P in the far field, where (ζ, η) are the local Cartesian coordinates on P , instead of a target intensity $g(\psi, \chi)$. In this section, we explain how we calculate $g(\psi, \chi)$ and $\tilde{g}(\mathbf{y})$ from $L(\zeta, \eta)$ and how we convert the entire target domain on the screen P to a domain in stereographic coordinate space.

Figure 3.5 shows a target screen P for the point-to-far-field reflector discussed in Section 3.4. Let d be the distance from the origin to the plane P and \mathbf{n}_P be the normal vector to the plane P directed to the origin. Then $d = |\mathbf{n}_P|$, and $\hat{\mathbf{n}}_P = \mathbf{n}_P/d$. Let $\hat{\mathbf{e}}_\zeta, \hat{\mathbf{e}}_\eta$ be an orthonormal basis in P . We say that a point $\zeta \hat{\mathbf{e}}_\zeta + \eta \hat{\mathbf{e}}_\eta - d \hat{\mathbf{n}}_P$ has coordinates (ζ, η) on the plane P . Figure 3.5 shows the plane P in the xz -plane and accompanying vectors are $\hat{\mathbf{e}}_\zeta = (1, 0, 0)$, $\hat{\mathbf{e}}_\eta = (0, 0, 1)$ and $\hat{\mathbf{n}}_P = (0, -1, 0)$. We come back to this example later in this section.

A ray with direction vector $\hat{\mathbf{t}}$ intersects the plane at (ζ, η) if

$$\hat{\mathbf{t}} = \frac{\zeta \hat{\mathbf{e}}_\zeta + \eta \hat{\mathbf{e}}_\eta - d \hat{\mathbf{n}}_P}{\sqrt{\zeta^2 + \eta^2 + d^2}}, \quad (3.10)$$

where we assume that $\hat{\mathbf{t}}$ is emitted from O (far-field approximation), as explained in Section 3.1.1. Taking the inner products with $\hat{\mathbf{e}}_\zeta$, $\hat{\mathbf{e}}_\eta$ and $\hat{\mathbf{n}}_P$, respectively, gives three equations, from which we derive that

$$\zeta = -d \frac{\hat{\mathbf{t}} \cdot \hat{\mathbf{e}}_\zeta}{\hat{\mathbf{t}} \cdot \hat{\mathbf{n}}_P}, \quad \eta = -d \frac{\hat{\mathbf{t}} \cdot \hat{\mathbf{e}}_\eta}{\hat{\mathbf{t}} \cdot \hat{\mathbf{n}}_P}. \quad (3.11)$$

Substituting $\hat{\mathbf{t}}$ from Equation (3.9) we see that ζ and η are functions of \mathbf{y} . Note that we use a plus or minus sign in the stereographic projection in (3.8) depending on the optical system, e.g., we use a minus sign for the reflector system in Figure 3.5.

Next, we need to derive the luminous intensity $\tilde{g}(\mathbf{y}) = g(\psi, \chi)$ [lm/sr] from the illuminance $L(\xi, \eta)$ [lm/m²]. The spherical coordinates (ψ, χ) denote the inclination with respect to $-\hat{\mathbf{n}}_P$ and the azimuth, respectively, with

$$\psi = \arccos(-\hat{\mathbf{t}} \cdot \hat{\mathbf{n}}_P), \quad 0 \leq \psi < \pi, \quad (3.12a)$$

$$\chi = \tan^{-1}(\hat{\mathbf{t}} \cdot \hat{\mathbf{e}}_\xi, \hat{\mathbf{t}} \cdot \hat{\mathbf{e}}_\eta), \quad 0 \leq \chi \leq 2\pi, \quad (3.12b)$$

where the inverse function $\tan^{-1}(t_1, t_3)$ is defined as

$$\tan^{-1}(t_1, t_3) = \begin{cases} \arctan(t_3/t_1), & t_1, t_3 \geq 0, \\ \arctan(t_3/t_1) + \pi, & t_1 < 0, \\ \arctan(t_3/t_1) + 2\pi, & t_1 \geq 0, t_3 < 0. \end{cases} \quad (3.13)$$

Note that the range of \arctan is $(-\pi/2, \pi/2)$, so that $\tan^{-1}(t_1, t_3)$ has range $[0, 2\pi)$.

First, we find an expression for the coordinates of the intersection of $\hat{\mathbf{t}}$ and the target plane P . A ray with angles ψ and χ in this coordinate system intersects the plane at a distance $d/\cos(\psi)$ from the origin and the relation between the planar coordinates ξ, η , and the spherical coordinates ψ, χ is given by

$$\xi = d \tan(\psi) \cos(\chi), \quad (3.14a)$$

$$\eta = d \tan(\psi) \sin(\chi). \quad (3.14b)$$

Local energy conservation gives that $g(\psi, \chi) dS(\psi, \chi) = g(\psi, \chi) \sin(\psi) d\psi d\chi$ on the unit sphere, which we require to be equal to $L(\xi, \eta) d\xi d\eta$. Hence, by changing coordinates we require

$$g(\psi, \chi) = \left| \frac{1}{\sin(\psi)} \frac{\partial(\xi, \eta)}{\partial(\psi, \chi)} \right| L(\xi, \eta). \quad (3.15)$$

Using (3.14) we have

$$\frac{\partial(\xi, \eta)}{\partial(\psi, \chi)} = \frac{\partial \xi}{\partial \psi} \frac{\partial \eta}{\partial \chi} - \frac{\partial \xi}{\partial \chi} \frac{\partial \eta}{\partial \psi} = d^2 \frac{\sin(\psi)}{\cos^3(\psi)}. \quad (3.16)$$

Hence,

$$\tilde{g}(\psi, \chi) = \left| \frac{d^2}{\cos^3(\psi)} \right| L(\xi, \eta). \quad (3.17)$$

From (3.12a) and (3.10), we have

$$\cos(\psi) = \frac{d}{\sqrt{\xi^2 + \eta^2 + d^2}} \geq 0, \quad (3.18)$$

which is positive only if $\psi \leq \pi/2$. In this case, the ray intersects the target plane. Substituting into (3.17) gives

$$\tilde{g}(\mathbf{y}) = g(\psi, \chi) = \frac{(\tilde{\zeta}^2 + \eta^2 + d^2)^{3/2}}{d} L(\tilde{\zeta}, \eta). \quad (3.19)$$

Lastly, we want to know what the entire target domain looks like in the stereographic coordinate space. For this we need \mathbf{y} as a function of $\tilde{\zeta}$ and η in order to convert the boundary of the target domain on P to the boundary $\partial\mathcal{Y}$ of the domain in \mathbf{y} -coordinates. For instance, if we consider a target plane P parallel to the xz -plane at some $y = \ell > 0$ for a reflector system, as shown in Figure 3.5, we have $\hat{\mathbf{e}}_{\tilde{\zeta}} = (1, 0, 0)$, $\hat{\mathbf{e}}_{\eta} = (0, 0, 1)$ and $\hat{\mathbf{n}}_P = (0, -1, 0)$. Hence, from (3.11) and (3.9) we obtain

$$\tilde{\zeta} = d \frac{y_1}{y_2}, \quad \eta = d \frac{|\mathbf{y}|^2 - 1}{2y_2}, \quad (3.20)$$

using a north pole in the stereographic projections, since for a reflector we assume that the light rays are not reflected upwards. Substituting (3.20) into (3.19), we can calculate $\tilde{g}(\mathbf{y})$. To determine the boundary $\partial\mathcal{Y}$ we use (3.10) and (3.8) to get

$$y_1 = \frac{t_1}{1 - t_3} = \frac{\tilde{\zeta}}{\sqrt{\tilde{\zeta}^2 + \eta^2 + d^2}} \frac{1}{1 - \frac{\eta}{\sqrt{\tilde{\zeta}^2 + \eta^2 + d^2}}} = \frac{\tilde{\zeta}}{\sqrt{\tilde{\zeta}^2 + \eta^2 + d^2} - \eta}, \quad (3.21a)$$

$$y_2 = \frac{t_2}{1 - t_3} = \frac{d}{\sqrt{\tilde{\zeta}^2 + \eta^2 + d^2}} \frac{1}{1 - \frac{\eta}{\sqrt{\tilde{\zeta}^2 + \eta^2 + d^2}}} = \frac{d}{\sqrt{\tilde{\zeta}^2 + \eta^2 + d^2} - \eta}, \quad (3.21b)$$

and calculate $\mathbf{y} = (y_1, y_2)$ for all $(\tilde{\zeta}, \eta)$ that form the boundary of the specified target projection on screen P .

The above derivations were performed for the point-to-far-field reflector as displayed in Figure 3.5. If we instead consider a lens system where the light rays are refracted mainly in the upward z -direction, it would be handy to put the target plane P parallel to the xy -plane at some $z = \ell > 0$. In this case, we have $\hat{\mathbf{e}}_{\tilde{\zeta}} = (1, 0, 0)$, $\hat{\mathbf{e}}_{\eta} = (0, 1, 0)$ and $\hat{\mathbf{n}}_P = (0, 0, -1)$. Hence, from (3.11) and (3.9) we obtain

$$\tilde{\zeta} = d \frac{2y_1}{1 - |\mathbf{y}|^2}, \quad \eta = d \frac{2y_2}{1 - |\mathbf{y}|^2}, \quad (3.22)$$

using a south pole and plus sign in the stereographic projections, since we assume that a lens does not refract rays in the downwards z -direction. Substituting (3.22) into (3.19), we can calculate $\tilde{g}(\mathbf{y})$. To determine the boundary $\partial\mathcal{Y}$

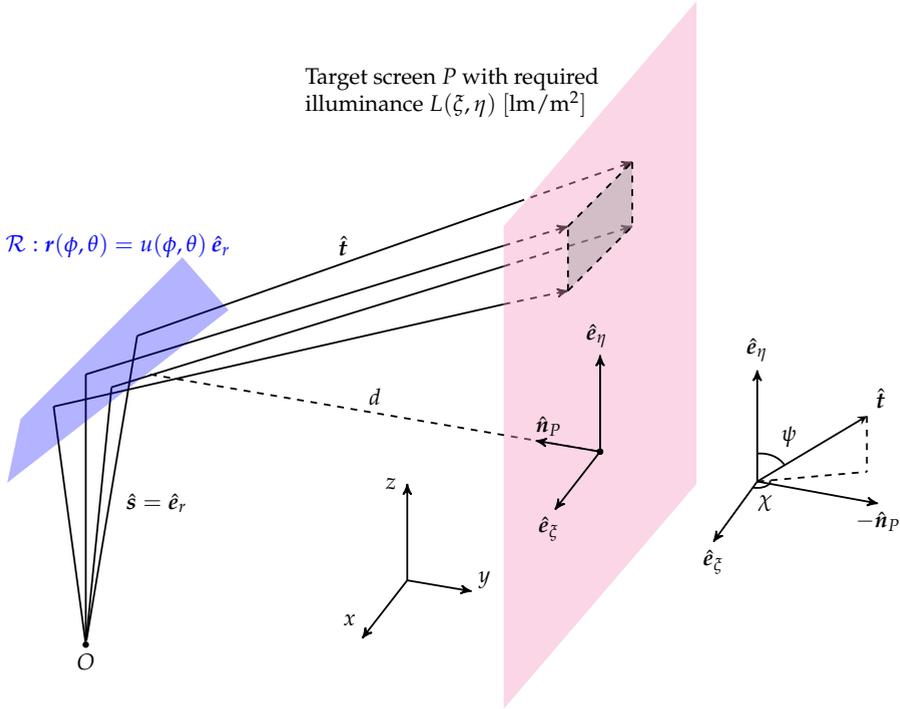


Figure 3.5: A target screen P in the xz -plane at distance d for a point-to-far-field reflector.

we use (3.10) and (3.8) to get

$$y_1 = \frac{t_1}{1 + t_3} = \frac{\xi}{\sqrt{\xi^2 + \eta^2 + d^2}} \frac{1}{1 + \frac{d}{\sqrt{\xi^2 + \eta^2 + d^2}}} = \frac{\xi}{\sqrt{\xi^2 + \eta^2 + d^2} + d}, \quad (3.23a)$$

$$y_2 = \frac{t_2}{1 + t_3} = \frac{\eta}{\sqrt{\xi^2 + \eta^2 + d^2}} \frac{1}{1 + \frac{d}{\sqrt{\xi^2 + \eta^2 + d^2}}} = \frac{\eta}{\sqrt{\xi^2 + \eta^2 + d^2} + d}, \quad (3.23b)$$

and calculate $\mathbf{y} = (y_1, y_2)$ for all (ξ, η) that form the boundary of the specified target projection on screen P .

3.2 Parallel-to-far-field reflector

In this section we consider a reflector with a parallel source and a far-field target. We will derive Hamilton's mixed characteristic $W = W(z_s, z_t, \mathbf{q}_s, \mathbf{p}_t)$, dependent on the source coordinates $\mathbf{x} = \mathbf{q}_s$ and target momentum \mathbf{p}_t . We consider the source plane $z = z_s = 0$ and target plane $z = z_t = -L$, i.e., the

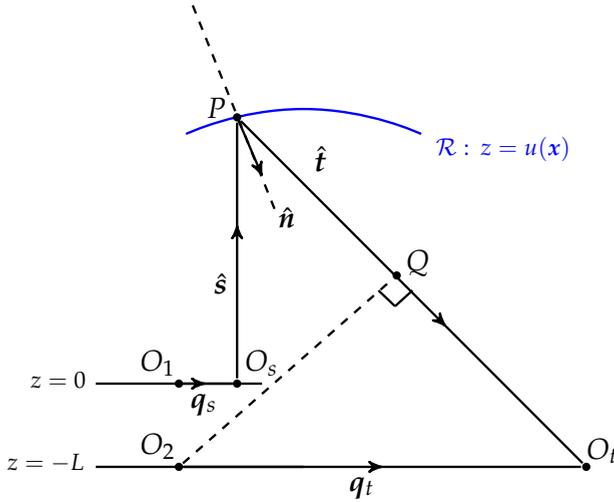


Figure 3.6: Illustration of the derivation of Hamilton's mixed characteristic W for a parallel source and far-field target.

target plane is located a distance L away from the source plane. The index of refraction n equals 1 everywhere.

The position and direction coordinates of the ray \hat{s} are given by the two-vectors $\mathbf{q}_s = \mathbf{x}$ and $\mathbf{p}_s = \mathbf{0}$, respectively. The position and direction coordinates of the ray \hat{t} are given by \mathbf{q}_t and $\mathbf{p}_t = (t_1, t_2)$, respectively. The point where the ray \hat{s} hits the reflector is given by $P(\mathbf{x}, u(\mathbf{x}))$, expressed using its position vector $(\mathbf{x}, u(\mathbf{x}))$, as shown schematically in Figure 3.6.

The mixed characteristic $W(\mathbf{p}_t)$

The point characteristic between a point $O_s(\mathbf{x}, 0)$ on the source plane and $O_t(\mathbf{q}_t, -L)$ on the target plane is given by

$$V(\mathbf{q}_s, \mathbf{q}_t) = u(\mathbf{x}) + d(P, O_t), \quad (3.24)$$

where the optical path length between P and O_t is

$$d(P, O_t) = \sqrt{|\mathbf{q}_t - \mathbf{x}|^2 + (u(\mathbf{x}) + L)^2}, \quad (3.25)$$

and $u(\mathbf{x})$ is the optical path length from $(\mathbf{x}, 0)$ to P . Both optical path lengths are equal to the Euclidean distances since $n = 1$. The mixed characteristic W

is given by the relation

$$\begin{aligned} W(\mathbf{q}_s, \mathbf{p}_t) &= V(\mathbf{q}_s, \mathbf{q}_t) - \mathbf{q}_t \cdot \mathbf{p}_t \\ &= u(\mathbf{x}) + \sqrt{|\mathbf{q}_t - \mathbf{x}|^2 + (L + u(\mathbf{x}))^2} - \mathbf{q}_t \cdot \mathbf{p}_t. \end{aligned} \quad (3.26)$$

We can show that the mixed characteristic $W = W(\mathbf{p}_t)$ is independent of \mathbf{q}_s since $\mathbf{p}_s = \mathbf{0}$ using the results in Section 2.7.6, summarized in Figure 3.1 (and Figure 3.2). Similar results hold for the other optical systems in this chapter and we will simply refer to Figure 3.1 and 3.2.

Using Figure 3.6 we find that

$$\mathbf{p}_t = \frac{\mathbf{q}_t - \mathbf{x}}{d(P, O_t)}, \quad t_3 = \frac{-L - u(\mathbf{x})}{d(P, O_t)}, \quad (3.27)$$

and consequently,

$$\begin{aligned} W(\mathbf{p}_t) &= u(\mathbf{x}) + \frac{1}{d(P, O_t)} \times \left(|\mathbf{q}_t - \mathbf{x}|^2 - \mathbf{q}_t \cdot (\mathbf{q}_t - \mathbf{x}) + (L + u(\mathbf{x}))^2 \right) \\ &= u(\mathbf{x}) - \mathbf{x} \cdot \mathbf{p}_t - u(\mathbf{x})t_3 - L t_3 \\ &= u(\mathbf{x})(1 - t_3) - x_1 t_1 - x_2 t_2 - L t_3, \end{aligned} \quad (3.28)$$

using (3.27) for the second equal sign. We can rewrite the above expression as

$$\frac{W(\mathbf{p}_t) + L t_3}{1 - t_3} = u(\mathbf{x}) - \frac{x_1 t_1 + x_2 t_2}{1 - t_3}, \quad (3.29)$$

where the left-hand side only depends on $\hat{\mathbf{t}}$, and hence on the stereographic coordinates \mathbf{y} . Note that the numerator can be rewritten using the results on the interpretation of the characteristic functions from the previous chapter in Section 2.7.5, Equation (2.132), as

$$W(\mathbf{p}_t) + L t_3 = u(\mathbf{x}) + d(P, Q) + L t_3 = V(\mathbf{q}_s, \mathbf{q}_t) - \mathbf{r}_{O_t} \cdot \hat{\mathbf{t}}, \quad (3.30)$$

with \mathbf{r}_{O_t} the position vector of O_t . Moreover, straightforward differentiation and using (3.28) gives

$$\frac{\partial(W(\mathbf{p}_t) + L t_3)}{\partial L} = 0, \quad (3.31)$$

i.e., we can choose the location of the target plane $z = -L$ wherever we like. For the remaining optical systems considered in this chapter, we can perform a similar analysis. For this reason, we set $L = 0$ and always choose the target plane $z = z_t = z_s$ to coincide with the source plane $z = z_s$.

The cost function

The reflected direction $\hat{\mathbf{t}}$ can be thought of as a mapping $\hat{\mathbf{s}} \mapsto \hat{\mathbf{t}}$ from the source domain \mathcal{X} to the unit sphere \mathbb{S}^2 . We use the stereographic coordinate transformation in (3.8) to obtain $\mathbf{y}(\hat{\mathbf{t}})$ using a north pole and corresponding minus sign, since we assume a reflector does not reflect light rays in the upward z -direction. Consequently, (3.29) leads to

$$u(\mathbf{x}) = \mathbf{x} \cdot \mathbf{y} + w(\mathbf{y}), \quad (3.32)$$

where $w(\mathbf{y}) = \frac{W(\mathbf{p}_t) + L t_3}{1 - t_3}$ is a function of \mathbf{y} by transforming $\frac{W(\mathbf{p}_t) + L t_3}{1 - t_3}$ to stereographic coordinates (the specific form is not important in subsequent derivations). Defining $u_1(\mathbf{x}) = u(\mathbf{x})$ and $u_2(\mathbf{y}) = w(\mathbf{y})$ we obtain the optimal transport formulation

$$u_2(\mathbf{y}) - u_1(\mathbf{x}) = -\mathbf{x} \cdot \mathbf{y} = c(\mathbf{x}, \mathbf{y}), \quad (3.33)$$

in accordance with (3.1).

We can find another optimal-transport relation, which is frequently encountered in literature, by performing a set of substitutions. The scalar product satisfies

$$\mathbf{x} \cdot \mathbf{y} = \frac{1}{2}(|\mathbf{x}|^2 + |\mathbf{y}|^2 - |\mathbf{x} - \mathbf{y}|^2), \quad (3.34)$$

and using the auxiliary functions

$$u_1(\mathbf{x}) = u(\mathbf{x}) - \frac{1}{2}|\mathbf{x}|^2, \quad u_2(\mathbf{y}) = w(\mathbf{y}) + \frac{1}{2}|\mathbf{y}|^2, \quad (3.35)$$

we obtain

$$u_2(\mathbf{y}) - u_1(\mathbf{x}) = -\frac{1}{2}|\mathbf{x} - \mathbf{y}|^2 = \tilde{c}(\mathbf{x}, \mathbf{y}). \quad (3.36)$$

This cost function more frequently appears in optimal-transport literature than (3.33). This literature will be discussed in Chapter 4. Both cost functions $c(\mathbf{x}, \mathbf{y}) = -\mathbf{x} \cdot \mathbf{y}$ and $\tilde{c}(\mathbf{x}, \mathbf{y}) = -\frac{1}{2}|\mathbf{x} - \mathbf{y}|^2$ are quadratic and they can be used to describe this optical system.

The generating function

Next, we construct the generating function $G(\mathbf{x}, \mathbf{y}, w) = u(\mathbf{x})$ using (3.32) as

$$u(\mathbf{x}) = G(\mathbf{x}, \mathbf{y}, w) = \mathbf{x} \cdot \mathbf{y} + w, \quad (3.37)$$

where $w = w(\mathbf{y})$ is a function of \mathbf{y} . For this system, the inverse function H is simply the function $w(\mathbf{y})$, which is related to the characteristic function $W(\mathbf{p}_t)$.

We denote the inverse $w(\mathbf{y}) = H(\mathbf{x}, \mathbf{y}, G(\mathbf{x}, \mathbf{y}, w(\mathbf{y})))$ as

$$H(\mathbf{x}, \mathbf{y}, w) = -\mathbf{x} \cdot \mathbf{y} + w, \quad (3.38)$$

where $w = u(\mathbf{x})$ is now a function of \mathbf{x} .

The mapping

We can derive the mapping using the vectorial law of reflection. We first write the downward normal vector to the reflector as

$$\mathbf{n} = \begin{pmatrix} u_{x_1} \\ u_{x_2} \\ -1 \end{pmatrix}, \quad \hat{\mathbf{n}} = \frac{\mathbf{n}}{|\mathbf{n}|}. \quad (3.39)$$

The direction $\hat{\mathbf{t}}$ can be found directly from the vectorial law of reflection (2.39) as

$$\hat{\mathbf{t}} = \frac{1}{|\nabla u|^2 + 1} \begin{pmatrix} 2u_{x_1} \\ 2u_{x_2} \\ |\nabla u|^2 - 1 \end{pmatrix}, \quad (3.40)$$

and comparing this expression with the stereographic projection (3.9) using a north pole and minus sign, we find that

$$\mathbf{y} = \mathbf{m}(\mathbf{x}) = \nabla u(\mathbf{x}), \quad (3.41)$$

which is a mapping from the Cartesian source coordinates $\mathbf{x} \in \mathcal{X}$ to the stereographic target coordinates $\mathbf{y} \in \mathcal{Y}$. This mapping connects a coordinate on the source $\mathbf{x} \in \mathcal{X}$ with a coordinate on the target $\mathbf{y} \in \mathcal{Y}$ and represents the trajectory of a light ray.

Energy conservation

By transferring the light from source to target we require that all light from the source ends up at the target so that energy is conserved, i.e.,

$$\int_{\mathcal{A}} f(\mathbf{x}) \, d\mathbf{x} = \int_{\hat{\mathbf{t}}(\mathcal{A})} g(\psi, \chi) \, d\mathcal{S}(\psi, \chi), \quad (3.42)$$

for an arbitrary set $\mathcal{A} \subset \mathcal{X}$ and image set $\hat{\mathbf{t}}(\mathcal{A}) \subset \mathcal{S}^2$. Note that this image set corresponds to the far-field approximation. If we substitute $\hat{\mathbf{t}} = \hat{\mathbf{t}}(\mathbf{y})$ from (3.9) we can write (3.42) as

$$\int_{\mathcal{A}} f(\mathbf{x}) \, d\mathbf{x} = \int_{\mathbf{y}(\hat{\mathbf{t}}(\mathcal{A}))} \tilde{g}(\mathbf{y}) \left| \frac{\partial \hat{\mathbf{t}}}{\partial y_1} \times \frac{\partial \hat{\mathbf{t}}}{\partial y_2} \right| \, d\mathbf{y}, \quad (3.43)$$

where $\tilde{g}(\mathbf{y}) = g(\psi, \chi)$ and $\mathbf{y}(\hat{\mathbf{t}}(\mathcal{A}))$ denotes the stereographic projection of the set \mathcal{A} under the mapping $\hat{\mathbf{t}}$, cf. (3.8). Note that for global energy conservation we choose $\mathcal{A} = \mathcal{X}$ and $\mathbf{y}(\hat{\mathbf{t}}(\mathcal{A})) = \mathcal{Y}$. We calculate the Jacobian of the coordinate transformation to stereographic coordinates at the target as

$$\left| \frac{\partial \hat{\mathbf{t}}}{\partial y_1} \times \frac{\partial \hat{\mathbf{t}}}{\partial y_2} \right| = \frac{4}{(1 + |\mathbf{y}|^2)^2}. \quad (3.44)$$

Substituting (3.44) and the mapping $\mathbf{y} = \mathbf{m}(\mathbf{x})$ into the energy conservation relation (3.43) gives

$$\int_{\mathcal{A}} f(\mathbf{x}) \, d\mathbf{x} = \int_{\mathcal{A}} \tilde{g}(\mathbf{m}(\mathbf{x})) \frac{4}{(1 + |\mathbf{m}(\mathbf{x})|^2)^2} \det(D\mathbf{m}(\mathbf{x})) \, d\mathbf{x}, \quad (3.45)$$

where we omit the absolute value sign of the determinant and restrict ourselves to a positive Jacobian $\det(D\mathbf{m})$ of the mapping. Since (3.45) holds for every $\mathcal{A} \subseteq \mathcal{X}$, pointwise it follows that

$$\det(D\mathbf{m}(\mathbf{x})) = \frac{1}{4} (1 + |\mathbf{m}(\mathbf{x})|^2)^2 \frac{f(\mathbf{x})}{\tilde{g}(\mathbf{m}(\mathbf{x}))}, \quad (3.46)$$

pointwise. Substituting the mapping $\mathbf{m} = \nabla u$ gives the *standard Monge-Ampère equation*

$$\det(D^2u) = \frac{1}{4} (1 + |\nabla u(\mathbf{x})|^2)^2 \frac{f(\mathbf{x})}{\tilde{g}(\nabla u(\mathbf{x}))}, \quad (3.47)$$

where D^2u is the Hessian matrix containing the second-order partial derivatives with respect to \mathbf{x} .

We define the corresponding transport boundary condition as

$$m(\partial\mathcal{X}) = \partial\mathcal{Y}, \quad (3.48)$$

i.e., $\nabla u(\partial\mathcal{X}) = \partial\mathcal{Y}$, stating that all light from the boundary of the source \mathcal{X} is mapped to the boundary of the target \mathcal{Y} . In Section 4.5 we will show that the transport boundary condition follows from the implicit boundary condition $m(\mathcal{X}) = \mathcal{Y}$, stating that all the light from the source \mathcal{X} must be transferred to the target domain \mathcal{Y} . The equivalence of the boundary conditions follows from the edge-ray principle [128] and convexity of the optical surface.

For the parallel-to-far-field reflector problem we can construct the quadratic optimal-transport cost functions

$$c(\mathbf{x}, \mathbf{y}) = -\mathbf{x} \cdot \mathbf{y}, \quad \text{or} \quad \tilde{c}(\mathbf{x}, \mathbf{y}) = -\frac{1}{2}|\mathbf{x} - \mathbf{y}|^2, \quad (3.49)$$

the generating function

$$G(\mathbf{x}, \mathbf{y}, w) = \mathbf{x} \cdot \mathbf{y} + w, \quad (3.50)$$

with corresponding inverse

$$H(\mathbf{x}, \mathbf{y}, w) = -\mathbf{x} \cdot \mathbf{y} + w, \quad (3.51)$$

and mapping

$$\mathbf{y} = \mathbf{m}(\mathbf{x}) = \nabla u(\mathbf{x}). \quad (3.52)$$

*Combining the mapping with energy conservation gives the **standard Monge-Ampère equation***

$$\det(D^2u) = \frac{1}{4} (1 + |\nabla u(\mathbf{x})|^2)^2 \frac{f(\mathbf{x})}{\tilde{g}(\nabla u(\mathbf{x}))}. \quad (3.53)$$

The standard Monge-Ampère equation can also be used for a lens system. For details we refer to [161, p. 38–41].

3.3 Parallel-to-near-field reflector

We now consider a parallel source, a reflector and a near-field target. The position and direction coordinates on the source plane are given by the two-vectors $\mathbf{q}_s = \mathbf{x}$ and $\mathbf{p}_s = \mathbf{0}$, respectively. The position and direction coordinates on the target plane are given by $\mathbf{q}_t = \mathbf{y}$ and $\mathbf{p}_t = (t_1, t_2)$, respectively. The point

where the ray hits the reflector is given by $P(x, u(x))$, as shown schematically in Figure 3.7.

The point characteristic $V(q_t)$

We choose the source plane to coincide with the target plane. The point characteristic between a point $O_s(x, 0)$ on the source plane and $O_t(y, 0)$ on the target plane is given by

$$\begin{aligned} V(q_s, q_t) &= u(x) + d(P, O_t) \\ &= u(x) + \sqrt{|y - x|^2 + u(x)^2}, \end{aligned} \quad (3.54)$$

where $u(x)$ is the optical path length from $(x, 0)$ to P , and $d(P, O_t)$ denotes the Euclidean distance between P and O_t . Using Figure 3.1, we know that the point characteristic V is independent of the position coordinate q_s . Hence, using $y = q_t$ we write

$$V(y) = u(x) + \sqrt{|y - x|^2 + u(x)^2}. \quad (3.55)$$

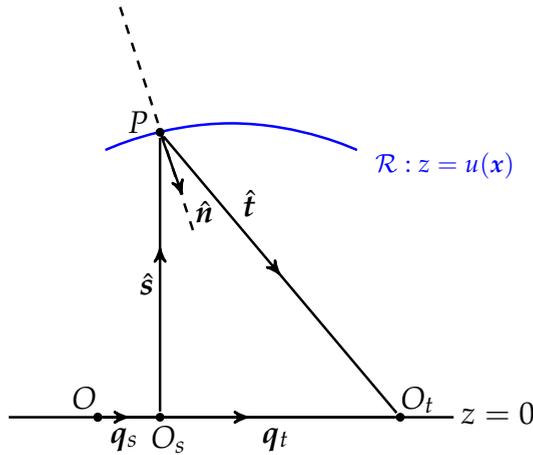


Figure 3.7: Illustration for the derivation of Hamilton's point characteristic V for a parallel source and near-field target.

The cost function

We cannot change variables in (3.55) and define a $u_1(x)$ and $u_2(y)$ such that we obtain a relation of the form (3.1) with a cost function.

The generating function

Solving (3.55) for $u(\mathbf{x})$ we obtain

$$u(\mathbf{x}) = \frac{1}{2} V(\mathbf{y}) - \frac{1}{2V(\mathbf{y})} |\mathbf{x} - \mathbf{y}|^2. \quad (3.56)$$

We construct a generating function $G(\mathbf{x}, \mathbf{y}, w) = u(\mathbf{x})$ by letting $w(\mathbf{y}) = \frac{1}{V(\mathbf{y})}$, in order to find the same generating functions as presented in [145], as

$$u(\mathbf{x}) = G(\mathbf{x}, \mathbf{y}, w) = \frac{1}{2w} - \frac{w}{2} |\mathbf{x} - \mathbf{y}|^2, \quad (3.57)$$

where $w(\mathbf{y}) = \frac{1}{V(\mathbf{y})}$ is a function of \mathbf{y} .

The function H is the reciprocal value of the point characteristic $V(\mathbf{y})$ rewritten in stereographic coordinates. Using (3.55) we find that

$$H(\mathbf{x}, \mathbf{y}, w) = \left(w + \sqrt{|\mathbf{y} - \mathbf{x}|^2 + w^2} \right)^{-1}, \quad (3.58)$$

where $w = u(\mathbf{x})$ is now a function of \mathbf{x} .

The mapping

The normal to the reflector surface can be written as in (3.39) for the parallel-to-far-field case. Subsequently, we wrote the outgoing direction $\hat{\mathbf{t}}$ as in (3.40) using the normal (3.39). For the near-field problem we can calculate $\mathbf{y} = \mathbf{m}(\mathbf{x})$ by solving the system

$$\mathbf{y} = \mathbf{x} - d \mathbf{p}_t, \quad (3.59a)$$

$$0 = u(\mathbf{x}) - d t_3, \quad (3.59b)$$

where $d = d(P, O_t)$ is the distance from P to O_t . Using $\mathbf{p}_t = (t_1, t_2)$ and substituting (3.40) into (3.59) gives

$$\mathbf{y} = \mathbf{x} - d \frac{2 \nabla u}{|\nabla u|^2 + 1}, \quad (3.60a)$$

$$0 = u(\mathbf{x}) - d \frac{|\nabla u|^2 - 1}{|\nabla u|^2 + 1}. \quad (3.60b)$$

Solving (3.60b) for d and substituting into (3.60a) gives

$$\mathbf{y} = \mathbf{m}(\mathbf{x}) = \mathbf{x} + \frac{2u \nabla u}{1 - |\nabla u|^2}, \quad (3.61)$$

under the condition that $|\nabla u|^2 \neq 1$.

Energy conservation

We have the energy balance

$$\int_{\mathcal{A}} f(\mathbf{x}) \, d\mathbf{x} = \int_{\mathbf{y}(\mathcal{A})} g(\mathbf{y}) \, d\mathbf{y}, \quad (3.62)$$

for an arbitrary set $\mathcal{A} \subset \mathcal{X}$ and image set $\mathbf{y}(\mathcal{A}) \subset \mathcal{Y}$. Substituting the mapping $\mathbf{y} = \mathbf{m}(\mathbf{x})$ gives

$$\int_{\mathcal{A}} f(\mathbf{x}) \, d\mathbf{x} = \int_{\mathcal{A}} g(\mathbf{m}(\mathbf{x})) |\det(D\mathbf{m}(\mathbf{x}))| \, d\mathbf{x}. \quad (3.63)$$

Analogous to the standard Monge-Ampère equation, we can use (3.63) to find the Jacobian equation

$$\det(D\mathbf{m}(\mathbf{x})) = \frac{f(\mathbf{x})}{g(\mathbf{m}(\mathbf{x}))}, \quad (3.64)$$

assuming $\det(D\mathbf{m}) > 0$. Substituting the mapping in (3.61) into this equation gives a second-order nonlinear PDE for u . In Chapter 4, we will show that this second-order nonlinear PDE can be written as a generated Jacobian equation.

As in the previous section, the corresponding transport boundary condition is defined as $\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}$.

For the parallel-to-near-field reflector problem we cannot construct an optimal-transport cost function. The generating function is

$$G(\mathbf{x}, \mathbf{y}, w) = \frac{1}{2w} - \frac{w}{2} |\mathbf{x} - \mathbf{y}|^2, \quad (3.65)$$

with corresponding inverse

$$H(\mathbf{x}, \mathbf{y}, w) = \left(w + \sqrt{|\mathbf{y} - \mathbf{x}|^2 + w^2} \right)^{-1}, \quad (3.66)$$

and mapping

$$\mathbf{y} = \mathbf{m}(\mathbf{x}) = \mathbf{x} + \frac{2u \nabla u}{1 - |\nabla u|^2}. \quad (3.67)$$

Combining the mapping with energy conservation gives the Jacobian equation

$$\det(D\mathbf{m}(\mathbf{x})) = \frac{f(\mathbf{x})}{g(\mathbf{m}(\mathbf{x}))}. \quad (3.68)$$

3.4 Point-to-far-field reflector

We now consider a point source, a reflector surface and a far-field target. The position and direction coordinate vectors on the source plane are given by the two-vectors $\mathbf{q}_s = \mathbf{0}$ and $\mathbf{p}_s = (s_1, s_2)$, respectively. The position and direction coordinates on the target plane are given by \mathbf{q}_t and $\mathbf{p}_t = (t_1, t_2)$, respectively. We write the radial parameter $u(\phi, \theta)$ as $u(\hat{\mathbf{s}})$. The point where the ray hits the reflector is given by $P(u(\hat{\mathbf{s}}) \mathbf{p}_s, u(\hat{\mathbf{s}}) s_3)$, as shown schematically in Figure 3.8.

The angular characteristic $T(\mathbf{p}_t)$

The point characteristic between point $O_s(\mathbf{q}_s, 0)$, which is the same point as the origin O , and $O_t(\mathbf{q}_t, 0)$ is given by

$$\begin{aligned} V(\mathbf{q}_s, \mathbf{q}_t) &= u(\hat{\mathbf{s}}) + d(P, O_t) \\ &= u(\hat{\mathbf{s}}) + \sqrt{|\mathbf{q}_t - u(\hat{\mathbf{s}}) \mathbf{p}_s|^2 + (u(\hat{\mathbf{s}}) s_3)^2}, \end{aligned} \quad (3.69)$$

where $u(\hat{\mathbf{s}})$ is the optical path length from O to P , and $d(P, O_t)$ denotes the Euclidean distance between P and O_t .

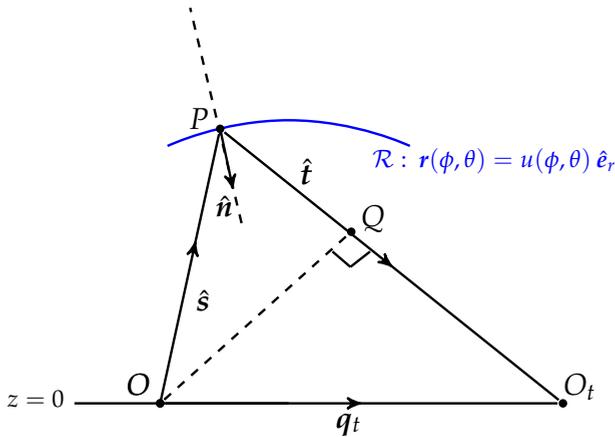


Figure 3.8: Illustration for the derivation of Hamilton's angular characteristic T for a point source and far-field target.

Hamilton's angular characteristic, which depends on the direction of the ray at the source plane and the direction at the target plane, is given by

$$T(\mathbf{p}_s, \mathbf{p}_t) = V(\mathbf{q}_s, \mathbf{q}_t) + \mathbf{q}_s \cdot \mathbf{p}_s - \mathbf{q}_t \cdot \mathbf{p}_t. \quad (3.70)$$

Figure 3.1 shows that the angular characteristic T is independent of the direction coordinate \mathbf{p}_s , and the expression for T reads

$$\begin{aligned} T(\mathbf{p}_t) &= V(\mathbf{q}_s, \mathbf{q}_t) - \mathbf{q}_t \cdot \mathbf{p}_t \\ &= u(\hat{\mathbf{s}}) + \sqrt{|\mathbf{q}_t - u(\hat{\mathbf{s}}) \mathbf{p}_s|^2 + (u(\hat{\mathbf{s}}) s_3)^2} - \mathbf{q}_t \cdot \mathbf{p}_t. \end{aligned} \quad (3.71)$$

Using Figure 3.8 we have that

$$\mathbf{p}_t = \frac{\mathbf{q}_t - u(\hat{\mathbf{s}}) \mathbf{p}_s}{d(P, O_t)}, \quad t_3 = -\frac{u(\hat{\mathbf{s}}) s_3}{d(P, O_t)}, \quad (3.72)$$

and consequently,

$$\begin{aligned} T(\mathbf{p}_t) &= u(\hat{\mathbf{s}}) + \frac{1}{d(P, O_t)} \left[|\mathbf{q}_t - u(\hat{\mathbf{s}}) \mathbf{p}_s|^2 - \mathbf{q}_t \cdot (\mathbf{q}_t - u(\hat{\mathbf{s}}) \mathbf{p}_s) + (u(\hat{\mathbf{s}}) s_3)^2 \right] \\ &= u(\hat{\mathbf{s}}) - u(\hat{\mathbf{s}}) (\mathbf{p}_t \cdot \mathbf{p}_s + s_3 t_3). \end{aligned} \quad (3.73)$$

Hence, we arrive at

$$T(\mathbf{p}_t) = u(\hat{\mathbf{s}})(1 - \hat{\mathbf{s}} \cdot \hat{\mathbf{t}}). \quad (3.74)$$

The cost function

In (3.74) we have that $u(\hat{\mathbf{s}}) > 0$ and $1 - \hat{\mathbf{s}} \cdot \hat{\mathbf{t}} > 0$. Taking the logarithm and defining the new functions $\tilde{u}_1(\hat{\mathbf{s}}) = -\log u(\hat{\mathbf{s}})$ and $\tilde{u}_2(\hat{\mathbf{t}}) = -\log(T(\mathbf{p}_t))$ results in the relation

$$\tilde{u}_2(\hat{\mathbf{t}}) - \tilde{u}_1(\hat{\mathbf{s}}) = -\log(1 - \hat{\mathbf{s}} \cdot \hat{\mathbf{t}}) = \tilde{c}(\hat{\mathbf{s}}, \hat{\mathbf{t}}). \quad (3.75)$$

Changing to stereographic coordinates and using the auxiliary functions

$$u_1(\mathbf{x}) = \tilde{u}_1(\hat{\mathbf{s}}) + \log(1 + |\mathbf{x}|^2), \quad u_2(\mathbf{y}) = \tilde{u}_2(\hat{\mathbf{t}}) + \log\left(\frac{2}{1 + |\mathbf{y}|^2}\right), \quad (3.76)$$

we arrive at the relation

$$u_2(\mathbf{y}) - u_1(\mathbf{x}) = -\log(N(\mathbf{x}, \mathbf{y})) = c(\mathbf{x}, \mathbf{y}), \quad (3.77a)$$

where

$$N(\mathbf{x}, \mathbf{y}) = 1 - 2 \mathbf{x} \cdot \mathbf{y} + |\mathbf{x}|^2 |\mathbf{y}|^2. \quad (3.77b)$$

The generating function

Solving (3.74) for $u(\hat{\mathbf{s}})$ we obtain

$$u(\hat{\mathbf{s}}) = \frac{T(\mathbf{p}_t)}{1 - \hat{\mathbf{s}} \cdot \hat{\mathbf{t}}}. \quad (3.78)$$

Changing to stereographic coordinates for $\hat{\mathbf{s}}$ and $\hat{\mathbf{t}}$ using (3.9), with a north pole and minus sign for $\hat{\mathbf{t}}$, gives

$$u(\mathbf{x}) = \frac{T(\mathbf{p}_t) (1 + |\mathbf{x}|^2) (1 + |\mathbf{y}|^2)}{2 (1 - 2 \mathbf{x} \cdot \mathbf{y} + |\mathbf{x}|^2 |\mathbf{y}|^2)}. \quad (3.79)$$

Now, we construct the generating function $G(\mathbf{x}, \mathbf{y}, w) = u$ with $w = T(\mathbf{p}_t)$ as

$$u(\mathbf{x}) = G(\mathbf{x}, \mathbf{y}, w) = \frac{w (1 + |\mathbf{x}|^2) (1 + |\mathbf{y}|^2)}{2 (1 - 2 \mathbf{x} \cdot \mathbf{y} + |\mathbf{x}|^2 |\mathbf{y}|^2)}. \quad (3.80)$$

Note that $w = T(\mathbf{p}_t)$ depends on the outgoing ray $\hat{\mathbf{t}}$ and hence $w = w(\mathbf{y})$ is a function of \mathbf{y} . We can also simply write the generating function as, cf. (3.78),

$$u(\mathbf{x}) = G(\mathbf{x}, \mathbf{y}, w) = w (1 - \hat{\mathbf{s}}(\mathbf{x}) \cdot \hat{\mathbf{t}}(\mathbf{y}))^{-1}. \quad (3.81)$$

The function H is the angular characteristic $T(\mathbf{p}_t)$ rewritten in stereographic coordinates. Using (3.79) we find

$$H(\mathbf{x}, \mathbf{y}, w) = \frac{2 w (1 - 2 \mathbf{x} \cdot \mathbf{y} + |\mathbf{x}|^2 |\mathbf{y}|^2)}{(1 + |\mathbf{x}|^2) (1 + |\mathbf{y}|^2)}, \quad (3.82)$$

where $w = u(\mathbf{x})$ is now a function of \mathbf{x} .

The mapping

Deriving the mapping $\mathbf{y} = \mathbf{m}(\mathbf{x})$ is not straightforward but we can do it, so let's go! The mapping \mathbf{m} can be determined by tracing a typical ray through the optical system. We consider an incident ray propagating in the direction $\hat{\mathbf{s}} = \hat{\mathbf{e}}_r$, which intercepts the reflector \mathcal{R} and reflects off in direction $\hat{\mathbf{t}}$. The unit surface normal of the parametrized reflector surface $\mathbf{r}(\phi, \theta) = u(\phi, \theta) \hat{\mathbf{e}}_r$, directed towards the point source, is given by

$$\hat{\mathbf{n}} = \frac{\frac{\partial \mathbf{r}}{\partial \theta} \times \frac{\partial \mathbf{r}}{\partial \phi}}{\left| \frac{\partial \mathbf{r}}{\partial \theta} \times \frac{\partial \mathbf{r}}{\partial \phi} \right|} = \frac{-\hat{\mathbf{e}}_r + \nabla_r u}{\sqrt{1 + |\nabla_r u|^2}}, \quad (3.83a)$$

with

$$\nabla_r u = \frac{1}{u} \frac{\partial u}{\partial \phi} \hat{e}_\phi + \frac{1}{u \sin(\phi)} \frac{\partial u}{\partial \theta} \hat{e}_\theta, \quad (3.83b)$$

which is the gradient of u restricted to the surface $r = \text{constant}$. Using the vectorial law of reflection $\hat{t} = \hat{s} - 2(\hat{s} \cdot \hat{n})\hat{n}$ we obtain the direction \hat{t} of the reflected ray

$$\hat{t} = \hat{e}_r + \frac{2}{1 + |\nabla_r u|^2} (-\hat{e}_r + \nabla_r u). \quad (3.84)$$

First, let $u_1(x) = -\log(u(x)/(1 + |x|^2))$. We rewrite the gradient of u restricted to the surface $r = \text{constant}$ in (3.83b) to

$$\begin{aligned} \nabla_r u &= \nabla_r [e^{-u_1} (1 + |x|^2)] \\ &= \left(-\frac{\partial u_1}{\partial \phi} + \frac{2x_1}{1 + |x|^2} \frac{\partial x_1}{\partial \phi} + \frac{2x_2}{1 + |x|^2} \frac{\partial x_2}{\partial \phi} \right) \hat{e}_\phi \\ &\quad + \frac{1}{\sin(\phi)} \left(-\frac{\partial u_1}{\partial \theta} + \frac{2x_1}{1 + |x|^2} \frac{\partial x_1}{\partial \theta} + \frac{2x_2}{1 + |x|^2} \frac{\partial x_2}{\partial \theta} \right) \hat{e}_\theta \\ &= \left(-\frac{\partial u_1}{\partial \phi} + s_1 \frac{\partial x_1}{\partial \phi} + s_2 \frac{\partial x_2}{\partial \phi} \right) \hat{e}_\phi + \frac{1}{\sin(\phi)} \left(-\frac{\partial u_1}{\partial \theta} + s_1 \frac{\partial x_1}{\partial \theta} + s_2 \frac{\partial x_2}{\partial \theta} \right) \hat{e}_\theta, \end{aligned} \quad (3.85)$$

eliminating the term u in the denominator, cf. (3.83b). We write $u_1 = u_1(x)$ and apply the chain rule to the partial derivatives of u_1 to express the partial derivatives in terms of x , i.e.,

$$\frac{\partial u_1}{\partial \phi} = \frac{\partial u_1}{\partial x_1} \frac{\partial x_1}{\partial \phi} + \frac{\partial u_1}{\partial x_2} \frac{\partial x_2}{\partial \phi}, \quad \frac{\partial u_1}{\partial \theta} = \frac{\partial u_1}{\partial x_1} \frac{\partial x_1}{\partial \theta} + \frac{\partial u_1}{\partial x_2} \frac{\partial x_2}{\partial \theta}. \quad (3.86)$$

Using the definition of the stereographic coordinates of the source in Equation (3.7) we can rewrite the partial derivatives of x occurring in (3.85) and (3.86) in \hat{s} -coordinates as

$$\frac{\partial x_1}{\partial \phi} = \frac{\cos(\theta)}{1 + \cos(\phi)} = \frac{s_1}{(1 + s_3) \sqrt{s_1^2 + s_2^2}}, \quad \frac{\partial x_1}{\partial \theta} = -\frac{\sin(\phi) \sin(\theta)}{1 + \cos(\phi)} = -\frac{s_2}{1 + s_3}, \quad (3.87a)$$

$$\frac{\partial x_2}{\partial \phi} = \frac{\sin(\theta)}{1 + \cos(\phi)} = \frac{s_2}{(1 + s_3) \sqrt{s_1^2 + s_2^2}}, \quad \frac{\partial x_2}{\partial \theta} = \frac{\sin(\phi) \cos(\theta)}{1 + \cos(\phi)} = \frac{s_1}{1 + s_3}. \quad (3.87b)$$

Second, we can express the basis vectors \hat{e}_ϕ and \hat{e}_θ in \mathbf{x} -coordinates ($\hat{e}_r = \hat{\mathbf{s}}$, see (3.9)) as follows:

$$\begin{aligned}\hat{e}_\phi &= \begin{pmatrix} \cos(\phi) \cos(\theta) \\ \cos(\phi) \sin(\theta) \\ -\sin(\phi) \end{pmatrix} = \frac{1}{\sqrt{s_1^2 + s_2^2}} \begin{pmatrix} s_1 s_3 \\ s_2 s_3 \\ -s_1^2 - s_2^2 \end{pmatrix} \\ &= \frac{1}{|\mathbf{x}|(1 + |\mathbf{x}|^2)} \begin{pmatrix} x_1(1 - |\mathbf{x}|^2) \\ x_2(1 - |\mathbf{x}|^2) \\ -2|\mathbf{x}|^2 \end{pmatrix}, \\ \hat{e}_\theta &= \begin{pmatrix} -\sin(\theta) \\ \cos(\theta) \\ 0 \end{pmatrix} = \frac{1}{\sqrt{s_1^2 + s_2^2}} \begin{pmatrix} -s_2 \\ s_1 \\ 0 \end{pmatrix} = \frac{1}{|\mathbf{x}|} \begin{pmatrix} -x_2 \\ x_1 \\ 0 \end{pmatrix}.\end{aligned}\quad (3.87c)$$

Third, substituting $\sin(\phi) = \sqrt{s_1^2 + s_2^2} = 2|\mathbf{x}|/(1 + |\mathbf{x}|^2)$, (3.86) and (3.87) into (3.85) and using the inverse projection in (3.9), we get an expression for $\nabla_r u$ in terms of \mathbf{x} and the partial derivatives of u_1 with respect to \mathbf{x} , as

$$\nabla_r u = \frac{1}{2} \begin{pmatrix} -\frac{\partial u_1}{\partial x_1} (1 - x_1^2 + x_2^2) - 2x_1 + 2\frac{\partial u_1}{\partial x_2} x_1 x_2 \\ -\frac{\partial u_1}{\partial x_2} (1 + x_1^2 - x_2^2) - 2x_2 + 2\frac{\partial u_1}{\partial x_1} x_1 x_2 \\ -4 + 2\mathbf{x} \cdot \nabla u_1 \end{pmatrix} + \frac{2}{1 + |\mathbf{x}|^2} \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix}, \quad (3.88a)$$

where

$$\nabla u_1 = \begin{pmatrix} \frac{\partial u_1}{\partial x_1} \\ \frac{\partial u_1}{\partial x_2} \end{pmatrix}. \quad (3.88b)$$

Fourth, substituting this expression into (3.84) and applying the inverse stereographic projection in (3.9) to $\hat{e}_r = \hat{\mathbf{s}}$, we find the outgoing ray $\hat{\mathbf{t}}(\mathbf{x})$ expressed in \mathbf{x} -coordinates

$$\hat{\mathbf{t}}(\mathbf{x}) = \frac{1}{D(\mathbf{x})} \begin{pmatrix} -4\frac{\partial u_1}{\partial x_1} + 2|\nabla u_1|^2 x_1 \\ -4\frac{\partial u_1}{\partial x_2} + 2|\nabla u_1|^2 x_2 \\ 2|\nabla u_1|^2 - D(\mathbf{x}) \end{pmatrix}, \quad (3.89a)$$

where

$$D(\mathbf{x}) = 4 - 4\mathbf{x} \cdot \nabla u_1 + |\nabla u_1|^2 (1 + |\mathbf{x}|^2). \quad (3.89b)$$

Finally, transforming $\hat{\mathbf{t}}(\mathbf{x})$ to \mathbf{y} using the definition of the stereographic coordinates of the target in (3.8), we arrive at a compact expression for the

mapping $\mathbf{y} = \mathbf{m}(\mathbf{x})$

$$\mathbf{y} = \frac{-2 \nabla u_1 + \mathbf{x} |\nabla u_1|^2}{4 - 4 \mathbf{x} \cdot \nabla u_1 + (|\mathbf{x}| |\nabla u_1|)^2}. \quad (3.90)$$

Energy conservation

We require that all light from the source ends up at the target and that energy is conserved, i.e.,

$$\int_{\mathcal{A}} f(\phi, \theta) \, d\mathcal{S}(\phi, \theta) = \int_{\hat{\mathbf{t}}(\mathcal{A})} g(\psi, \chi) \, d\mathcal{S}(\psi, \chi), \quad (3.91)$$

for an arbitrary set $\mathcal{A} \subset S^2$ and image set $\hat{\mathbf{t}}(\mathcal{A}) \subset S^2$. If we substitute $\hat{\mathbf{s}} = \hat{\mathbf{s}}(\mathbf{x})$ and $\hat{\mathbf{t}} = \hat{\mathbf{t}}(\mathbf{y})$ from (3.9) we can write (3.91) as

$$\int_{\mathbf{x}(\mathcal{A})} \tilde{f}(\mathbf{x}) \left| \frac{\partial \hat{\mathbf{s}}}{\partial x_1} \times \frac{\partial \hat{\mathbf{s}}}{\partial x_2} \right| \, d\mathbf{x} = \int_{\mathbf{y}(\hat{\mathbf{t}}(\mathcal{A}))} \tilde{g}(\mathbf{y}) \left| \frac{\partial \hat{\mathbf{t}}}{\partial y_1} \times \frac{\partial \hat{\mathbf{t}}}{\partial y_2} \right| \, d\mathbf{y}. \quad (3.92)$$

The Jacobians of the coordinate transformations to stereographic coordinates are

$$\left| \frac{\partial \hat{\mathbf{s}}}{\partial x_1} \times \frac{\partial \hat{\mathbf{s}}}{\partial x_2} \right| = \frac{4}{(1 + |\mathbf{x}|^2)^2}, \quad \left| \frac{\partial \hat{\mathbf{t}}}{\partial y_1} \times \frac{\partial \hat{\mathbf{t}}}{\partial y_2} \right| = \frac{4}{(1 + |\mathbf{y}|^2)^2}. \quad (3.93)$$

Substituting (3.93) and the mapping $\mathbf{y} = \mathbf{m}(\mathbf{x})$ into the energy conservation relation (3.92) gives

$$\int_{\mathbf{x}(\mathcal{A})} \frac{4 \tilde{f}(\mathbf{x})}{(1 + |\mathbf{x}|^2)^2} \, d\mathbf{x} = \int_{\mathbf{x}(\mathcal{A})} \frac{4 \tilde{g}(\mathbf{m}(\mathbf{x}))}{(1 + |\mathbf{m}(\mathbf{x})|^2)^2} |\det(D\mathbf{m}(\mathbf{x}))| \, d\mathbf{x}. \quad (3.94)$$

We can rewrite (3.94) as the Jacobian equation

$$\det(D\mathbf{m}(\mathbf{x})) = \frac{(1 + |\mathbf{m}(\mathbf{x})|^2)^2}{(1 + |\mathbf{x}|^2)^2} \frac{\tilde{f}(\mathbf{x})}{\tilde{g}(\mathbf{m}(\mathbf{x}))}, \quad (3.95)$$

where we omit the absolute value sign of the determinant and restrict ourselves to a positive Jacobian of the mapping.

Finally, we note that the corresponding transport boundary condition is defined as $\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}$.

For the point-to-far-field reflector problem we can construct the non-quadratic, logarithmic, optimal-transport cost function

$$c(\mathbf{x}, \mathbf{y}) = -\log(1 - 2 \mathbf{x} \cdot \mathbf{y} + |\mathbf{x}|^2 |\mathbf{y}|^2), \quad (3.96)$$

the generating function

$$G(\mathbf{x}, \mathbf{y}, w) = \frac{w (1 + |\mathbf{x}|^2) (1 + |\mathbf{y}|^2)}{2 (1 - 2 \mathbf{x} \cdot \mathbf{y} + |\mathbf{x}|^2 |\mathbf{y}|^2)}, \quad (3.97)$$

with corresponding inverse

$$H(\mathbf{x}, \mathbf{y}, w) = \frac{2 w (1 - 2 \mathbf{x} \cdot \mathbf{y} + |\mathbf{x}|^2 |\mathbf{y}|^2)}{(1 + |\mathbf{x}|^2) (1 + |\mathbf{y}|^2)}, \quad (3.98)$$

and mapping

$$\mathbf{y} = \mathbf{m}(\mathbf{x}) = \frac{-2 \nabla u_1 + \mathbf{x} |\nabla u_1|^2}{4 - 4 \mathbf{x} \cdot \nabla u_1 + (|\mathbf{x}| |\nabla u_1|)^2}, \quad (3.99)$$

where ∇u_1 is the gradient of $u_1(\mathbf{x}) = \log(u(\mathbf{x}) / (1 + |\mathbf{x}|^2))$ with respect to \mathbf{x} . Combining the mapping with energy conservation gives the Jacobian equation

$$\det(\mathbf{Dm}(\mathbf{x})) = \frac{(1 + |\mathbf{m}(\mathbf{x})|^2)^2}{(1 + |\mathbf{x}|^2)^2} \frac{\tilde{f}(\mathbf{x})}{\tilde{g}(\mathbf{m}(\mathbf{x}))}. \quad (3.100)$$

3.5 Point-to-far-field lens

We consider a lens with refractive index $n > 1$, e.g., of glass or plastic material such that n is approximately between 1.3 and 1.7 [78, p. 103]. The first surface of the lens is spherical and it does not alter the direction of the incoming rays.

The position and direction coordinate vectors on the source plane are given by the two-vectors $\mathbf{q}_s = \mathbf{0}$ and $\mathbf{p}_s = (s_1, s_2)$, respectively. We consider the spherical surface to be positioned infinitesimally close to O , i.e., the optical path length between O and the point P , where the incoming ray hits the freeform surface of the lens, is equal to $n u(\hat{s})$. The point P has coordinates $P(u(\hat{s}) \mathbf{p}_s, u(\hat{s}) s_3)$, as shown schematically in Figure 3.9.

The position and direction coordinates on the target plane are given by \mathbf{q}_t and $\mathbf{p}_t = (t_1, t_2)$, respectively.

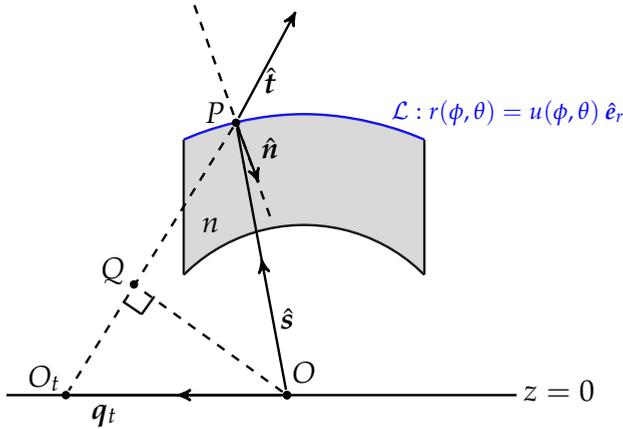


Figure 3.9: Illustration for the derivation of Hamilton's angular characteristic T for a point source and far-field target.

The angular characteristic $T(\mathbf{p}_t)$

The point characteristic (or optical path length) between point $O_s(\mathbf{q}_s, 0)$, which is the same point as the origin O , and the virtual point $O_t(\mathbf{q}_t, 0)$ is given by

$$\begin{aligned} V(\mathbf{q}_s, \mathbf{q}_t) &= n u(\hat{\mathbf{s}}) - d(P, O_t), \\ d(P, O_t) &= \sqrt{|\mathbf{q}_t - u(\hat{\mathbf{s}}) \mathbf{p}_s|^2 + (-u(\hat{\mathbf{s}}) s_3)^2}, \end{aligned} \quad (3.101)$$

where $n u(\hat{\mathbf{s}})$ is the optical path length from O to P , and $d(P, O_t)$ denotes the Euclidean distance between P and O_t . Note that the minus sign in front of $d(P, O_t)$ is a consequence of O_t being a virtual image point. The final portion of the ray travels in the opposite direction of $\overrightarrow{P O_t}$, as explained in Section 2.7.5.

Hamilton's angular characteristic, which depends on the direction of the ray at the source plane and the direction at the target plane, is given by

$$T(\mathbf{p}_s, \mathbf{p}_t) = V(\mathbf{q}_s, \mathbf{q}_t) + \mathbf{q}_s \cdot \mathbf{p}_s - \mathbf{q}_t \cdot \mathbf{p}_t. \quad (3.102)$$

Using Figure 3.2 the angular characteristic T is independent of the direction coordinate \mathbf{p}_s , and the expression for T reads

$$T(\mathbf{p}_t) = V(\mathbf{q}_s, \mathbf{q}_t) - \mathbf{q}_t \cdot \mathbf{p}_t = n u(\hat{\mathbf{s}}) - d(P, O_t) - \mathbf{q}_t \cdot \mathbf{p}_t. \quad (3.103)$$

Using Figure 3.9 we find

$$\mathbf{p}_t = -\frac{\mathbf{q}_t - u(\hat{\mathbf{s}}) \mathbf{p}_s}{d(P, O_t)}, \quad t_3 = \frac{u(\hat{\mathbf{s}}) s_3}{d(P, O_t)}, \quad (3.104)$$

and,

$$\begin{aligned}
 T(\mathbf{p}_t) &= n u(\hat{\mathbf{s}}) - \frac{1}{d(P, O_t)} \left[|\mathbf{q}_t - u(\hat{\mathbf{s}}) \mathbf{p}_s|^2 - \mathbf{q}_t \cdot (\mathbf{q}_t - u(\hat{\mathbf{s}}) \mathbf{p}_s) + (u(\hat{\mathbf{s}}) s_3)^2 \right] \\
 &= n u(\hat{\mathbf{s}}) - u(\hat{\mathbf{s}}) (\mathbf{p}_t \cdot \mathbf{p}_s + s_3 t_3) \\
 &= u(\hat{\mathbf{s}}) (n - \hat{\mathbf{s}} \cdot \hat{\mathbf{t}}).
 \end{aligned} \tag{3.105}$$

Note that $n - \hat{\mathbf{s}} \cdot \hat{\mathbf{t}} > 0$ for $n > 1$.

The cost function

Introducing $u_1(x) = -\log(u(x))$ and $-u_2(y) = \log(T(\mathbf{p}_t))$ we obtain

$$u_2(\mathbf{y}) - u_1(\mathbf{x}) = -\log \left(n - 1 + \frac{2|\mathbf{x} - \mathbf{y}|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}|^2)} \right) = c(\mathbf{x}, \mathbf{y}), \tag{3.106}$$

where $c(\mathbf{x}, \mathbf{y})$ is a logarithmic cost function in optimal transport theory of the stereographic coordinates $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$.

The generating function

Solving (3.105) for $u(\hat{\mathbf{s}})$ we obtain

$$u(\hat{\mathbf{s}}) = \frac{T(\mathbf{p}_t)}{n - \hat{\mathbf{s}} \cdot \hat{\mathbf{t}}}. \tag{3.107}$$

We change to stereographic coordinates using (3.9) and use a south pole and plus sign for \mathbf{y} , since we assume that the lens does not refract rays downwards. Using stereographic coordinates and the relation (3.34) we can rewrite (3.107) to

$$u(x) = T(\mathbf{p}_t) \left(n - 1 + \frac{2|\mathbf{x} - \mathbf{y}|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}|^2)} \right)^{-1}. \tag{3.108}$$

We construct the generating function from the relation $u(x) = G(x, \mathbf{y}, w)$ with $w = T(\mathbf{p}_t)$ as

$$u(x) = G(x, \mathbf{y}, w) = w \left(n - 1 + \frac{2|\mathbf{x} - \mathbf{y}|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}|^2)} \right)^{-1}. \tag{3.109}$$

Note that $w = T(\mathbf{p}_t)$ is dependent on the outgoing ray $\hat{\mathbf{t}}$ and hence $w = w(\mathbf{y})$ is a function of \mathbf{y} . We can also write the generating function as, cf. (3.107),

$$u(x) = G(x, \mathbf{y}, w) = w (n - \hat{\mathbf{s}}(x) \cdot \hat{\mathbf{t}}(\mathbf{y}))^{-1}. \tag{3.110}$$

The function H is the angular characteristic $T(\mathbf{p}_t)$ rewritten in stereographic coordinates, i.e.,

$$H(\mathbf{x}, \mathbf{y}, w) = w \left(n - 1 + \frac{2 |\mathbf{x} - \mathbf{y}|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}|^2)} \right). \quad (3.111)$$

The mapping

The mapping can be found by tracing a typical ray through the optical system and using the law of refraction. We will not deduce the mapping in this thesis, because the solution is a long and complicated expression. In the next chapter, we will show that we can find the mapping implicitly from (3.1) or (3.2).

Energy conservation

The energy conservation relation can be found in a similar way as in the reflector-case of the previous section. The south pole of the stereographic projection $\mathbf{y}(\hat{\mathbf{t}})$ does not change the Jacobians in the integral balances. Analogously, the Jacobian equation for \mathbf{m} is

$$\det(D\mathbf{m}(\mathbf{x})) = \frac{(1 + |\mathbf{m}(\mathbf{x})|^2)^2}{(1 + |\mathbf{x}|^2)^2} \frac{\tilde{f}(\mathbf{x})}{\tilde{g}(\mathbf{m}(\mathbf{x}))}, \quad (3.112)$$

where we omit the absolute value sign of the determinant and restrict ourselves to a positive Jacobian of the mapping.

Again, the transport boundary condition is $\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}$.

For the point-to-far-field lens problem we can construct the non-quadratic, logarithmic, optimal-transport cost function

$$c(\mathbf{x}, \mathbf{y}) = -\log \left(n - 1 + \frac{2 |\mathbf{x} - \mathbf{y}|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}|^2)} \right), \quad (3.113)$$

the generating function

$$G(\mathbf{x}, \mathbf{y}, w) = w \left(n - 1 + \frac{2 |\mathbf{x} - \mathbf{y}|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}|^2)} \right)^{-1}, \quad (3.114)$$

with corresponding inverse

$$H(\mathbf{x}, \mathbf{y}, w) = w \left(n - 1 + \frac{2 |\mathbf{x} - \mathbf{y}|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}|^2)} \right). \quad (3.115)$$

Combining the mapping with energy conservation gives the Jacobian equation

$$\det(D\mathbf{m}(\mathbf{x})) = \frac{(1 + |\mathbf{m}(\mathbf{x})|^2)^2}{(1 + |\mathbf{x}|^2)^2} \frac{\tilde{f}(\mathbf{x})}{\tilde{g}(\mathbf{m}(\mathbf{x}))}. \quad (3.116)$$

3.6 Point-to-parallel reflector

In this section, we consider a point source, a parallel target and two reflector surfaces.

For single surface systems, which we considered in the previous sections, the source and target coordinates are related by one of Hamilton's characteristic functions, where the characteristic function can be shown to only be a function of the position or direction coordinates at the target plane. For double surface systems with collimated and/or spherical source and target wavefronts the optical path length for all rays is equal to the same constant. This is a consequence of the *Theorem of Malus and Dupin* [12, p. 130].

In Figure 3.10, the light source is a point source at the origin O of the Cartesian coordinate system with $(x, y, z) \in \mathbb{R}^3$. The source emits light in the direction $\hat{\mathbf{s}} = \hat{\mathbf{e}}_r$. The first reflector surface is described by the parametrization $\mathcal{R}_1 : \mathbf{r}(\phi, \theta) = u(\phi, \theta) \hat{\mathbf{e}}_r$, where $u(\phi, \theta) > 0$ is the radial parameter that describes the location of the reflector surface, $0 \leq \phi \leq \pi$ is the zenith and $0 \leq \theta < 2\pi$ is the azimuth in the spherical coordinate system. The second reflector surface is described by $\mathcal{R}_2 : z = v(\mathbf{y})$ where \mathbf{y} are the Cartesian coordinates of the target plane $z = L$. The surface \mathcal{R}_1 reflects the ray $\hat{\mathbf{s}}$ in

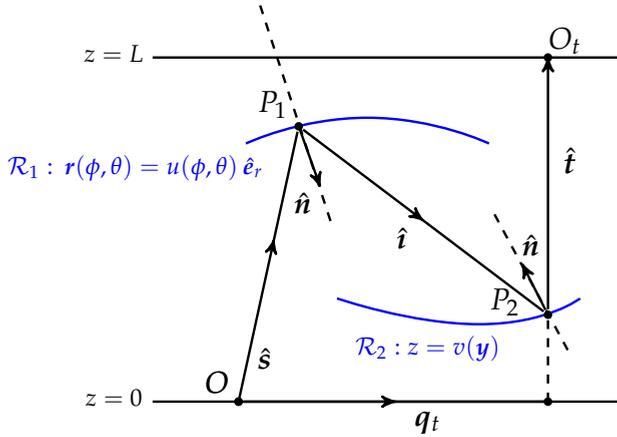


Figure 3.10: Illustration for the derivation of Hamilton's mixed characteristic W^* for a point source and parallel target.

direction \hat{i} and the surface \mathcal{R}_2 reflects the ray \hat{i} in direction $\hat{t} = \hat{e}_z$.

The position and direction coordinate vectors on the source plane are given by the two-vectors $\mathbf{q}_s = \mathbf{0}$ and $\mathbf{p}_s = (s_1, s_2)$, respectively. The position and direction coordinates on the target plane are given by $\mathbf{q}_t = \mathbf{y}$ and $\mathbf{p}_t = (t_1, t_2) = \mathbf{0}$, respectively. We define the position vectors $P_1(u(\hat{s}), \mathbf{p}_s, u(\hat{s}) s_3)$ and $P_2(\mathbf{q}_t, v(\mathbf{y}))$.

In this example, we choose the target plane not to coincide with the source plane, i.e., $L = z_t \neq z_s = 0$.

*The mixed characteristic W^**

The point characteristic (or optical path length) between point $(\mathbf{q}_s, 0)$ and $O_t(\mathbf{q}_t, L)$ is given by

$$\begin{aligned} V(\mathbf{q}_s, \mathbf{q}_t) &= u(\hat{s}) + d(P_1, P_2) + L - v(\mathbf{y}) \\ &= u(\hat{s}) + \sqrt{|\mathbf{q}_t - u(\hat{s}) \mathbf{p}_s|^2 + (v(\mathbf{y}) - u(\hat{s}) s_3)^2} + L - v(\mathbf{y}), \end{aligned} \quad (3.117)$$

where $u(\hat{s})$ is the distance from O to P_1 and $v(\mathbf{y})$ is the distance from P_2 to $(\mathbf{q}_t, 0)$.

Hamilton's mixed characteristic, which depends on the direction of the ray at the source plane and the position at the target plane, is given by

$$W^*(\mathbf{p}_s, \mathbf{q}_t) = V(\mathbf{q}_s, \mathbf{q}_t) + \mathbf{q}_s \cdot \mathbf{p}_s = V(\mathbf{q}_s, \mathbf{q}_t), \quad (3.118)$$

since $\mathbf{q}_s = \mathbf{0}$. Figure 3.1 shows that the mixed characteristic W^* is independent of both the direction coordinate \mathbf{p}_s and the spatial coordinate \mathbf{q}_t , so that W^* is a constant length. We obtain

$$W^* = u(\hat{\mathbf{s}}) + \sqrt{|\mathbf{q}_t - u(\hat{\mathbf{s}}) \mathbf{p}_s|^2 + (v(\mathbf{y}) - u(\hat{\mathbf{s}}) s_3)^2} + L - v(\mathbf{y}). \quad (3.119)$$

From (3.117) and (3.118) we find two equations for the distance $d(P_1, P_2)$ as

$$d(P_1, P_2)^2 = (W^* - L - u(\hat{\mathbf{s}}) + v(\mathbf{y}))^2, \quad (3.120a)$$

$$d(P_1, P_2)^2 = |\mathbf{q}_t - u(\hat{\mathbf{s}}) \mathbf{p}_s|^2 + (v(\mathbf{y}) - u(\hat{\mathbf{s}}) s_3)^2. \quad (3.120b)$$

Combining (3.120a) and (3.120b) gives

$$(W^* - L - u(\hat{\mathbf{s}}) + v(\mathbf{y}))^2 - |\mathbf{q}_t - u(\hat{\mathbf{s}}) \mathbf{p}_s|^2 - (v(\mathbf{y}) - u(\hat{\mathbf{s}}) s_3)^2 = 0,$$

and using $\mathbf{p}_s = (s_1, s_2)$ and $\mathbf{q}_t = \mathbf{y}$ results in

$$\begin{aligned} (W^* - L)^2 - 2u(\hat{\mathbf{s}})(W^* - L) + 2v(\mathbf{y})(W^* - L) - 2u(\hat{\mathbf{s}})v(\mathbf{y}) \\ - |\mathbf{y}|^2 + 2u(\hat{\mathbf{s}})(s_1 y_1 + s_2 y_2 + s_3 v(\mathbf{y})) = 0. \end{aligned} \quad (3.121)$$

Rewriting gives

$$\begin{aligned} (W^* - L)^2 - 2u(\hat{\mathbf{s}})(W^* - L - s_1 y_1 - s_2 y_2 - s_3 v(\mathbf{y})) \\ + 2v(\mathbf{y})(W^* - L) - 2u(\hat{\mathbf{s}})v(\mathbf{y}) - |\mathbf{y}|^2 = 0. \end{aligned} \quad (3.122)$$

We continue and use the reduced optical path length $\beta = W^* - L$.

The cost function

We can find a non-quadratic, logarithmic, cost function in optimal transport theory. In general, we can write (3.122) as

$$a(\hat{\mathbf{s}}, \mathbf{y}) u(\hat{\mathbf{s}}) + b v(\mathbf{y}) + c(\hat{\mathbf{s}}) u(\hat{\mathbf{s}}) v(\mathbf{y}) + d(\mathbf{y}) = 0, \quad (3.123)$$

with

$$a(\hat{\mathbf{s}}, \mathbf{y}) = 2(-\beta + s_1 y_1 + s_2 y_2), \quad b = 2\beta, \quad (3.124)$$

$$c(\hat{\mathbf{s}}) = -2(1 - s_3), \quad d(\mathbf{y}) = \beta^2 - |\mathbf{y}|^2. \quad (3.125)$$

To separate the coefficient $a(\hat{\mathbf{s}}, \mathbf{y})$ from $u(\hat{\mathbf{s}})$ we perform the change of variables $h = 1/u$ to get

$$a(\hat{\mathbf{s}}, \mathbf{y}) + b v(\mathbf{y}) h(\hat{\mathbf{s}}) + c(\hat{\mathbf{s}}) v(\mathbf{y}) + d(\mathbf{y}) h(\hat{\mathbf{s}}) = 0. \quad (3.126)$$

In order to set the coefficient of $v(\mathbf{y}) h(\hat{\mathbf{s}})$ equal to 1, we let $\tilde{a} = a/b$, $\tilde{c} = c/b$, and $\tilde{d} = d/b$ to obtain

$$\tilde{a}(\hat{\mathbf{s}}, \mathbf{y}) + v(\mathbf{y}) h(\hat{\mathbf{s}}) + \tilde{c}(\hat{\mathbf{s}}) v(\mathbf{y}) + \tilde{d}(\mathbf{y}) h(\hat{\mathbf{s}}) = 0. \quad (3.127)$$

Factoring this equation gives

$$(v(\mathbf{y}) + \tilde{d}(\mathbf{y})) (h(\hat{\mathbf{s}}) + \tilde{c}(\hat{\mathbf{s}})) + \tilde{a}(\hat{\mathbf{s}}, \mathbf{y}) - \tilde{d}(\mathbf{y}) \tilde{c}(\hat{\mathbf{s}}) = 0. \quad (3.128)$$

Letting $\tilde{u}_1(\hat{\mathbf{s}}) = h(\hat{\mathbf{s}}) + \tilde{c}(\hat{\mathbf{s}})$ and $\tilde{u}_2(\mathbf{y}) = v(\mathbf{y}) + \tilde{d}(\mathbf{y})$ and $\bar{c}(\hat{\mathbf{s}}, \mathbf{y}) = -\tilde{a}(\hat{\mathbf{s}}, \mathbf{y}) + \tilde{d}(\mathbf{y}) \tilde{c}(\hat{\mathbf{s}})$ gives

$$\tilde{u}_1(\hat{\mathbf{s}}) \tilde{u}_2(\mathbf{y}) = \bar{c}(\hat{\mathbf{s}}, \mathbf{y}). \quad (3.129)$$

It can be shown that all terms in the above equation are positive [135], so that we can take the logarithm and set $u_1(\hat{\mathbf{s}}) = -\log(\tilde{u}_1(\hat{\mathbf{s}}))$, $u_2(\mathbf{y}) = \log(\tilde{u}_2(\mathbf{y}))$, $c(\hat{\mathbf{s}}, \mathbf{y}) = \log(\bar{c}(\hat{\mathbf{s}}, \mathbf{y}))$ in order to find a relation of the form

$$u_2(\mathbf{y}) - u_1(\hat{\mathbf{s}}) = c(\hat{\mathbf{s}}, \mathbf{y}), \quad (3.130)$$

and transforming to stereographic coordinates using (3.7), we end up with the non-quadratic cost function

$$c(\mathbf{x}, \mathbf{y}) = \log \left(\frac{\beta^2 - \beta \mathbf{x} \cdot \mathbf{y} + |\mathbf{x}|^2 |\mathbf{y}|^2}{\beta^2 (1 + |\mathbf{x}|^2)} \right). \quad (3.131)$$

More details of the derivation of this cost function are shown by my colleague Teun van Roosmalen [135].

The generating function

Solving (3.122) for u we obtain

$$u(\hat{\mathbf{s}}) = \frac{\beta^2 - |\mathbf{y}|^2 + 2 v(\mathbf{y}) \beta}{2 (\beta - s_1 y_1 - s_2 y_2 + v(\mathbf{y}) (1 - s_3))}. \quad (3.132)$$

Changing to stereographic coordinates \mathbf{x} using (3.9) gives

$$u(\mathbf{x}) = \frac{1}{2} \frac{(1 + |\mathbf{x}|^2) (\beta^2 - |\mathbf{y}|^2 + 2 \beta v(\mathbf{y}))}{\beta - 2 \mathbf{x} \cdot \mathbf{y} + (\beta + 2 v(\mathbf{y})) |\mathbf{x}|^2}. \quad (3.133)$$

Now, we construct the generating function $G(\mathbf{x}, \mathbf{y}, w) = u$ with $w = v(\mathbf{y})$ as

$$G(\mathbf{x}, \mathbf{y}, w) = \frac{1}{2} \frac{(1 + |\mathbf{x}|^2) (\beta^2 - |\mathbf{y}|^2 + 2 \beta w)}{\beta - 2 \mathbf{x} \cdot \mathbf{y} + (\beta + 2 w) |\mathbf{x}|^2}. \quad (3.134)$$

The function H can be found by inverting this relation. We obtain

$$H(\mathbf{x}, \mathbf{y}, w) = \frac{4 w \mathbf{x} \cdot \mathbf{y} - 2 w \beta (1 + |\mathbf{x}|^2) - (1 + |\mathbf{x}|^2)(|\mathbf{y}|^2 - \beta^2)}{2 |\mathbf{x}|^2 (2 w - \beta) - 2 \beta}. \quad (3.135)$$

The mapping

We will not present the mapping for this optical system in this work but continue straight to energy conservation. In the next chapter, we will show that we can derive the mapping implicitly from (3.1) or (3.2).

Energy conservation

The energy conservation relation can be derived similarly as in the previous sections. The Jacobian equation for \mathbf{m} is

$$\det(\mathbf{D}\mathbf{m}(\mathbf{x})) = \frac{4}{(1 + |\mathbf{x}|^2)^2} \frac{\tilde{f}(\mathbf{x})}{g(\mathbf{m}(\mathbf{x}))}, \quad (3.136)$$

where we omit the absolute value sign of the determinant and restrict ourselves to a positive Jacobian of the mapping.

The corresponding transport boundary condition is $\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}$.

For the point-to-parallel reflector problem we can construct the non-quadratic optimal-transport cost function

$$c(\mathbf{x}, \mathbf{y}) = \log \left(\frac{\beta^2 - \beta \mathbf{x} \cdot \mathbf{y} + |\mathbf{x}|^2 |\mathbf{y}|^2}{\beta^2 (1 + |\mathbf{x}|^2)} \right), \quad (3.137)$$

the generating function

$$G(\mathbf{x}, \mathbf{y}, w) = \frac{1}{2} \frac{(1 + |\mathbf{x}|^2) (\beta^2 - |\mathbf{y}|^2 + 2 \beta w)}{\beta - 2 \mathbf{x} \cdot \mathbf{y} + (\beta + 2 w) |\mathbf{x}|^2}, \quad (3.138)$$

with corresponding inverse

$$H(\mathbf{x}, \mathbf{y}, w) = \frac{4 w \mathbf{x} \cdot \mathbf{y} - 2 w \beta (1 + |\mathbf{x}|^2) - (1 + |\mathbf{x}|^2)(|\mathbf{y}|^2 - \beta^2)}{2 |\mathbf{x}|^2 (2 w - \beta) - 2 \beta}. \quad (3.139)$$

Combining the mapping with energy conservation gives the Jacobian equation

$$\det(\mathbf{D}\mathbf{m}(\mathbf{x})) = \frac{4}{(1 + |\mathbf{x}|^2)^2} \frac{\tilde{f}(\mathbf{x})}{g(\mathbf{m}(\mathbf{x}))}. \quad (3.140)$$

3.7 Summary

In Table 3.1 we give a complete overview of the 16 base cases showing the source, target, generating function and cost function if it exists. All generating functions in these tables were derived by myself. I derived the cost functions together with some colleagues [124, 131, 133, 135, 161, 162, 164].

We list a few conventions and remarks:

- System 4 has a quadratic cost function and can also be described using the standard Monge-Ampère equation [163].
- The parallel-to-point systems (System 3 and 11) are effectively the same as the point-to-parallel systems (System 8 and 16), respectively, with source and target coordinates interchanged. Hence, strictly speaking, our 16 base cases reduce to 14 base cases.
- For lens systems with parallel or point sources *and* parallel or point targets, i.e., System 11, 12, 15 and 16, we can think of two layouts: one single lens with two freeform surfaces, or two separate lenses each consisting of one flat and one freeform surface [164]. In the icons and equations we restrict ourselves to the first layout and consider one single lens of refractive index n .
- The \pm -signs for System 11, 13, 15 and 16 indicate that there are multiple solutions. For the parallel-to-parallel lens (System 12), a plus sign in front of the square root was found to be a consequence of the first layout as described above, i.e., one single lens with two freeform surfaces [161, p. 46–47]. Whether a similar result holds for the other systems can be determined using geometrical arguments, but is beyond the scope of this thesis.
- For the parallel-to-parallel reflectors and lenses (4 and 12), we simply defined the generating function as $u(\mathbf{x}) = G(\mathbf{x}, \mathbf{y}, w) = -c(\mathbf{x}, \mathbf{y}) + w$. My colleague Nitin Yadav [161, 162, 164] considers these systems in more detail taking $u_1(\mathbf{x}) = u(\mathbf{x})$ and $u_2(\mathbf{y}) = -w(\mathbf{y})$.
- For the other systems that have cost functions, we could also have chosen to define $G(\mathbf{x}, \mathbf{y}, w) = u_1(\mathbf{x}) = -c(\mathbf{x}, \mathbf{y}) + w$ where $u_1(\mathbf{x}) \neq u(\mathbf{x})$. The formulation of the generating function is not unique and a matter of convention, but each generating function G has a unique inverse H . In Chapter 4 we continue to use the convention $u(\mathbf{x}) = G(\mathbf{x}, \mathbf{y}, w)$, i.e., the generating function is always equal to the optical surface $u(\mathbf{x})$ that

we are interested in. In the next chapter, we will also elaborate on the properties of generating functions.

Of course, we can think of many more optical systems. More generally, the source can be thought of as a surface from which the light is emitted in the direction normal to that surface. In case of a parallel bundle, this surface is a part of a plane, and in case of a point source, this surface is a part of the unit sphere. Double freeform systems could include a reflector and a lens as the required two freeform surfaces, which adds another eight options, i.e., for parallel-to-point, parallel-to-parallel, point-to-point and point-to-parallel we can choose the first freeform surface as a reflector and the second as a lens, or the other way around.

In Chapter 8, we move beyond the base cases and consider a double freeform lens with a point source and far-field target.

Source	Target	Icon	Generating function	Cost function
1. Parallel	Near-field		$G(\mathbf{x}, \mathbf{y}, w) = \frac{1}{2w} - \frac{w}{2} \mathbf{x} - \mathbf{y} ^2$	—
2. Parallel	Far-field		$G(\mathbf{x}, \mathbf{y}, w) = \mathbf{x} \cdot \mathbf{y} + w$	$c(\mathbf{x}, \mathbf{y}) = -\mathbf{x} \cdot \mathbf{y}$
3. Parallel	Point		$G(\mathbf{x}, \mathbf{y}, w) = \frac{1}{2} \frac{(1+ \mathbf{y} ^2)(\beta^2 - \mathbf{x} ^2 + 2\beta w)}{\beta - 2\mathbf{x} \cdot \mathbf{y} + (\beta + 2w) \mathbf{y} ^2}$	$c(\mathbf{x}, \mathbf{y}) = \log \left(\frac{\beta^2 - \beta \mathbf{x} \cdot \mathbf{y} + \mathbf{x} ^2 \mathbf{y} ^2}{\beta^2 (1+ \mathbf{y} ^2)} \right)$
4. Parallel	Parallel		$G(\mathbf{x}, \mathbf{y}, w) = -c(\mathbf{x}, \mathbf{y}) + w$	$c(\mathbf{x}, \mathbf{y}) = -\frac{\beta^2 + 2\beta L - \mathbf{x} - \mathbf{y} ^2}{2\beta}$

Table 3.1: Reflector systems with parallel sources. Cartesian coordinates are black and stereographic coordinates are blue.

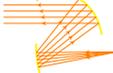
Source	Target	Icon	Generating function	Cost function
5. Point	Near-field		$G(\mathbf{x}, \mathbf{y}, w) = \frac{w^{-2} - \frac{1}{4} \mathbf{y} ^2}{w^{-1} - \frac{1}{2} \mathbf{p}_s(\mathbf{x}) \cdot \mathbf{y}}$	—
6. Point	Far-field		$G(\mathbf{x}, \mathbf{y}, w) = w (1 - \hat{\mathbf{s}}(\mathbf{x}) \cdot \hat{\mathbf{f}}(\mathbf{y}))^{-1}$	$c(\mathbf{x}, \mathbf{y}) = -\log(1 - \hat{\mathbf{s}}(\mathbf{x}) \cdot \hat{\mathbf{f}}(\mathbf{y}))$
7.* Point	Point		$G(\mathbf{x}, \mathbf{y}, w) = \frac{- \mathbf{d} ^2 + T^2 - 2T w + 2w \mathbf{d} \cdot \hat{\mathbf{f}}(\mathbf{y})}{2(T - \mathbf{d} \cdot \hat{\mathbf{s}}(\mathbf{x}) - w(1 - \hat{\mathbf{s}}(\mathbf{x}) \cdot \hat{\mathbf{f}}(\mathbf{y})))}$	$c(\mathbf{x}, \mathbf{y}) = \log \left(1 - \frac{(1 - \hat{\mathbf{s}}(\mathbf{x}) \cdot \hat{\mathbf{f}}(\mathbf{y})) (T^2 - \mathbf{d} ^2)}{2(T - \mathbf{d} \cdot \hat{\mathbf{s}}(\mathbf{x})) (T - \mathbf{d} \cdot \hat{\mathbf{f}}(\mathbf{y}))} \right)$
8. Point	Parallel		$G(\mathbf{x}, \mathbf{y}, w) = \frac{1}{2} \frac{(1 + \mathbf{x} ^2) (\beta^2 - \mathbf{y} ^2 + 2\beta w)}{\beta - 2 \mathbf{x} \cdot \mathbf{y} + (\beta + 2w) \mathbf{x} ^2}$	$c(\mathbf{x}, \mathbf{y}) = \log \left(\frac{\beta^2 - \beta \mathbf{x} \cdot \mathbf{y} + \mathbf{x} ^2 \mathbf{y} ^2}{\beta^2 (1 + \mathbf{x} ^2)} \right)$

Table 3.1: Reflector systems with point sources. Cartesian coordinates are black and stereographic coordinates are blue.

*Here \mathbf{d} is the vector connecting the two points and T the angular characteristic independent of any position or direction coordinate.

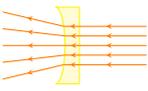
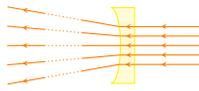
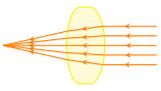
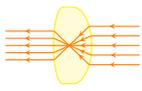
Source	Target	Icon	Generating function	Cost function
9. Parallel	Near-field		$G(\mathbf{x}, \mathbf{y}, w) = \frac{1}{n^2-1} (n w + \sqrt{w^2 + (n^2 - 1) \mathbf{x} - \mathbf{y} ^2})$	—
10. Parallel	Far-field		$G(\mathbf{x}, \mathbf{y}, w) = \mathbf{x} \cdot \mathbf{y} + w$	$c(\mathbf{x}, \mathbf{y}) = -\mathbf{x} \cdot \mathbf{y}$
11.* Parallel	Point		$G(\mathbf{x}, \mathbf{y}, w) = \frac{1}{n^2-1} (-w - \beta + n^2 (\mathbf{p}_t(\mathbf{y}) \cdot \mathbf{x} + t_3(\mathbf{y}) w) \pm S_1)$	—
12.** Parallel	Parallel		$G(\mathbf{x}, \mathbf{y}, w) = -c(\mathbf{x}, \mathbf{y}) + w$	$c(\mathbf{x}, \mathbf{y}) = -L + \frac{n\beta}{1-n^2} + \frac{1}{n^2-1} S_2$

Table 3.1: Lens systems with parallel sources. Cartesian coordinates are black and stereographic coordinates are blue.

* $S_1 = \sqrt{[w - n^2 (\mathbf{p}_t(\mathbf{y}) \cdot \mathbf{x} + t_3(\mathbf{y}) w) - \beta]^2 + (n^2 - 1) [-n^2 (|\mathbf{x}|^2 + w^2) + (w + \beta)^2]}$ with $n(z_s) = n(z_t) = 1$.

** $S_2 = \sqrt{\beta^2 + (n^2 - 1) |\mathbf{x} - \mathbf{y}|^2}$.

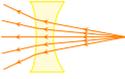
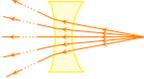
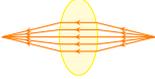
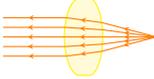
Source	Target	Icon	Generating function	Cost function
13.* Point	Near-field		$G(\mathbf{x}, \mathbf{y}, w) = \frac{1}{n^2-1} (n w - \mathbf{p}_s(\mathbf{x}) \cdot \mathbf{y} \pm S_3)$	—
14. Point	Far-field		$G(\mathbf{x}, \mathbf{y}, w) = w (n - \hat{\mathbf{s}}(\mathbf{x}) \cdot \hat{\mathbf{f}}(\mathbf{y}))^{-1}$	$c(\mathbf{x}, \mathbf{y}) = -\log(n - \hat{\mathbf{s}}(\mathbf{x}) \cdot \hat{\mathbf{f}}(\mathbf{y}))$
15.** Point	Point		$G(\mathbf{x}, \mathbf{y}, w) = \frac{1}{n^2-1} (-T + n^2 \mathbf{d} \cdot \hat{\mathbf{s}}(\mathbf{x}) + w(1 - n^2 \hat{\mathbf{s}}(\mathbf{x}) \cdot \hat{\mathbf{f}}(\mathbf{y})) \pm S_4)$	—
16.*** Point	Parallel		$G(\mathbf{x}, \mathbf{y}, w) = \frac{1}{n^2-1} (-w - \beta + n^2 \times (\mathbf{p}_s(\mathbf{x}) \cdot \mathbf{y} + s_3(\mathbf{x}) w) \pm S_5)$	—

Table 3.1: Lens systems with point sources. Cartesian coordinates are black and stereographic coordinates are blue.

* $S_3 = \sqrt{(-n w + \mathbf{p}_s(\mathbf{x}) \cdot \mathbf{y})^2 - (n^2 - 1)(w^2 - |\mathbf{y}|^2)}$ with $n(z_s) = n$.

** $S_4 = \sqrt{[T - n^2 \mathbf{d} \cdot \hat{\mathbf{s}}(\mathbf{x}) - w(1 - n^2 \hat{\mathbf{s}}(\mathbf{x}) \cdot \hat{\mathbf{f}}(\mathbf{y}))]^2 - (n^2 - 1)[T^2 - n^2 |\mathbf{d}|^2 - 2 w T + (1 - n^2) w^2 + 2 n^2 w \mathbf{d} \cdot \hat{\mathbf{f}}(\mathbf{y})]}$ with $n(z_s) = n(z_t) = 1$.

*** $S_5 = \sqrt{[w - n^2 (\mathbf{p}_s(\mathbf{x}) \cdot \mathbf{y} + s_3(\mathbf{x}) w) - \beta]^2 + (n^2 - 1)[-n^2 (|\mathbf{y}|^2 + w^2) + (w + \beta)^2]}$ with $n(z_s) = n(z_t) = 1$.

Chapter 4

Generated Jacobian Equations

The Jacobian equations for the mapping seen thus far are all equations of the form

$$\det(D\mathbf{m}(x)) = \frac{f(x)}{g(\mathbf{m}(x))}, \quad (4.1)$$

where $D\mathbf{m}$ is the Jacobi matrix of \mathbf{m} , and we take $f(x)$ and $g(\mathbf{m}(x))$ to incorporate the Jacobians of the coordinate transformation to stereographic coordinates in case of a point source and/or a far-field target, respectively.

In Chapter 3, we have seen that the Jacobian equation reduces to the *standard Monge-Ampère equation* for a parallel-to-far-field single optical surface $z = u(x)$, i.e., System 2 and 10 in Table 3.1. The mapping for these systems is $\mathbf{m} = \nabla u$. In the following, we use the numbering of the optical systems in Table 3.1. Some optical systems (System 2 – 4, 6 – 8, 10, 12 and 14) can be parametrized by an equation of the form

$$u_2(\mathbf{y}) - u_1(x) = c(x, \mathbf{y}), \quad (4.2)$$

where we have $u_1(x) = u(x)$ or $u_1(x) \neq u(x)$ is related to the location of the surface $u(x)$ via a change of variables, $u_2(\mathbf{y})$ is an auxiliary variable related to one of Hamilton's characteristic functions (System 2, 6, 10 and 14) or a second optical surface (System 3, 4, 7, 8 and 12), and $c(x, \mathbf{y})$ is an optimal-transport cost function. We will show that the the Jacobian equation for the mapping \mathbf{m} in (4.1) can be written as a so-called *generalized Monge-Ampère equation*, also known as a *partial differential equation of optimal transport* [148, p. 282].

However, relation (4.2) does not exist for many optical systems, e.g., for systems with near-field targets. As shown in Table 3.1, for 7 out of the 16 base cases we cannot derive a cost function. More generally, all optical systems presented in this thesis can be described using a generating function, i.e.,

$$u(x) = G(x, \mathbf{y}, w(\mathbf{y})), \quad (4.3)$$

with $u(\mathbf{x})$ the parameter describing one of the optical surfaces of the system and $w(\mathbf{y})$ an auxiliary variable related to one of Hamilton's characteristic functions (System 1, 2, 5, 6, 9, 10, 13 and 14) or a second optical surface (System 3, 4, 7, 8, 11, 12, 15 and 16). We will show that these systems have an associated *generated Jacobian equation*. Again, $w(\mathbf{y})$ is related to a second optical surface if we have a parallel target, as in Section 3.6, or a point target.

In this chapter, we present the underlying theory of these equations, which lays the foundation of our numerical algorithms presented later in this thesis. In Section 4.1, we will present the general problem statement underpinning all Jacobian equations in this thesis.

In Section 4.2, we will use definitions and theorems from the field of *convex analysis* [13]. We will see that this theory applies to the parallel-to-far-field systems (System 2 and 10) with corresponding standard Monge-Ampère equation and generating function $G(\mathbf{x}, \mathbf{y}, w) = \mathbf{x} \cdot \mathbf{y} + w$.

In Section 4.3, this theory is extended to partial differential equations of optimal transport with generating functions that can be rewritten to the form $u_1(\mathbf{x}) = G(\mathbf{x}, \mathbf{y}, u_2(\mathbf{y})) = -c(\mathbf{x}, \mathbf{y}) + u_2(\mathbf{y})$. We will call this theory *c-convex analysis*.

For generic generating functions, we further generalize the theory in Section 4.4 to what we will call *G-convex analysis*. The field of G-convexity theory is quite new [69, 70, 82, 145], while c-convexity theory in optimal transport theory is well-established [138, 147, 148].

Lastly, we end this chapter in Section 4.5 with a proof showing that the transport boundary condition $\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}$, first introduced in (3.48), is equivalent to the implicit boundary condition $\mathbf{m}(\mathcal{X}) = \mathcal{Y}$, stating that all the light from the source domain \mathcal{X} must be transferred to the target domain \mathcal{Y} . The equivalence of the boundary conditions follows from the edge-ray principle [128] and convexity of the optical surface.

4.1 Measure-preserving mappings

A common problem across different disciplines in mathematics is that of finding a mapping from a given space \mathcal{X} to a given space \mathcal{Y} , within a certain family of admissible mappings. Examples of this problem are the rearrangement of mass from one distribution into another in an optimal way [18], the analysis of semigeostrophic flows [34], the coupling of probability measures to maximize covariance [138, p. 41], stable matching problems in economics [110], and the inverse design of freeform optical surfaces.

Consider \mathcal{X} and \mathcal{Y} to be two separable metric spaces such that any meas-

ures μ on \mathcal{X} and ν on \mathcal{Y} are Radon measures, i.e., $\mu \in \mathcal{R}(\mathcal{X})$ and $\nu \in \mathcal{R}(\mathcal{Y})$, satisfying the overall mass balance condition $\mu(\mathcal{X}) = \nu(\mathcal{Y}) < \infty$, i.e.,

$$\int_{\mathcal{X}} d\mu(x) = \int_{\mathcal{Y}} d\nu(y) = 1. \quad (4.4)$$

We also call these measures *probability measures*.

The aim is to find a mapping $\mathbf{m} : \mathcal{X} \rightarrow \mathcal{Y}$ that maps μ into ν , i.e., for every (Borel) set $\mathcal{A} \subset \mathcal{X}$ we have the conservation relation

$$\int_{m(\mathcal{A})} h(\mathbf{y}) d\nu(\mathbf{y}) = \int_{\mathcal{A}} h(\mathbf{m}(x)) d\mu(x), \quad (4.5)$$

for all continuous functions $h \in C(\mathcal{Y})$. Equivalently, it holds for any set $\mathcal{B} \subset \mathcal{Y}$ that [122, Ch. 2]

$$\nu(\mathcal{B}) = \mu(\{x \in \mathcal{X} : \mathbf{m}(x) \in \mathcal{B}\}) = \mu(\mathbf{m}^{-1}(\mathcal{B})). \quad (4.6)$$

A mapping $\mathbf{m} : \mathcal{X} \rightarrow \mathcal{Y}$ satisfying (4.6) is called a *measure-preserving* mapping, which moves a single point from a measurable space to another. The notation $\mathbf{m}_\#(\mu)$ is frequently used to denote the so-called *push-forward* of μ by \mathbf{m} , i.e., each element of mass of a measure μ on \mathcal{X} is pushed forward via the mapping \mathbf{m} to get an element of mass in \mathcal{Y} so that we obtain an aggregated measure $\mathbf{m}_\#(\mu)$ on \mathcal{Y} [122, Ch. 2]. Hence, $\nu = \mu \circ \mathbf{m}^{-1} = \mathbf{m}_\#(\mu)$.

In our case, \mathcal{X} and \mathcal{Y} are domains in \mathbb{R}^2 and we consider the continuous densities $f : \mathcal{X} \rightarrow [0, \infty)$ and $g : \mathcal{Y} \rightarrow [0, \infty)$ such that $d\mu(x) = f(x) dx$ and $d\nu(y) = g(y) dy$, where dx and dy are infinitesimal area elements on \mathbb{R}^2 . The integral condition in (4.5) with $h(y) = 1$ becomes the mass balance condition

$$\int_{\mathcal{A}} f(x) dx = \int_{m(\mathcal{A})} g(y) dy. \quad (4.7)$$

Substituting the mapping $\mathbf{y} = \mathbf{m}(x)$ we can derive the Jacobian equation

$$\det(D\mathbf{m}(x)) = \frac{f(x)}{g(\mathbf{m}(x))}, \quad (4.8)$$

where $D\mathbf{m}(x)$ is the 2×2 Jacobi matrix of \mathbf{m} with respect to x and we assume a positive Jacobian of the mapping $\det(D\mathbf{m}) > 0$. The coordinates $x \in \mathcal{X}$ with image $\mathbf{m}(x) \in \mathcal{Y}$ determine the Jacobian. If \mathbf{m} is unknown, this equation is a nonlinear first-order PDE for the unknown \mathbf{m} .

For our freeform design problems the domains \mathcal{X} , \mathcal{Y} are in \mathbb{R}^2 , in Cartesian (or polar) coordinates or in Cartesian (or polar) stereographic coordinates.

Later on in this thesis, we will provide more details on the use of polar (stereographic) coordinates. The above energy conservation relation in (4.7) is exactly similar to the relation (3.62) for the parallel-to-near-field system in Section 3.3, i.e., System 1 in Table 3.1. For the other optical systems the integral balance includes Jacobians for the change of coordinates from spherical to stereographic coordinates. In the remainder of this chapter, we omit these Jacobians for generality, e.g., for the point-to-far-field systems in Section 3.4 and in Section 3.5, we think of $f(\mathbf{x})$ to incorporate the Jacobian as $f(\mathbf{x}) = 4\tilde{f}(\mathbf{x})/(1 + |\mathbf{x}|^2)^2$ and $g(\mathbf{y}) = 4\tilde{g}(\mathbf{y})/(1 + |\mathbf{m}(\mathbf{x})|^2)^2$, cf. (3.94).

4.2 Convex analysis

In this section, we will give a definition of the Legendre-Fenchel transform and show that a convex or concave function u that forms a so-called *conjugate pair* with its transform satisfies the standard Monge-Ampère equation. First, we start with the basic definitions of a convex set and convex function.

Definition 4.2.1 (Convex set). *A set $\mathcal{C} \subset \mathbb{R}^2$ is convex if it contains all line segments between any two points $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}$, i.e.,*

$$\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}, \forall t \in [0, 1] : \quad t \mathbf{x}_1 + (1 - t) \mathbf{x}_2 \in \mathcal{C}. \quad (4.9)$$

Definition 4.2.2 (Convex and concave function). *A function $u : \mathcal{X} \rightarrow \mathbb{R}$ on a convex set \mathcal{X} is called convex if for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and all $t \in [0, 1]$ we have*

$$u(t \mathbf{x}_1 + (1 - t) \mathbf{x}_2) \leq t u(\mathbf{x}_1) + (1 - t) u(\mathbf{x}_2). \quad (4.10)$$

The function is strictly convex if for all $t \in (0, 1)$ and $\mathbf{x}_1 \neq \mathbf{x}_2 \in \mathcal{X}$ we have the inequality $<$ in (4.10). Analogous definitions for a concave and strictly concave function on a convex set can be formulated by changing \leq and $<$ in the above definitions to \geq and $>$.

A twice differentiable function is convex or concave if and only if its Hessian matrix is positive or negative semi-definite, respectively, in the domain. Convex and concave functions are continuous and locally Lipschitz in the interior of their domain. However, discontinuities can occur at the boundary [138, p. 24].

The Legendre-Fenchel transform of a concave function $w : \mathcal{Y} \rightarrow \mathbb{R}$ is given

by

$$\forall \mathbf{x} \in \mathcal{X} : \quad w^*(\mathbf{x}) = \sup_{\mathbf{y} \in \mathcal{Y}} (\mathbf{x} \cdot \mathbf{y} + w(\mathbf{y})). \quad (4.11a)$$

The function $w^* : \mathcal{X} \rightarrow \mathbb{R}$ is called the *conjugate* of w . Similarly, we can introduce the Legendre-Fenchel transform of a function $u : \mathcal{X} \rightarrow \mathbb{R}$ as

$$\forall \mathbf{y} \in \mathcal{Y} : \quad u^*(\mathbf{y}) = \inf_{\mathbf{x} \in \mathcal{X}} (-\mathbf{x} \cdot \mathbf{y} + u(\mathbf{x})). \quad (4.11b)$$

The function $u^* : \mathcal{Y} \rightarrow \mathbb{R}$ is called the *conjugate* of u .

The Legendre-Fenchel transforms are also frequently defined as a sup/sup pair, i.e., the infimum in (4.11b) becomes

$$\forall \mathbf{y} \in \mathcal{Y} : \quad \tilde{u}^*(\mathbf{y}) = \sup_{\mathbf{x} \in \mathcal{X}} (\mathbf{x} \cdot \mathbf{y} - u(\mathbf{x})), \quad (4.12)$$

i.e., $\tilde{u}^*(\mathbf{y}) = -u^*(\mathbf{y})$. This is a matter of convention. We adopt (4.11b).

A property of the Legendre-Fenchel transform is that the transform u^* of a convex function u is concave, since $\mathbf{y} \mapsto -\mathbf{x} \cdot \mathbf{y} + u(\mathbf{x})$ is an affine function, which is concave and continuous. The hypograph of these functions is therefore closed and concave. Subsequently, u^* is the pointwise infimum of these functions, which has a hypograph as the intersection of the affine functions' hypographs. Each of those hypographs is closed and concave, so that the hypograph of u^* is also closed and concave.

Furthermore, $u = u^{**}$ if and only if u is convex and lower semi-continuous by the Fenchel-Moreau theorem [42]. We will now present the definition of lower (and upper) semi-continuity.

Definition 4.2.3 (Semi-continuity). *A function $u : \mathcal{X} \rightarrow \mathbb{R}$ is lower-semicontinuous at a point $\mathbf{x}_0 \in \mathcal{X}$ if for every $\epsilon > 0$ there exists a neighborhood \mathcal{U} of \mathbf{x}_0 such that $u(\mathbf{x}_0) - \epsilon \leq u(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{U}$. For a metric space, this can also be written as*

$$\liminf_{\mathbf{x} \rightarrow \mathbf{x}_0} u(\mathbf{x}) \geq u(\mathbf{x}_0).$$

Similarly, a function $u : \mathcal{X} \rightarrow \mathbb{R}$ is upper-semicontinuous at a point $\mathbf{x}_0 \in \mathcal{X}$ if for every $\epsilon > 0$ there exists a neighborhood \mathcal{U} of \mathbf{x}_0 such that $u(\mathbf{x}_0) + \epsilon \geq u(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{U}$. For a metric space, this can also be written as

$$\limsup_{\mathbf{x} \rightarrow \mathbf{x}_0} u(\mathbf{x}) \leq u(\mathbf{x}_0).$$

It is clear that u is continuous at \mathbf{x}_0 if and only if u is lower semi-continuous and upper semi-continuous at this point.

We will show that the Legendre-Fenchel transformation (4.11b) of a strictly convex differentiable u can be interpreted as a mapping between the graph of the function and the family of tangents to the graph. Assuming \mathcal{X} and \mathcal{Y} are compact, the supremum in (4.11a) and infimum in (4.11b) become a maximum and a minimum, respectively.

The tangent plane to $z = u(\mathbf{x})$ in $\mathbf{x} = \mathbf{x}_0$ is given by

$$z = (\mathbf{x} - \mathbf{x}_0) \cdot \nabla u(\mathbf{x}_0) + u(\mathbf{x}_0). \quad (4.13)$$

A cross-section of the surface $u(\mathbf{x})$ and tangent plane is drawn in Figure 4.1. For a strictly convex function u , for any \mathbf{y} , the function $\mathbf{x} \mapsto -\mathbf{x} \cdot \mathbf{y} + u(\mathbf{x})$ is strictly convex in \mathbf{x} , and thus the infimum in (4.11b) is unique. As a result, we can write

$$u(\mathbf{x}_0) = \mathbf{x}_0 \cdot \mathbf{y} + u^*(\mathbf{y}), \quad (4.14)$$

where $\mathbf{y} = \nabla u(\mathbf{x}_0)$ for some unique $\mathbf{x}_0 \in \mathcal{X}$. Thus, the coordinates $\mathbf{y} \in \mathcal{Y}$ are the slopes of the tangent planes to $u(\mathbf{x})$. From (4.14) we can also show that the intercept with the z -axis of the tangent plane is $u^*(\mathbf{y})$ as follows. Substituting $\mathbf{x} = \mathbf{0}$ in (4.13) gives $z = -\mathbf{x}_0 \cdot \nabla u(\mathbf{x}_0) + u(\mathbf{x}_0)$. Subsequently, substituting $u(\mathbf{x}_0)$ from (4.14) and $\mathbf{y} = \nabla u(\mathbf{x}_0)$ into this expression gives $z = u^*(\mathbf{y})$. This value $z = u^*(\mathbf{y})$ is unique, since the infimum in (4.11b) is unique.

Combining (4.13) and (4.14) and using $\mathbf{y} = \nabla u(\mathbf{x}_0)$, the family of tangent planes to the function $u(\mathbf{x})$ is given by

$$F(\mathbf{x}, \mathbf{y}, z) := z - \mathbf{x} \cdot \mathbf{y} - u^*(\mathbf{y}) = 0, \quad (4.15)$$

parametrized by \mathbf{y} . We can find the graph of the original function $u(\mathbf{x})$ back as the envelope of this family of tangent planes by requiring

$$\left. \frac{\partial F(\mathbf{x}, \mathbf{y}, z)}{\partial \mathbf{y}} \right|_{\mathbf{x}=\mathbf{x}_0} = -\mathbf{x}_0 - \nabla_{\mathbf{y}} u^*(\mathbf{y}) = \mathbf{0}, \quad (4.16)$$

where $\nabla_{\mathbf{y}}$ denotes the gradient with respect to \mathbf{y} . We keep writing ∇ without subscript to denote the gradient with respect to \mathbf{x} . From (4.16) we see that the relation $\mathbf{y} = \nabla u(\mathbf{x}_0)$ can be inverted to $\mathbf{x}_0 = -\nabla_{\mathbf{y}} u^*(\mathbf{y})$, which is the stationary point of the transform in (4.11a) with $w(\mathbf{y}) = u^*(\mathbf{y})$.

For the optical systems considered in this thesis, the source domain \mathcal{X} and target domain \mathcal{Y} are compact. For the parallel-to-far-field problem explained in Section 3.2 we have the relation $u(\mathbf{x}) = \mathbf{x} \cdot \mathbf{y} + w(\mathbf{y})$, cf. (3.37), and we would like to find $\mathbf{y} = \mathbf{m}(\mathbf{x})$ such that the integral condition (4.7) is satisfied.

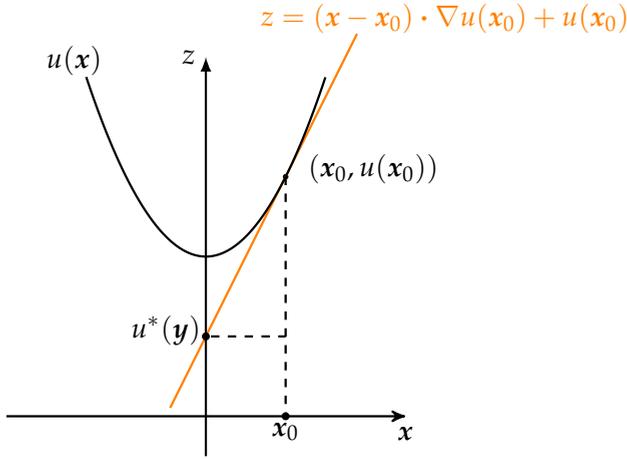


Figure 4.1: Cross-section of a strictly convex function $u(x)$, whose graph can be supported from below by the tangent functions.

We assume that u and w are a so-called *conjugate pair*, defined as

$$\forall x \in \mathcal{X} : \quad u(x) = \max_{y \in \mathcal{Y}} (x \cdot y + w(y)), \quad (4.17a)$$

$$\forall y \in \mathcal{Y} : \quad w(y) = \min_{x \in \mathcal{X}} (-x \cdot y + u(x)). \quad (4.17b)$$

Then $u^{**} = u$ is automatically satisfied. By the Fenchel-Moreau theorem [42] we know that u is convex and lower-semicontinuous.

We attempt to compute $u(x)$ such that the maximum and minimum in (4.17) are unique. Assuming u is twice differentiable, we find the minimum of (4.17b) as $y = \nabla u(x)$, and substituting this expression into (4.8) gives the standard Monge-Ampère equation

$$\det(D^2u(x)) = \frac{f(x)}{g(\nabla u(x))}, \quad (4.18)$$

where $D^2u(x)$ is the Hessian of u . The standard Monge-Ampère equation describes the parallel-to-far-field single surface systems (System 2 and 10), fully written for System 2 in Equation (3.53) with the inclusion of the stereographic Jacobian. System 4 can also be described using the standard Monge-Ampère equation [162].

A sufficient condition for a minimum in (4.17b) is that $P = -D^2u(x)$ is symmetric negative semi-definite (SND). We choose to introduce a minus sign

to conform with the SND/SPD matrix P that we will introduce later in this chapter for c -convex and G -convex pairs.

Following the arguments above, it is no surprise that the Legendre-Fenchel transform u^* also satisfies a standard Monge-Ampère equation involving the inverse mapping $x = (\nabla u)^{-1}(y) = -\nabla_y u^*(y)$ [124, p. 94–96].

We can repeat the above discussion by swapping the supremum and infimum in the definitions of the Legendre transforms (4.11) and consider a concave solution u . Thus, using the Fenchel-Moreau theorem [42] the transform (4.17) becomes

$$\forall x \in \mathcal{X} : \quad u(x) = \min_{y \in \mathcal{Y}} (x \cdot y + w(y)), \quad (4.19a)$$

$$\forall y \in \mathcal{Y} : \quad w(y) = \max_{x \in \mathcal{X}} (-x \cdot y + u(x)). \quad (4.19b)$$

In particular, if u is strictly concave, then $-x \cdot y + u(x)$ is strictly concave in x and the maximum $y = \nabla u$ is unique. A sufficient condition for a maximum in (4.19b) requires $P = -D^2u(x)$ to be symmetric positive semi-definite (SPD).

4.3 C-convex analysis

In this section, we generalize the Legendre-Fenchel transform to a so-called c -transform, which can be written down for optical systems with a cost function in optimal transport theory. The conjugate pairs in the previous section generalize to a c -convex and c -concave pair. Subsequently, we explain some theoretical background on cost functions and conjugate pairs in optimal transport theory. Lastly, in Section 4.3.1, we will derive the mapping $y = m(x)$ implicitly and we present conditions on the computation of a c -convex/ c -concave solution u_1 .

The standard Monge-Ampère equation corresponds to the cost function $c(x, y) = -x \cdot y$. If we replace $-x \cdot y$ in the Legendre-Fenchel transform (4.11) by a general cost function and let $u_1(x) = u(x)$ and $u_2(y) = w(y)$ we obtain the c -transform $u_2^* : \mathcal{X} \rightarrow \mathbb{R}$ of a function $u_2 : \mathcal{Y} \rightarrow \mathbb{R}$ defined as

$$\forall x \in \mathcal{X} : \quad u_2^*(x) = \sup_{y \in \mathcal{Y}} (-c(x, y) + u_2(y)). \quad (4.20a)$$

The c -transform $u_1^* : \mathcal{Y} \rightarrow \mathbb{R}$ of a function $u_1 : \mathcal{X} \rightarrow \mathbb{R}$ is defined as

$$\forall y \in \mathcal{Y} : \quad u_1^*(y) = \inf_{x \in \mathcal{X}} (c(x, y) + u_1(x)). \quad (4.20b)$$

Definition 4.3.1 (c-convexity). A real-valued function $u_1 : \mathcal{X} \rightarrow \mathbb{R}$ is called *c-convex* if there exists $u_2 : \mathcal{Y} \rightarrow \mathbb{R}$ such that $u_1 = u_2^*$ as defined in (4.20a). Defining a conjugate pair (u_1, u_2) as $u_1 = u_2^*$ and $u_1^* = u_2$ gives $u_1^{**} = u_1$. Assuming \mathcal{X} and \mathcal{Y} are compact, we define a *c-convex pair* as

$$\forall \mathbf{x} \in \mathcal{X} : \quad u_1(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{Y}} (u_2(\mathbf{y}) - c(\mathbf{x}, \mathbf{y})), \quad (4.21a)$$

$$\forall \mathbf{y} \in \mathcal{Y} : \quad u_2(\mathbf{y}) = \min_{\mathbf{x} \in \mathcal{X}} (u_1(\mathbf{x}) + c(\mathbf{x}, \mathbf{y})). \quad (4.21b)$$

Definition 4.3.2 (c-concavity). A function $u_1 : \mathcal{X} \rightarrow \mathbb{R}$ is called *c-concave* if the supremum and infimum in (4.20a) and (4.20b) are interchanged and there exists $u_2 : \mathcal{Y} \rightarrow \mathbb{R}$ such that $u_1 = u_2^*$. Defining a conjugate pair (u_1, u_2) as $u_1 = u_2^*$ and $u_1^* = u_2$ gives $u_1^{**} = u_1$. Assuming \mathcal{X} and \mathcal{Y} are compact, we define a *c-concave pair* as

$$\forall \mathbf{x} \in \mathcal{X} : \quad u_1(\mathbf{x}) = \min_{\mathbf{y} \in \mathcal{Y}} (u_2(\mathbf{y}) - c(\mathbf{x}, \mathbf{y})), \quad (4.22a)$$

$$\forall \mathbf{y} \in \mathcal{Y} : \quad u_2(\mathbf{y}) = \max_{\mathbf{x} \in \mathcal{X}} (u_1(\mathbf{x}) + c(\mathbf{x}, \mathbf{y})). \quad (4.22b)$$

We have not given u_2 a name in the above definitions. For a c-convex pair we have that u_1 is c-convex, but we could call u_2 c-concave since we have $u_1^* = u_2$ and we consider a sup/inf pair in (4.20). Similarly, using this convention for a c-concave pair, a c-concave u_1 is paired with a c-convex u_2 .

Example 4.3.1. The Legendre-Fenchel transform in (4.11) corresponds to the cost function $c(\mathbf{x}, \mathbf{y}) = -\mathbf{x} \cdot \mathbf{y}$. c-convexity means regular convexity in this case, since $u_1^* = u_2$ gives

$$u_1(\mathbf{x}_0) = \mathbf{x}_0 \cdot \mathbf{y}_0 + u_2(\mathbf{y}_0),$$

where $\mathbf{y}_0 = \nabla u_1(\mathbf{x}_0)$. Rearranging gives $u_2(\mathbf{y}_0) = -\mathbf{x}_0 \cdot \mathbf{y}_0 + u_1(\mathbf{x}_0)$. Using this equation and that $u_1 = u_2^*$ gives

$$u_1(\mathbf{x}) \geq \mathbf{x} \cdot \mathbf{y}_0 + u_2(\mathbf{y}_0) = \mathbf{x} \cdot \mathbf{y}_0 - \mathbf{x}_0 \cdot \mathbf{y}_0 + u_1(\mathbf{x}_0),$$

which results in the inequality

$$u_1(\mathbf{x}) - u_1(\mathbf{x}_0) \geq (\mathbf{x} - \mathbf{x}_0) \cdot \mathbf{y}_0.$$

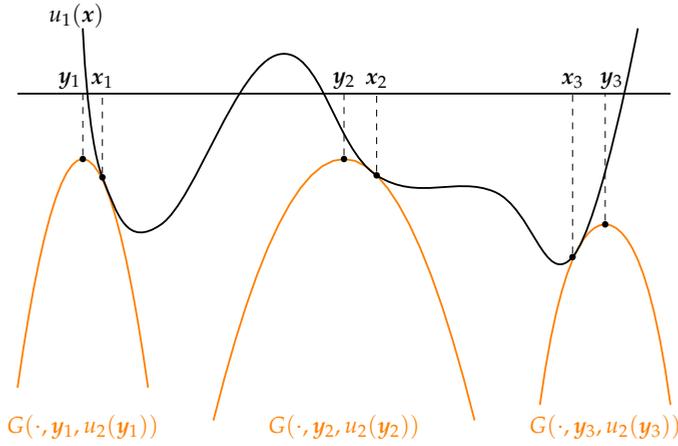


Figure 4.2: $u_1(x)$ is a function whose graph can be supported from below by the functions $G(\cdot, \mathbf{y}, u_2(\mathbf{y})) = -c(\cdot, \mathbf{y}) + u_2(\mathbf{y})$, and for all $x_0 \in \mathcal{X}$ there exists $\mathbf{y}_0 \in \mathcal{Y}$ such that $u_1(x_0) = G(x_0, \mathbf{y}_0, u_2(\mathbf{y}_0))$.

Substituting $\mathbf{y}_0 = \nabla u_1(x_0)$ gives that $u_1(x)$ should lie above its tangent planes, which holds if and only if u_1 is convex.

Using the same reasoning as in the previous section, for a conjugate pair (u_1, u_2) we have that u_1 is enveloped by graphs of the generating function $G(\cdot, \mathbf{y}, u_2(\mathbf{y})) = -c(\cdot, \mathbf{y}) + u_2(\mathbf{y})$. For a c-convex function, for any $x_0 \in \mathcal{X}$ there exists a $\mathbf{y}_0 \in \mathcal{Y}$ such that

$$u_1(x_0) = -c(x_0, \mathbf{y}_0) + u_2(\mathbf{y}_0), \quad (4.23)$$

$$\forall x \in \mathcal{X} : \quad u_1(x) \geq -c(x, \mathbf{y}_0) + u_2(\mathbf{y}_0), \quad (4.24)$$

i.e., u_1 is supported from below by graphs of the function $-c(x, \mathbf{y}) + u_2(\mathbf{y})$, as shown in Figure 4.2. Frequently c-convex and c-concave functions arise in the field of optimal transport theory [148, p. 54].

Optimal transport theory

In optimal transport theory, the main problem is to find a mapping $\mathbf{m} : \mathcal{X} \rightarrow \mathcal{Y}$ that transforms a measure μ on \mathcal{X} to a measure ν on \mathcal{Y} , where \mathcal{X} and \mathcal{Y} are two separable metric spaces and the measures $\mu \in \mathcal{R}(\mathcal{X})$ and $\nu \in \mathcal{R}(\mathcal{Y})$ are probability measures. Let $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ be a Borel-measurable function. The aim is to find a mapping $\mathbf{m} : \mathcal{X} \rightarrow \mathcal{Y}$ that attains the infimum

$$\inf_{\mathbf{m} \in \mathcal{M}} I[\mathbf{m}] = \inf_{\mathbf{m} \in \mathcal{M}} \left\{ \int_{\mathcal{X}} c(x, \mathbf{m}(x)) \, d\mu(x) \mid \mathbf{m}_\#(\mu) = \nu \right\}, \quad (4.25)$$

where \mathcal{M} is the set of all measure-preserving mappings, $m_{\#}(\mu)$ denotes the push-forward of μ by m (recall $\nu = \mu \circ m^{-1}$ in (4.6)), and $c(x, y)$ is a cost function which indicates how much it costs to move one unit of mass from $x \in \mathcal{X}$ to $y \in \mathcal{Y}$.

Problem (4.25) was first formulated by the French mathematician *Gaspard Monge* (1781) [107], also regarded as the father of differential geometry. Originally Monge described the problem as the transport of a pile of soil or *déblais* to an excavation or *remblais*. Monge's formulation of the optimal transportation problem can be ill-posed, because sometimes there is no m satisfying $m_{\#}(\mu) = \nu$, e.g., when μ is a Dirac measure but ν is not.

The Soviet mathematician *Leonid Vitaliyevich Kantorovich* (1942) reformulated the problem as an infinite-dimensional linear-programming problem by relaxing the problem to the *Monge-Kantorovich problem*

$$\inf_{\gamma \in \Gamma(\mu, \nu)} J[\gamma] = \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\gamma(x, y), \quad (4.26)$$

where $\Gamma(\mu, \nu)$ is the collection of all probability measures on $\mathcal{X} \times \mathcal{Y}$ with marginals μ on \mathcal{X} and ν on \mathcal{Y} , which means that for all subsets $\mathcal{A} \subset \mathcal{X}$ we have $\gamma(\mathcal{A} \times \mathcal{Y}) = \mu(\mathcal{A})$ and for all subsets $\mathcal{B} \subset \mathcal{Y}$ we have $\gamma(\mathcal{X} \times \mathcal{B}) = \nu(\mathcal{B})$. We can interpret $\Gamma(\mu, \nu)$ as the set of all possible couplings between μ and ν , i.e., it includes *mass splitting*: $x \in \mathcal{X}$ can be mapped to multiple points $y \in \mathcal{Y}$ and vice versa. A solution can always be found, since a possible solution is for instance $\mu \otimes \nu$, which is the product measure on $\mathcal{X} \times \mathcal{Y}$, which can be interpreted as every point $x \in \mathcal{X}$ is mapped to every point $y \in \mathcal{Y}$. It is defined such that

$$\int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d(\mu \otimes \nu)(x, y) = \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\mu(x) d\nu(y). \quad (4.27)$$

It can be shown [2, p. 135] that a minimizer for this problem always exists when the cost function c is lower semi-continuous.

For all the cost functions which we considered in Table 3.1 we can formulate the Monge-Kantorovich problem. All cost functions are continuous and differentiable, and so also lower semi-continuous.

An important theorem in optimal transport theory is the duality theorem for the Monge-Kantorovich problem.

Theorem 1 (Kantorovich duality theorem). *The infimum of $J[\gamma]$ in the Monge-Kantorovich problem is equal to*

$$\sup_{u_1, u_2} L[u_1, u_2] = \sup_{u_1, u_2} \left\{ - \int_{\mathcal{X}} u_1(\mathbf{x}) \, d\mu(\mathbf{x}) + \int_{\mathcal{Y}} u_2(\mathbf{y}) \, d\nu(\mathbf{y}) \mid -u_1(\mathbf{x}) + u_2(\mathbf{y}) \leq c(\mathbf{x}, \mathbf{y}) \right\}, \quad (4.28)$$

i.e., $\inf J[\gamma] = \sup L[u_1, u_2]$, where the supremum runs over all bounded and continuous functions $u_1 : \mathcal{X} \rightarrow \mathbb{R}$ and $u_2 : \mathcal{Y} \rightarrow \mathbb{R}$ subject to a constraint.

Proof of Theorem 1. We refer to the optimal-transport book written by the French politician and mathematician *Cédric Villani* [148, p. 57–75] for the proof. \square

Corollary 1.1. *A mapping $\tilde{\gamma} \in \Gamma(\mu, \nu)$ is optimal, i.e., attains the minimal $J[\gamma]$ for a bounded and continuous cost c , if and only if there is a c -convex u_1 such that $u_1^*(\mathbf{y}) - u_1(\mathbf{x}) = c(\mathbf{x}, \mathbf{y})$ [148, p. 58].*

We can interpret the dual Kantorovich problem informally [148, p. 53]. We will refer to this interpretation as the *shipper's problem*. Imagine my supervisor Jan wants to minimize the transportation cost of the export of a large amount of wholesale flowers to multiple countries. Jan can hire boats for the transport, but he has to pay them $c(\mathbf{x}, \mathbf{y})$ for each amount of flowers which is transported from warehouse \mathbf{x} to country \mathbf{y} . A shipper offers to take care of the transportation problem, buying flowers at warehouses and selling them to buyers overseas. What happens along the journey is not Jan's concern. Let $u_1(\mathbf{x})$ be the price at which a flower container is bought at warehouse \mathbf{x} , and $u_2(\mathbf{y})$ the price at which the container is sold in country \mathbf{y} . In total, the price which the warehouse and country together pay for the transport is $u_2(\mathbf{y}) - u_1(\mathbf{x})$, i.e., sales price minus buying price, which includes the shipper's profit, instead of the planned cost $c(\mathbf{x}, \mathbf{y})$. This is for each unit of flower: if there is mass $d\mu(\mathbf{x})$ at \mathbf{x} , then the total price for the transport from that location will be $u_1(\mathbf{x}) \, d\mu(\mathbf{x})$. Note that the total mass of all locations $\mathbf{x} \in \mathcal{X}$ is $\int_{\mathcal{X}} d\mu(\mathbf{x})$ and the total buying price is $\int_{\mathcal{X}} u_1(\mathbf{x}) \, d\mu(\mathbf{x})$. Similarly, the total mass of all locations $\mathbf{y} \in \mathcal{Y}$ is $\int_{\mathcal{Y}} d\nu(\mathbf{y})$ and the total sales price is $\int_{\mathcal{Y}} u_2(\mathbf{y}) \, d\nu(\mathbf{y})$. The shipper wants to be competitive, i.e., Jan wants to hire the shipper to minimize his cost, which results in

$$\forall \mathbf{x} \in \mathcal{X}, \forall \mathbf{y} \in \mathcal{Y} : \quad u_2(\mathbf{y}) - u_1(\mathbf{x}) \leq c(\mathbf{x}, \mathbf{y}). \quad (4.29)$$

The shipper wants to maximize his/her/their profits, which naturally leads to the dual Kantorovich problem in (4.28).

Example 4.3.2. *The parallel-to-far-field reflector and lens (System 2 and 10) have the generating function $G(\mathbf{x}, \mathbf{y}, w) = \mathbf{x} \cdot \mathbf{y} + w(\mathbf{y}) = u(\mathbf{x})$. The cost function is $c(\mathbf{x}, \mathbf{y}) = -\mathbf{x} \cdot \mathbf{y}$. Then the c -transform is just the Legendre-Fenchel transform with $u_1(\mathbf{x}) = u(\mathbf{x})$ and $u_2(\mathbf{y}) = w(\mathbf{y})$. However, we also proceeded by defining $u_1(\mathbf{x}) = u(\mathbf{x}) - \frac{1}{2}|\mathbf{x}|^2$ and $u_2(\mathbf{y}) = w(\mathbf{y}) + \frac{1}{2}|\mathbf{y}|^2$, deriving the relation*

$$u_2(\mathbf{y}) - u_1(\mathbf{x}) = -\frac{1}{2}|\mathbf{x} - \mathbf{y}|^2 = \tilde{c}(\mathbf{x}, \mathbf{y}).$$

We rewrote the cost function to the quadratic form, because this cost function frequently arises in classical optimal transport results. The French mathematician Yann Brenier (1987) (among others [25, 33, 38, 126, 136, 140, 146]) proved a theorem, which is now known as Brenier's theorem, that guarantees the existence of a unique mapping \mathbf{m} satisfying (4.7), and minimizing (4.25), which is the gradient of a convex potential [17].

Existence, uniqueness and regularity

Under certain conditions of differentiability of c , combined with the injectivity of $\nabla_{\mathbf{x}}c(\mathbf{x}, \cdot)$ (twist condition), there exist both a conjugate pair (u_1, u_2) that maximizes the Kantorovich functional L and an associated optimal mapping \mathbf{m} minimizing (4.25) which is unique. In fact, there exists a c -convex or c -concave function u_1 such that

$$\nabla u_1(\mathbf{x}) + \nabla_{\mathbf{x}}c(\mathbf{x}, \mathbf{m}(\mathbf{x})) = \mathbf{0}, \quad (4.30)$$

which is the stationary point of (4.21b) or (4.22b). The mapping $\mathbf{m}(\mathbf{x})$ can be found using the implicit function theorem if the twist condition is satisfied. More details can be found in Villani's book [148, p. 215–267]. The theory of existence and uniqueness of the solution of optimal transport problems has also been studied in [30, 54, 92, 96]. Existence proofs aim to show that the infima in (4.25) and (4.26) are equal. This holds under convexity assumptions on the cost function [30]. Conditions for existence, uniqueness and smoothness of a solution to reflector-type problems with a point source were derived in [62, 67, 119, 151, 152], which also have a logarithmic cost function. The existence and uniqueness of weak solutions to the point-to-far-field lens system is established in [71, 72].

The regularity of the functions (u_1, u_2) solving the optimal transport problem can only hold under certain stringent assumptions on the cost function and on the geometry of the target domain. The assumptions on the geometry

include a generalization of a convex set in (4.2.1) to a *c-convex set*, where any two points in the target domain can be connected by a so-called *c-segment*, which is the image of a usual line segment in \mathcal{Y} by a map $(\nabla_x c(x, \cdot))^{-1}$.

There are various counterexamples to the regularity of optimal transport problems, e.g., Caffarelli's counterexample [148, p. 283] and Loeper's counterexample [148, p. 285]. Caffarelli's counterexample gives an example of a discontinuous optimal mapping for smooth and compactly supported densities f and g , and Loeper's counterexample shows that geometric obstructions in the target domain prevent smoothness. Regularity estimations of solutions (u_1, u_2) follow from the Urbas-Trudinger-Wang regularity theory [148, p. 318] and Loeper-Ma-Trudinger-Wang regularity theory [148, p. 318]. The regularity of the cost function is sufficient to build a strong regularity theory. Proving existence, uniqueness and regularity for each optical system is beyond the scope of this thesis.

4.3.1 An optimal-transport mapping

For optimal-transport relations of the form $u_2(\mathbf{y}) - u_1(\mathbf{x}) = c(\mathbf{x}, \mathbf{y})$ we are looking for a mapping $\mathbf{y} = \mathbf{m}(\mathbf{x})$ which maps \mathcal{X} to \mathcal{Y} and satisfies the Jacobian equation (4.8). In the previous section, we motivated that such a mapping exists and can be found using a *c-convex/c-concave pair* (4.21) or (4.22) as the stationary point of $u_1(\mathbf{x}) + c(\mathbf{x}, \mathbf{y})$. In this section, we will present the conditions on deriving the mapping implicitly and find the general formulation of a *generalized Monge-Ampère equation*.

For continuously differentiable cost functions, the expression for the optical map $\mathbf{y} = \mathbf{m}(\mathbf{x})$ is implicitly given by (4.30), under the condition that the mixed Hessian matrix C , defined by

$$C = C(\mathbf{x}, \mathbf{m}(\mathbf{x})) = D_{xy}c = \begin{pmatrix} \frac{\partial^2 c}{\partial x_1 \partial y_1} & \frac{\partial^2 c}{\partial x_1 \partial y_2} \\ \frac{\partial^2 c}{\partial x_2 \partial y_1} & \frac{\partial^2 c}{\partial x_2 \partial y_2} \end{pmatrix} \quad (4.31)$$

is invertible. This invertibility is required for the mapping \mathbf{y} , defined as $\mathbf{y} = \mathbf{m}(\mathbf{x}) = (\nabla_x c(x, \cdot))^{-1} \circ (-\nabla u_1(x))$, to exist locally, where $(\nabla_x c)^{-1}$ is the local inverse of $(\mathbf{x}, \mathbf{y}) \mapsto \nabla_x c(\mathbf{x}, \mathbf{y})$ for fixed \mathbf{x} . In other words, for every point \mathbf{y} there exists a small neighbourhood for which $D_{xy}c(\mathbf{x}, \mathbf{y})$ is invertible. This condition for the existence of a globally injective mapping is called the *twist condition* [148, p. 234]. In Section 4.5 we will discuss the requirements for a globally invertible mapping \mathbf{m} , i.e., a bijective mapping, for the point-to-far-field reflector problem in particular.

In fact, we can verify by substituting the expression for the cost function c of an optical system into (4.30) and solving for $\mathbf{y} = \mathbf{m}(\mathbf{x})$, that this implicit

mapping is identical to the mapping derived via the law of reflection or refraction, e.g., the mappings in (3.52) and (3.99). This holds for all optical systems considered in this thesis.

A sufficient condition for a minimum/maximum solution in (4.21b) and (4.22b) requires

$$-D^2u_1(x) - D_{xx}c(x, \mathbf{m}(x)) = \mathbf{P}(x), \quad (4.32)$$

to be SND/SPD, respectively, where D^2u_1 is the Hessian matrix of u_1 and $D_{xx}c$ is the Hessian matrix of c with respect \mathbf{x} . Hence, the requirements for a c-convex or c-concave solution can be written down as follows:

- For a c-convex pair we require \mathbf{P} to be SND and consequently, $\text{tr}(\mathbf{P}) \leq 0$ and $\det(\mathbf{P}) \geq 0$.
- For a c-concave pair we need an SPD matrix \mathbf{P} with $\text{tr}(\mathbf{P}) \geq 0$ and $\det(\mathbf{P}) \geq 0$.

Note that \mathbf{P} is a symmetric matrix. We choose to introduce a minus sign in (4.32), to conform with the SND/SPD matrix \mathbf{P} that we will introduce later in this chapter for G-convex pairs.

Differentiating (4.30) again with respect to \mathbf{x} gives

$$D_{xx}c(x, \mathbf{m}(x)) + \mathbf{C}(x, \mathbf{m}(x)) D\mathbf{m}(x) + D^2u_1(x) = \mathbf{O}, \quad (4.33)$$

where $D\mathbf{m}(x)$ is the 2×2 Jacobi matrix of \mathbf{m} with respect to \mathbf{x} . Combining (4.32) and (4.33) gives the matrix equation

$$\mathbf{C}(x, \mathbf{m}(x)) D\mathbf{m}(x) = \mathbf{P}(x). \quad (4.34)$$

Combining (4.34) with the Jacobian equation of energy conservation (4.8) gives

$$\det(D\mathbf{m}(x)) = \frac{\det(\mathbf{P}(x))}{\det(\mathbf{C}(x, \mathbf{m}(x)))} = \frac{f(x)}{g(\mathbf{m}(x))}. \quad (4.35)$$

Substituting \mathbf{P} as defined in (4.32) and \mathbf{C} in (4.31) gives

$$\det(D^2u_1(x) + D_{xx}c(x, \mathbf{m}(x))) = \det(D_{xy}c(x, \mathbf{m}(x))) \frac{f(x)}{g(\mathbf{m}(x))}, \quad (4.36)$$

since $\det(-\mathbf{P}) = \det(\mathbf{P})$ for a 2×2 matrix \mathbf{P} . Equation (4.36) is the general form of a *generalized Monge-Ampère equation*. We will use the matrices \mathbf{P} and \mathbf{C} in our numerical method in Chapter 6.

In this section, we wrote the mapping $\mathbf{y} = (\nabla_x c(x, \cdot))^{-1} \circ (-\nabla u_1(x))$ as $\mathbf{y} = \mathbf{m}(x)$, i.e., since $\nabla u_1(x)$ depends on \mathbf{x} we wrote it as a function of \mathbf{x}

only. For instance, we found the mapping of the point-to-far-field reflector in Section 3.4 as

$$\mathbf{y} = \mathbf{m}(\mathbf{x}) = \frac{-2 \nabla u_1(\mathbf{x}) + \mathbf{x} |\nabla u_1(\mathbf{x})|^2}{4 - 4 \mathbf{x} \cdot \nabla u_1(\mathbf{x}) + (|\mathbf{x}| |\nabla u_1(\mathbf{x})|)^2}, \quad (4.37)$$

which can also be written as $\mathbf{y} = \mathbf{m}(\mathbf{x}) = \tilde{\mathbf{m}}(\mathbf{x}, \nabla u(\mathbf{x}))$ since $u_1(\mathbf{x})$ is related to $u(\mathbf{x})$ by a simple change of variables.

4.4 G-convex analysis

We go back to the relation $u(\mathbf{x}) = G(\mathbf{x}, \mathbf{y}, w(\mathbf{y}))$, for a generic generating function G , where $u(\mathbf{x})$ is the optical surface we are looking for. In this section, we will present the theory on *generated Jacobian equations* (GJE), coined by the Australian mathematician *Neil Trudinger* (2012) [145]. These equations have their own notions of transforms and convexity.

We denote the mapping most generally as $\mathbf{y} = \mathbf{m}(\mathbf{x}) = \bar{\mathbf{m}}(\mathbf{x}, u(\mathbf{x}), \nabla u(\mathbf{x}))$, now also as a function of u . For example, the mapping for the parallel-to-near-field system in Section 3.3 is

$$\mathbf{y} = \mathbf{m}(\mathbf{x}) = \bar{\mathbf{m}}(\mathbf{x}, u(\mathbf{x}), \nabla u(\mathbf{x})) = \mathbf{x} + \frac{2 u(\mathbf{x}) \nabla u(\mathbf{x})}{1 - |\nabla u(\mathbf{x})|^2}. \quad (4.38)$$

Generalizing the concept of transforms even further, we introduce the G-transform $w^* : \mathcal{X} \rightarrow \mathbb{R}$ of $w : \mathcal{Y} \rightarrow \mathbb{R}$ and H-transform $u^* : \mathcal{Y} \rightarrow \mathbb{R}$ of $u : \mathcal{X} \rightarrow \mathbb{R}$ as

$$\forall \mathbf{x} \in \mathcal{X} : \quad w^*(\mathbf{x}) = \sup_{\mathbf{y} \in \mathcal{Y}} G(\mathbf{x}, \mathbf{y}, w(\mathbf{y})), \quad (4.39a)$$

$$\forall \mathbf{y} \in \mathcal{Y} : \quad u^*(\mathbf{y}) = \inf_{\mathbf{x} \in \mathcal{X}} H(\mathbf{x}, \mathbf{y}, u(\mathbf{x})), \quad (4.39b)$$

where $u(\mathbf{x})$ is the optical surface and $H(\mathbf{x}, \mathbf{y}, G(\mathbf{x}, \mathbf{y}, w(\mathbf{y}))) = w(\mathbf{y})$ is the unique inverse of G for a given $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$, assuming the unique inverse exists. In Chapter 3, we derived G and H for a range of optical systems such that for all $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$ we have

$$u(\mathbf{x}) = G(\mathbf{x}, \mathbf{y}, w(\mathbf{y})) \iff w(\mathbf{y}) = H(\mathbf{x}, \mathbf{y}, u(\mathbf{x})), \quad (4.40)$$

i.e., for fixed \mathbf{x}, \mathbf{y} , we have that $G(\mathbf{x}, \mathbf{y}, \cdot)$ and $H(\mathbf{x}, \mathbf{y}, \cdot)$ are each other's inverses. We get that for all $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}, w \in \mathbb{R}$ we have $G_w > 0$ or $G_w < 0$, since G should be injective w.r.t. the third argument ($G_w = 0$ is possible at an isolated point).

We could also define the G-transform and H-transform as an inf/sup pair, sup/sup pair, or inf/inf pair, which will be discussed later in this section.

In all derivations of generating functions, we consider a scalar function $u : \mathcal{X} \rightarrow \mathbb{R}$ and a vector function $\mathbf{p} : \mathcal{X} \rightarrow \mathbb{R}^2$, representing $\nabla u(\mathbf{x})$, such that $\mathbf{y} = \mathbf{m}(\mathbf{x}) = \bar{\mathbf{m}}(\mathbf{x}, u(\mathbf{x}), \mathbf{p}(\mathbf{x}))$ for some $\bar{\mathbf{m}} : \mathcal{X} \times \mathbb{R} \times \mathbb{R}^2 \rightarrow \mathcal{Y}$. Using the chain rule on $\bar{\mathbf{m}} = \bar{\mathbf{m}}(\mathbf{x}, u, \mathbf{p})$ with variables $(\mathbf{x}, u, \mathbf{p}) \in \mathcal{X} \times \mathbb{R} \times \mathbb{R}^2$ gives

$$D\mathbf{m} = D_x \bar{\mathbf{m}} + \frac{\partial \bar{\mathbf{m}}}{\partial u} \otimes \nabla u + (D_p \bar{\mathbf{m}}) (D_x \mathbf{p}), \quad (4.41)$$

where $D_x \bar{\mathbf{m}}$ and $D_p \bar{\mathbf{m}}$ denote the matrices of first-order derivatives of $\bar{\mathbf{m}}$ with respect to \mathbf{x} and \mathbf{p} , respectively, $D_x \mathbf{p}$ is the matrix of first-order derivatives of \mathbf{p} with respect to \mathbf{x} , and $\frac{\partial \bar{\mathbf{m}}}{\partial u} \otimes \nabla u$ is the dyadic product of $\frac{\partial \bar{\mathbf{m}}}{\partial u}$ and ∇u , defined as $\frac{\partial \bar{\mathbf{m}}}{\partial u} (\nabla u)^T$.

Assuming $D_p \bar{\mathbf{m}}$ is invertible, we can write (4.41) as

$$D\mathbf{m} = (D_p \bar{\mathbf{m}}) (D_x \mathbf{p} - A(\mathbf{x}, u, \nabla u)), \quad (4.42a)$$

with

$$A(\mathbf{x}, u, \mathbf{p}) = -(D_p \bar{\mathbf{m}})^{-1} (D_x \bar{\mathbf{m}} + \frac{\partial \bar{\mathbf{m}}}{\partial u} \otimes \mathbf{p}). \quad (4.42b)$$

Note that we introduced two minus signs, in order to conform with the notation of Trudinger in [145]. Defining

$$\Psi(\mathbf{x}, u, \mathbf{p}) = \det((D_p \bar{\mathbf{m}})^{-1}(\mathbf{x}, u, \mathbf{p})) \frac{f(\mathbf{x})}{g(\bar{\mathbf{m}}(\mathbf{x}, u, \mathbf{p}))}, \quad (4.43a)$$

substituting (4.42a) into (4.8) and subsequently setting $\mathbf{p} = \nabla u$ results in the general formulation of a generated Jacobian equation as

$$\det(D^2 u(\mathbf{x}) - A(\mathbf{x}, u(\mathbf{x}), \nabla u(\mathbf{x}))) = \Psi(\mathbf{x}, u(\mathbf{x}), \nabla u(\mathbf{x})). \quad (4.43b)$$

If the mapping admits a solution $\mathbf{m}(\mathbf{x}) = \bar{\mathbf{m}}(\mathbf{x}, u(\mathbf{x}), \nabla u(\mathbf{x})) = \nabla u(\mathbf{x})$, it is easily verified that this equation reduces to the standard Monge-Ampère equation, since $D_x \bar{\mathbf{m}} = \mathbf{O}$, $\frac{\partial \bar{\mathbf{m}}}{\partial u} = \mathbf{0}$ and $D_p \bar{\mathbf{m}} = \mathbf{I}$. Consequently, we have $A(\mathbf{x}, u, \nabla u) = \mathbf{O}$ and $\Psi(\mathbf{x}, u, \mathbf{p}) = \frac{f(\mathbf{x})}{g(\nabla u(\mathbf{x}))}$.

The combination of a generated Jacobian equation with the boundary condition $\mathbf{m}(\mathcal{X}) = \mathcal{Y}$ is also known as a *second boundary value problem* [82]. We will show in Section 4.5 that this boundary condition is equivalent to the condition $\mathbf{m}(\partial \mathcal{X}) = \partial \mathcal{Y}$ under some convexity assumptions on the solution u .

In order to use (4.39) to derive the mapping \mathbf{m} , the generating function G needs to satisfy some conditions. In the following, we will give a formal

definition of a generating function. In Section 4.4.1, we will rewrite the generated Jacobian equation in (4.43b) in terms of the generating function. Subsequently, in Section 4.4.2, we derive an implicit relation for the mapping $\mathbf{y} = \mathbf{m}(\mathbf{x}) = \bar{\mathbf{m}}(\mathbf{x}, u(\mathbf{x}), \nabla u(\mathbf{x}))$.

Definition 4.4.1 (Generating function). *Consider the two-dimensional domains \mathcal{X}, \mathcal{Y} (open and bounded), which are subsets of \mathbb{R}^2 . Let G be a function*

$$G : \mathcal{G} \subset \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}. \quad (4.44)$$

Then G is called a generating function if it has the following properties:

1. G is C^1 in \mathcal{G} , which is an open set.
2. For all $(\mathbf{x}, \mathbf{y}, w) \in \mathcal{G}$, we have $G_w(\mathbf{x}, \mathbf{y}, w) < 0$ or $G_w(\mathbf{x}, \mathbf{y}, w) > 0$.
3. For a fixed $\mathbf{x} \in \mathcal{X}$, the map

$$(\mathbf{y}, w) \mapsto (\nabla_{\mathbf{x}} G(\mathbf{x}, \mathbf{y}, w), G(\mathbf{x}, \mathbf{y}, w)) \quad (4.45)$$

is differentiable and invertible in (\mathbf{y}, w) , and the inverse is also differentiable in (\mathbf{y}, w) .

The dual generating function H is defined by

$$G(\mathbf{x}, \mathbf{y}, H(\mathbf{x}, \mathbf{y}, u)) = u, \quad (4.46)$$

for all $(\mathbf{x}, \mathbf{y}, u)$ where G is well-defined and H also satisfies properties 1–3 above [69, 70].

We go back to the G -transform defined in (4.39). The pair $u : \mathcal{X} \rightarrow \mathbb{R}$ and $w : \mathcal{Y} \rightarrow \mathbb{R}$ is a conjugate pair if $u = w^*$ and $w = u^*$. The sup/inf pair as defined in (4.39) is only valid for a conjugate pair if $G_w > 0$. In Appendix A, we show that the condition $G_w > 0$ results in a max/min pair or a min/max pair for compact \mathcal{X} and \mathcal{Y} . After stating the definition of G -convexity below, we formally define these pairs. If $G_w < 0$, we get a max/max pair or a min/min pair. In the remainder of this chapter, we assume $G_w > 0$. The derivations for $G_w < 0$ are similar.

The interpretation of $u = w^*$ is that u is enveloped by graphs of functions $G(\cdot, \mathbf{y}, w(\mathbf{y}))$, as illustrated in Figure 4.2 replacing $u_1(\mathbf{x})$ with $u(\mathbf{x})$ and considering any $G(\mathbf{x}, \mathbf{y}, w)$ which is not necessarily of the optimal-transport form.

For any $\mathbf{x}_0 \in \mathcal{X}$ there exists a $\mathbf{y}_0 \in \mathcal{Y}$ such that

$$u(\mathbf{x}_0) = G(\mathbf{x}_0, \mathbf{y}_0, w(\mathbf{y}_0)), \quad (4.47a)$$

$$\forall \mathbf{x} \in \mathcal{X} : \quad u(\mathbf{x}) \geq G(\mathbf{x}, \mathbf{y}_0, w(\mathbf{y}_0)). \quad (4.47b)$$

Definition 4.4.2 (G-convexity). A real-valued function $u : \mathcal{X} \rightarrow \mathbb{R}$ is called G-convex if there exists $w : \mathcal{Y} \rightarrow \mathbb{R}$ such that $u = w^*$. A function $u : \mathcal{X} \rightarrow \mathbb{R}$ is called G-concave if the supremum and infimum in (4.39a) and (4.39b) are interchanged and there exists $w : \mathcal{Y} \rightarrow \mathbb{R}$ such that $u = w^*$.

A function $w : \mathcal{Y} \rightarrow \mathbb{R}$ is called H-concave if there exists $u : \mathcal{X} \rightarrow \mathbb{R}$ such that $w = u^*$. A function $w : \mathcal{Y} \rightarrow \mathbb{R}$ is called H-convex if the supremum and infimum in (4.39a) and (4.39b) are interchanged and there exists $u : \mathcal{Y} \rightarrow \mathbb{R}$ such that $w = u^*$.

It follows that if (u, w) form a conjugate pair as defined in (4.39), then u is G-convex and w is H-concave. If the supremum and infimum are interchanged, u is G-concave and w is H-convex.

Since the supremum and infimum are attained in (4.39a) and (4.39b) for compact \mathcal{X} and \mathcal{Y} , and we take \mathcal{X} and \mathcal{Y} to be compact sets, $u(\mathbf{x})$ is G-convex and $w(\mathbf{y})$ is H-concave if

$$\forall \mathbf{x} \in \mathcal{X} : \quad u(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{Y}} G(\mathbf{x}, \mathbf{y}, w(\mathbf{y})), \quad (4.48a)$$

$$\forall \mathbf{y} \in \mathcal{Y} : \quad w(\mathbf{y}) = \min_{\mathbf{x} \in \mathcal{X}} H(\mathbf{x}, \mathbf{y}, u(\mathbf{x})), \quad (4.48b)$$

or $u(\mathbf{x})$ is G-concave and $w(\mathbf{y})$ is H-convex if

$$\forall \mathbf{x} \in \mathcal{X} : \quad u(\mathbf{x}) = \min_{\mathbf{y} \in \mathcal{Y}} G(\mathbf{x}, \mathbf{y}, w(\mathbf{y})), \quad (4.49a)$$

$$\forall \mathbf{y} \in \mathcal{Y} : \quad w(\mathbf{y}) = \max_{\mathbf{x} \in \mathcal{X}} H(\mathbf{x}, \mathbf{y}, u(\mathbf{x})). \quad (4.49b)$$

Example 4.4.1. The Legendre-Fenchel transform in (4.11) corresponds to the generating function $G(\mathbf{x}, \mathbf{y}, w) = \mathbf{x} \cdot \mathbf{y} + w$ and $H(\mathbf{x}, \mathbf{y}, w) = -\mathbf{x} \cdot \mathbf{y} + w$. G-convexity means regular convexity in this case, since

$$u(\mathbf{x}_0) = \mathbf{x}_0 \cdot \mathbf{y}_0 + w_0,$$

for some $(\mathbf{x}_0, \mathbf{y}_0, w_0) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$. Rearranging gives $w_0 = -\mathbf{x}_0 \cdot \mathbf{y}_0 + u(\mathbf{x}_0)$ and using (4.47b) results in

$$u(\mathbf{x}) \geq \mathbf{x} \cdot \mathbf{y}_0 + w_0 = \mathbf{x} \cdot \mathbf{y}_0 - \mathbf{x}_0 \cdot \mathbf{y}_0 + u(\mathbf{x}_0),$$

and subsequently,

$$u(\mathbf{x}) - u(\mathbf{x}_0) \geq (\mathbf{x} - \mathbf{x}_0) \cdot \mathbf{y}_0.$$

Substituting $\mathbf{y}_0 = \nabla u(\mathbf{x}_0)$ gives that $u(\mathbf{x})$ should lie above its tangent planes, which holds if and only if u is convex.

In the following, we give a general definition of boundary value problems that are characterized by generating functions. First, we will give a definition for the G -exponential map for the mapping $\mathbf{y} = \overline{\mathbf{m}}(\mathbf{x}, u, \mathbf{p})$. Recall that for every fixed \mathbf{x} we require the map

$$(\mathbf{y}, w) \mapsto (\nabla_{\mathbf{x}}G(\mathbf{x}, \mathbf{y}, w), G(\mathbf{x}, \mathbf{y}, w)), \quad (4.50)$$

as defined in (4.45), to be invertible with a differentiable inverse.

Definition 4.4.3. For every pair of \mathbf{x} and $u(\mathbf{x})$ there is an open set in the tangent space $(T\mathcal{X})_{\mathbf{x}}$ of \mathcal{X} at \mathbf{x} and a differentiable map

$$\exp_{\mathbf{x},u} : (T\mathcal{X})_{\mathbf{x}} \rightarrow \mathcal{Y}, \quad (4.51)$$

such that we have

$$\nabla_{\mathbf{x}}G(\mathbf{x}, \mathbf{y}, H(\mathbf{x}, \mathbf{y}, u)) = \mathbf{p}, \quad (4.52)$$

where $\mathbf{y} = \exp_{\mathbf{x},u}(\mathbf{p})$. The map $\exp_{\mathbf{x},u}$ is called the G -exponential map at (\mathbf{x}, u) , i.e., the point $\mathbf{y} = \exp_{\mathbf{x},u}(\mathbf{p})$ is the point in \mathcal{Y} such that G has gradient \mathbf{p} at \mathbf{x} .

Example 4.4.2. The Legendre-Fenchel transform in (4.11) corresponds to the generating function $G(\mathbf{x}, \mathbf{y}, w) = \mathbf{x} \cdot \mathbf{y} + w$, resulting in

$$\exp_{\mathbf{x},u}(\mathbf{p}) = \mathbf{p}.$$

The shipper's problem or dual Kantorovich problem in optimal transport theory has the form $u_1(\mathbf{x}) = G(\mathbf{x}, \mathbf{y}, u_2(\mathbf{y})) = -c(\mathbf{x}, \mathbf{y}) + u_2(\mathbf{y})$. Hence, (4.52) with $\mathbf{y} = \exp_{\mathbf{x},u}(\mathbf{p})$ becomes

$$\mathbf{p} + \nabla_{\mathbf{x}}c(\mathbf{x}, \mathbf{p}) = \mathbf{0}.$$

Definition 4.4.4 (Second boundary value problem). We consider (1) the domains $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^2$, (2) a generating function $G : \mathcal{G} \subset \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}$, and (3) probability measures μ and ν . Recall, in our case $d\mu(\mathbf{x}) = f(\mathbf{x}) d\mathbf{x}$ and $d\nu(\mathbf{y}) = g(\mathbf{y}) d\mathbf{y}$. We aim to find a map $\mathbf{m} : \mathcal{X} \rightarrow \mathcal{Y}$ which is smooth and invertible given by

$$\mathbf{m}(x) = \exp_{x,u}(\nabla u(x)), \quad (4.53a)$$

since $\nabla_x G(x, \mathbf{y}, w) = \mathbf{p} = \nabla u(x)$, which will be shown in Section 4.4.1. The mapping for a G-convex function is such that $\nu = \mathbf{m}_\#(\mu)$, i.e., such that

$$\det(D\mathbf{m}(x)) = \frac{f(x)}{g(\mathbf{m}(x))}. \quad (4.53b)$$

We define the corresponding transport boundary condition as

$$\mathbf{m}(\mathcal{X}) = \mathcal{Y}. \quad (4.53c)$$

We say that u is a G-potential of \mathbf{m} if $\mathbf{m}(x) = \exp_{x,u}(\nabla u(x))$. The problem (4.53) is called a **second boundary value problem**, which means that the boundary condition is not of Dirichlet type. Instead, (4.53c) plays the role of boundary condition; see Section 4.5.

This is a generalization of the exponential map in Riemannian geometry where the mapping represents end points of geodesic segments [69]. The idea of introducing an exponential map is to obtain another notation for the mapping $\mathbf{y} = \mathbf{m}(x) = \bar{\mathbf{m}}(x, u, \nabla u) = \exp_{x,u}(\nabla u(x))$, i.e., the exponential denotes a function of x and u which takes as main argument the continuous compounding of small ‘actions’ ∇u .

Existence, uniqueness and regularity

The existence of smooth solutions to the second boundary value problem (4.53) has been studied by a number of authors [95, 145]. Their analysis follows the approach for the existence of dual potential functions in optimal transport theory. However, uniqueness for smooth or weak solutions has not been fully proven [69, 82, 145]. The existence of globally smooth solutions for the parallel-to-near-field reflector is established in [95], noting extensions to the lens problem. While a formulation as a linear Kantorovich problem for the parallel-to-near-field problem is not possible, a formulation as a nonlinear

Kantorovich problem and the corresponding generated Jacobian equation were derived in [68, 94].

Trudinger [145] generalizes the regularity theory for potentials (u, w) of Ma, Trudinger and Wang, using extended definitions of c -convex domains to G -convex domains. Karakhanyan et al. [83] prove the regularity for weak solutions for the point-to-near-field lens problem. Weak solutions to optimal transport problems are usually of the Aleksandrov- or Brenier-type. For a review of definitions, we refer to [147, p. 125–141]. There are still various open problems involving regularity, e.g., the regularity at the boundary of the source and target domains. Another question is whether a simple characterization of mappings m exists which admits a G -convex potential, if the generating function is not of the form $u_1(x) = G(x, y, u_2(y)) = -c(x, y) + u_2(y)$ [69].

4.4.1 Generated Jacobian equations

In this section, we proceed by rewriting the generated Jacobian equation (4.43b) in terms of the generating function G . We will obtain a general form of a generated Jacobian equation and will see that if we substitute the generating function $G(x, y, w) = -c(x, y) + w$ corresponding to optimal-transport problems into this equation we retrieve the generalized Monge-Ampère equation in (4.35).

If there is a solution u that is twice differentiable and G is twice differentiable, then the map $m(x) = \bar{m}(x, u(x), \nabla u(x)) = \exp_{x,u}(\nabla u(x))$ should be such that

$$\det(D\bar{m}(x, u(x), \nabla u(x))) = \frac{f(x)}{g(\bar{m}(x, u(x), \nabla u(x)))}. \quad (4.54)$$

Writing the mapping as $y = \exp_{x,u}(p) = \bar{m}(x, u, p)$ as a function of the variables $(x, u, p) \in \mathcal{X} \times \mathbb{R} \times \mathbb{R}^2$, we have the equation

$$G(x, \bar{m}(x, u, p), w) = u. \quad (4.55)$$

Writing $p = \nabla u(x)$ and differentiating this equation with respect to x gives

$$\nabla_x G + \left(D_x \bar{m} + \frac{\partial \bar{m}}{\partial u} \otimes \nabla u + (D_p \bar{m})(D^2 u) \right)^T (\nabla_y G + G_w \nabla_y w) = \nabla u, \quad (4.56)$$

where the dyadic product is the matrix $u \otimes v = u v^T$ for the vectors u, v . The stationary point of (4.48a) and (4.49a) for a G -convex and G -concave pair is given by

$$\nabla_y G + G_w \nabla_y w = \mathbf{0}. \quad (4.57)$$

Combining (4.56) and (4.57) gives

$$\nabla_x G(x, \bar{m}(x, u, \mathbf{p}), w) = \nabla u(x), \quad (4.58)$$

and we conclude that, indeed, using the property $\mathbf{p} = \nabla u(x)$ in Definition 4.4.4 was valid.

Consequently, for a solution u to (4.53) we can write down the system of equations

$$\nabla_x G(x, \bar{m}(x, u, \mathbf{p}), w) = \mathbf{p}, \quad (4.59a)$$

$$G(x, \bar{m}(x, u, \mathbf{p}), w) = u, \quad (4.59b)$$

where $w = H(x, \bar{m}(x, u, \mathbf{p}), u)$ and $\mathbf{p} = \nabla u$.

First, we consider the right-hand side of (4.43b). Differentiating (4.59a) and (4.59b) with respect to \mathbf{p} gives

$$(D_{xy}G) (D_{\mathbf{p}}\bar{m}) + \nabla_x G_w \left(\frac{\partial w}{\partial \mathbf{p}} \right)^T = \mathbf{I}, \quad (4.60a)$$

$$(D_{\mathbf{p}}\bar{m})^T \nabla_y G + G_w \frac{\partial w}{\partial \mathbf{p}} = \mathbf{0}. \quad (4.60b)$$

Substituting $w = H(x, \bar{m}(x, u, \mathbf{p}), u)$ gives

$$(D_{xy}G) (D_{\mathbf{p}}\bar{m}) + (\nabla_x G_w \otimes \nabla_y H) (D_{\mathbf{p}}\bar{m}) = \mathbf{I}, \quad (4.61a)$$

$$(D_{\mathbf{p}}\bar{m})^T \nabla_y G + G_w (D_{\mathbf{p}}\bar{m})^T \nabla_y H = \mathbf{0}, \quad (4.61b)$$

respectively. Note that D denotes matrices, while ∇ denotes vectors, and G_w is a scalar. Hence, from (4.61b), since $G_w \neq 0$ is a scalar by Definition 4.4.1, we have

$$\nabla_y H (D_{\mathbf{p}}\bar{m}) = -(G_w)^{-1} \nabla_y G (D_{\mathbf{p}}\bar{m}), \quad (4.62)$$

and substituting this into (4.61a) gives

$$(D_{xy}G) (D_{\mathbf{p}}\bar{m}) - (G_w)^{-1} (\nabla_x G_w \otimes \nabla_y G) (D_{\mathbf{p}}\bar{m}) = \mathbf{I}, \quad (4.63)$$

using that $(\mathbf{u} \otimes \mathbf{w}) A = \mathbf{u} (\mathbf{w}^T A)$ for vectors \mathbf{u}, \mathbf{w} and matrix A , where we have $\mathbf{u} = \nabla_x G_w$, $\mathbf{w} = \nabla_y H$ and $A = D_{\mathbf{p}}\bar{m}$. Rearranging terms gives

$$(D_{\mathbf{p}}\bar{m})^{-1} = D_{xy}G - (G_w)^{-1} (\nabla_x G_w \otimes \nabla_y G). \quad (4.64)$$

Hence, the right-hand side of the general form of the generated Jacobian equation in (4.43a) becomes

$$\begin{aligned} \Psi(x, u, \mathbf{p}) &= \det((D_{\mathbf{p}}\bar{m})^{-1}(x, u, \mathbf{p})) \frac{f(x)}{g(\bar{m}(x, u, \mathbf{p}))} \\ &= \det(D_{xy}G - (G_w)^{-1} (\nabla_x G_w \otimes \nabla_y G)) \frac{f(x)}{g(\bar{m}(x, u, \mathbf{p}))}. \end{aligned} \quad (4.65)$$

Next, we perform similar steps to rewrite the left-hand side in (4.43b). Substituting $\mathbf{p} = \nabla u$ and differentiating (4.59a) and (4.59b) with respect to \mathbf{x} gives

$$D_{xx}G + D_{xy}G \left(D_x \bar{\mathbf{m}} + \frac{\partial \bar{\mathbf{m}}}{\partial u} \otimes \nabla u + (D_p \bar{\mathbf{m}}) (D^2 u) \right) + \nabla_x G_w \otimes \frac{dw}{dx} = D^2 u, \quad (4.66a)$$

$$\nabla_x G + \left(D_x \bar{\mathbf{m}} + \frac{\partial \bar{\mathbf{m}}}{\partial u} \otimes \nabla u + (D_p \bar{\mathbf{m}}) (D^2 u) \right)^T (\nabla_y G) + G_w \frac{dw}{dx} = \nabla u, \quad (4.66b)$$

where $\frac{dw}{dx}$ is not written out in full (taking into account the dependencies of $\mathbf{y} = \bar{\mathbf{m}}(\mathbf{x}, u(\mathbf{x}), \nabla u(\mathbf{x}))$ on \mathbf{x}) because we will eliminate it from the equations shortly. Rearranging terms in (4.66b), using that $\nabla_x G = \nabla u$ as given in (4.59a) with $\mathbf{p} = \nabla u$, gives

$$\frac{dw}{dx} = -(G_w)^{-1} \left(D_x \bar{\mathbf{m}} + \frac{\partial \bar{\mathbf{m}}}{\partial u} \otimes \nabla u + (D_p \bar{\mathbf{m}}) (D^2 u) \right)^T (\nabla_y G). \quad (4.67)$$

Substituting this into (4.66a) gives

$$D_{xx}G + \left(D_{xy}G - (G_w)^{-1} \nabla_x G_w \otimes \nabla_y G \right) \left(D_x \bar{\mathbf{m}} + \frac{\partial \bar{\mathbf{m}}}{\partial u} \otimes \nabla u + D_p \bar{\mathbf{m}} \otimes D^2 u \right) = D^2 u, \quad (4.68)$$

since $\mathbf{u} \otimes (\mathbf{A}^T \mathbf{w}) = \mathbf{u} \otimes (\mathbf{w}^T \mathbf{A})^T = \mathbf{u} \mathbf{w}^T \mathbf{A} = (\mathbf{u} \otimes \mathbf{w}) \mathbf{A}$ for vectors \mathbf{u}, \mathbf{w} and matrix \mathbf{A} . Here, we have the vectors $\mathbf{u} = \nabla_x G_w$, $\mathbf{w} = \nabla_y G$ and matrix $\mathbf{A} = D_x \bar{\mathbf{m}} + \frac{\partial \bar{\mathbf{m}}}{\partial u} \otimes \nabla u + D_p \bar{\mathbf{m}} \otimes D^2 u$.

Substituting $(D_p \bar{\mathbf{m}})^{-1}$ from (4.64) into (4.68) gives

$$D_{xx}G + (D_p \bar{\mathbf{m}})^{-1} \left(D_x \bar{\mathbf{m}} + \frac{\partial \bar{\mathbf{m}}}{\partial u} \otimes \nabla u \right) = \mathbf{O}, \quad (4.69)$$

using $(D_p \bar{\mathbf{m}})^{-1} (D_p \bar{\mathbf{m}}) = \mathbf{I}$. Hence, we see that $D^2 u$ drops out from this equation and we can see that $\mathbf{A}(\mathbf{x}, u, \nabla u)$ in (4.42b) becomes

$$\mathbf{A}(\mathbf{x}, u, \nabla u) = -(D_p \bar{\mathbf{m}})^{-1} \left(D_x \bar{\mathbf{m}} + \frac{\partial \bar{\mathbf{m}}}{\partial u} \otimes \nabla u \right) = D_{xx}G. \quad (4.70)$$

With $\mathbf{y} = \mathbf{m}(\mathbf{x}) = \bar{\mathbf{m}}(\mathbf{x}, u, \nabla u)$ the generated Jacobian equation in (4.43b) now reads

$$\det(D^2 u - D_{xx}G) = \det(D_{xy}G - (G_w)^{-1} (\nabla_x G_w \otimes \nabla_y G)) \frac{f(\mathbf{x})}{g(\mathbf{m}(\mathbf{x}))}. \quad (4.71)$$

We can repeat the above derivations for the inverse H , starting from the equation

$$H((\bar{\mathbf{m}})^{-1}(\mathbf{y}, w, \mathbf{p}), \mathbf{y}, u) = w, \quad (4.72)$$

cf. (4.55), where we denote $\mathbf{x} = \mathbf{m}^{-1}(\mathbf{y}) = (\bar{\mathbf{m}})^{-1}(\mathbf{y}, w, \mathbf{p})$ as the inverse mapping, assuming it exists, and $\mathbf{p} = \nabla_{\mathbf{y}}w$. Then, we find an alternative form of the generated Jacobian equation as

$$\det(D^2w - D_{\mathbf{y}\mathbf{y}}H) = \det(D_{\mathbf{x}\mathbf{y}}H - (H_w)^{-1}(\nabla_{\mathbf{x}}H \otimes \nabla_{\mathbf{y}}H_w)) \frac{g(\mathbf{y})}{f(\mathbf{m}^{-1}(\mathbf{y}))}, \quad (4.73)$$

where D^2w is the Hessian matrix of w w.r.t. \mathbf{y} .

The general form of a **generated Jacobian equation** associated with generating function $G : \mathcal{G} \subset \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}$ and exponential map $\mathbf{y} = \mathbf{m}(\mathbf{x}) = \exp_{x,u}(\nabla u(\mathbf{x}))$ is given by

$$\det(D^2u - D_{\mathbf{x}\mathbf{x}}G) = \det(D_{\mathbf{x}\mathbf{y}}G - (G_w)^{-1}(\nabla_{\mathbf{x}}G_w \otimes \nabla_{\mathbf{y}}G)) \frac{f(\mathbf{x})}{g(\mathbf{m}(\mathbf{x}))}. \quad (4.74)$$

The shipper's problem or dual Kantorovich problem in optimal transport theory always has the form $u_1(\mathbf{x}) = G(\mathbf{x}, \mathbf{y}, u_2(\mathbf{y})) = -c(\mathbf{x}, \mathbf{y}) + u_2(\mathbf{y})$. Hence, $D_{\mathbf{x}\mathbf{x}}G = -D_{\mathbf{x}\mathbf{x}}c$ and

$$(D_{\mathbf{p}}\bar{\mathbf{m}})^{-1} = D_{\mathbf{x}\mathbf{y}}G - (G_w)^{-1}(\nabla_{\mathbf{x}}G_w \otimes \nabla_{\mathbf{y}}G) = D_{\mathbf{x}\mathbf{y}}G = -D_{\mathbf{x}\mathbf{y}}c. \quad (4.75)$$

The generated Jacobian equation in (4.43b) becomes the **generalized Monge-Ampère equation**

$$\det(D^2u_1(\mathbf{x}) + D_{\mathbf{x}\mathbf{x}}c(\mathbf{x}, \mathbf{m}(\mathbf{x}))) = \det(D_{\mathbf{x}\mathbf{y}}c(\mathbf{x}, \mathbf{m}(\mathbf{x}))) \frac{f(\mathbf{x})}{g(\mathbf{m}(\mathbf{x}))}, \quad (4.76)$$

for $\mathbf{y} = \mathbf{m}(\mathbf{x})$ since $\det(-D_{\mathbf{x}\mathbf{y}}c) = \det(D_{\mathbf{x}\mathbf{y}}c)$ for 2×2 matrices. Hence, we retrieved (4.36).

It is currently not known whether a variational characterization of solutions to generated Jacobian equations exist as in optimal transport theory, i.e., whether a functional $J(u, w)$ exists for a given G and probability measures μ and ν , such that its stationary points are a conjugate pair and there is a mapping sending μ to ν .

4.4.2 A generated-Jacobian mapping

In this section, we show how we can find the mapping $\mathbf{y} = \mathbf{m}(\mathbf{x}) = \overline{\mathbf{m}}(\mathbf{x}, u, \nabla u)$. The relation $u(\mathbf{x}) = G(\mathbf{x}, \mathbf{y}, w(\mathbf{y}))$ has many solutions for $u(\mathbf{x})$ and $w(\mathbf{y})$. However, we do not have a variational characterization of solutions and corresponding dual problem. For a solution u to the second boundary value problem as presented in Definition 4.4.4 we restrict ourselves to G-convex or G-concave solutions in (4.48) or (4.49), respectively.

By the implicit function theorem we can find a mapping as the stationary point of (4.48b) or (4.49b) written as $\mathbf{y} = \mathbf{m}(\mathbf{x}) = \overline{\mathbf{m}}(\mathbf{x}, u(\mathbf{x}), \nabla u(\mathbf{x}))$, i.e.,

$$\nabla_x H(\mathbf{x}, \mathbf{y}, u(\mathbf{x})) + H_w(\mathbf{x}, \mathbf{y}, u(\mathbf{x})) \nabla u(\mathbf{x}) = \mathbf{0}. \quad (4.77)$$

For simplicity, since $u = u(\mathbf{x})$, we define $\tilde{H}(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}, \mathbf{y}, u(\mathbf{x}))$ and rewrite (4.77) to the shorter form

$$\nabla_x \tilde{H}(\mathbf{x}, \mathbf{y}) = \mathbf{0}, \quad (4.78)$$

and use the implicit function theorem to denote the mapping $\mathbf{y} = \mathbf{m}(\mathbf{x})$ as a function of \mathbf{x} only.

A sufficient condition for a minimum in (4.48b) or a maximum in (4.49b) is for the Hessian matrix $-\mathbf{D}_{xx} \tilde{H}(\mathbf{x}, \mathbf{m}(\mathbf{x})) = \mathbf{P}$ to be SND or SPD, respectively. Hence, the requirements for a G-convex or G-concave solution can be written down as follows:

- For a G-convex $u(\mathbf{x})$ and H-concave $w(\mathbf{y})$ we need \mathbf{P} to be SND, i.e., $\text{tr}(\mathbf{P}) \leq 0$ and $\det(\mathbf{P}) \geq 0$.
- For a G-concave $u(\mathbf{x})$ and H-convex $w(\mathbf{y})$ we need \mathbf{P} to be SPD, i.e., $\text{tr}(\mathbf{P}) \geq 0$ and $\det(\mathbf{P}) \geq 0$.

Note that \mathbf{P} is symmetric.

In the above, we assumed that $G_w > 0$ and noted that for $G_w < 0$ we get a max/max pair or a min/min pair. In that case, the requirements on \mathbf{P} change as follows:

- For a G-convex $u(\mathbf{x})$ and H-convex $w(\mathbf{y})$ we need \mathbf{P} to be SPD, i.e., $\text{tr}(\mathbf{P}) \geq 0$ and $\det(\mathbf{P}) \geq 0$.
- For a G-concave $u(\mathbf{x})$ and H-concave $w(\mathbf{y})$ we need \mathbf{P} to be SND, i.e., $\text{tr}(\mathbf{P}) \leq 0$ and $\det(\mathbf{P}) \geq 0$.

Substituting $\mathbf{y} = \mathbf{m}(\mathbf{x})$ and differentiating (4.78) again with respect to \mathbf{x} gives

$$\mathbf{D}_{xx} \tilde{H}(\mathbf{x}, \mathbf{m}(\mathbf{x})) + \mathbf{D}_{xy} \tilde{H}(\mathbf{x}, \mathbf{m}(\mathbf{x})) \mathbf{D}\mathbf{m}(\mathbf{x}) = \mathbf{O}, \quad (4.79)$$

where $D_{xx}\tilde{H}$ is the Hessian matrix of \tilde{H} with respect to \mathbf{x} , $D_{xy}\tilde{H}$ is the matrix of mixed second-order partial derivatives with respect to \mathbf{x} and \mathbf{y} , which we will call the mixed Hessian matrix, and $D\mathbf{m}(\mathbf{x})$ is the 2×2 Jacobi matrix of \mathbf{m} with respect to \mathbf{x} . Using that $-D_{xx}\tilde{H}(\mathbf{x}, \mathbf{m}(\mathbf{x})) = \mathbf{P}$, we find

$$\begin{aligned} \mathbf{P} &= D_{xy}\tilde{H}(\mathbf{x}, \mathbf{m}(\mathbf{x})) D\mathbf{m}(\mathbf{x}) \\ &= \left(D_{xy}H(\mathbf{x}, \mathbf{y}, u(\mathbf{x})) + \nabla u(\mathbf{x}) \otimes \nabla_{\mathbf{y}}H_w(\mathbf{x}, \mathbf{y}, u(\mathbf{x})) \right) D\mathbf{m}(\mathbf{x}). \end{aligned} \quad (4.80)$$

We define the matrix $\mathbf{C} = D_{xy}\tilde{H}(\mathbf{x}, \mathbf{y})$ as the mixed Hessian matrix, and rewrite (4.80) as

$$\mathbf{P}(\mathbf{x}) = \mathbf{C}(\mathbf{x}, \mathbf{m}(\mathbf{x}), u(\mathbf{x})) D\mathbf{m}(\mathbf{x}), \quad (4.81)$$

where

$$\mathbf{C} = \mathbf{C}(\mathbf{x}, \mathbf{m}(\mathbf{x}), u(\mathbf{x})) = D_{xy}\tilde{H} = \begin{pmatrix} \frac{\partial^2 \tilde{H}}{\partial x_1 \partial y_1} & \frac{\partial^2 \tilde{H}}{\partial x_1 \partial y_2} \\ \frac{\partial^2 \tilde{H}}{\partial x_2 \partial y_1} & \frac{\partial^2 \tilde{H}}{\partial x_2 \partial y_2} \end{pmatrix}. \quad (4.82)$$

Assuming the mixed Hessian matrix \mathbf{C} is invertible, a mapping $\mathbf{m}(\mathbf{x})$ is given by the stationary point of (4.78).

Combining (4.81) with the Jacobian equation of energy conservation (4.54) gives the Jacobian equation

$$\det(D\mathbf{m}(\mathbf{x})) = \frac{\det(\mathbf{P}(\mathbf{x}))}{\det(\mathbf{C}(\mathbf{x}, \mathbf{m}(\mathbf{x}), u(\mathbf{x})))} = \frac{f(\mathbf{x})}{g(\mathbf{m}(\mathbf{x}))}. \quad (4.83)$$

Comparing this to the equivalent equation for optimal transport problems (4.35) we see that \mathbf{C} has an additional dependency on u in this formulation.

We can write (4.83) as the generated Jacobian equation in (4.73). Differentiating (4.72) w.r.t. \mathbf{y} twice results in $D^2w = D_{yy}\tilde{H} + D_{xy}\tilde{H} D\mathbf{m}^{-1}$, which gives

$$D^2w - D_{yy}\tilde{H} = -D_{xx}\tilde{H}(D\mathbf{m})^{-2}, \quad (4.84)$$

using (4.79) and $D\mathbf{m}^{-1} = (D\mathbf{m})^{-1}$. Hence, we can rewrite (4.83) with the matrices $\mathbf{P} = -D_{xx}\tilde{H}$ and $\mathbf{C} = D_{xy}\tilde{H}$ as

$$\det(D\mathbf{m})^{-2} \frac{\det(-D_{xx}\tilde{H})}{\det(D_{xy}\tilde{H})} = \frac{\det(D^2w - D_{yy}\tilde{H})}{\det(D_{xy}\tilde{H})} = \frac{g(\mathbf{y})}{f(\mathbf{m}^{-1}(\mathbf{y}))}, \quad (4.85)$$

using $\det(D\mathbf{m}) = f/g$. The cost function matrix $\mathbf{C} = D_{xy}\tilde{H}$ can be written as

$$D_{xy}\tilde{H} = D_{xy}H + \nabla u \otimes \nabla_{\mathbf{y}}H_w = D_{xy}H - (H_w)^{-1} \nabla_{\mathbf{x}}H \otimes \nabla_{\mathbf{y}}H_w, \quad (4.86)$$

using (4.77). Using $D_{yy}\tilde{H} = D_{yy}H$ and combining (4.85) and (4.86) gives

$$\det(D^2w - D_{yy}H) = \det(D_{xy}H - (H_w)^{-1} (\nabla_x H \otimes \nabla_y H_w)) \frac{g(\mathbf{y})}{f(\mathbf{m}^{-1}(\mathbf{y}))}, \quad (4.87)$$

as presented in (4.73).

Another possibility is to find the inverse mapping implicitly from (4.48a) and (4.49a), i.e., to find $\mathbf{x} = \mathbf{m}^{-1}(\mathbf{y}) = (\bar{\mathbf{m}})^{-1}(\mathbf{y}, w, \nabla_y w)$ using

$$\nabla_y G(\mathbf{x}, \mathbf{y}, w(\mathbf{y})) + G_w(\mathbf{x}, \mathbf{y}, w(\mathbf{y})) \nabla w(\mathbf{y}) = \mathbf{0}. \quad (4.88)$$

Completely analogous to the above discussion we can define the matrices $\mathbf{P} = -D_{yy}\tilde{G}$ and $\mathbf{C} = D_{xy}\tilde{G}$, conditions on the definiteness of \mathbf{P} and retrieve the generated Jacobian equation in (4.74). In the above, we chose to go via H because we are interested in the forward mapping (not the inverse) and will compute the surface u from relation (4.77) later in our numerical algorithm.

Instead of using the implicit function theorem in (4.77), we can also find the mapping $\bar{\mathbf{m}}(\mathbf{x}, u, \nabla u)$ and function $w(\mathbf{y})$ by simply solving the system of equations (4.59). We will show this below for the parallel-to-near-field reflector problem.

Example 4.4.3. *The mapping $\mathbf{y} = \mathbf{m}(\mathbf{x})$ for the parallel-to-near-field reflector system in Section 3.3 can be found by solving the system*

$$\begin{aligned} G(\mathbf{x}, \mathbf{y}, w) &= u(\mathbf{x}), \\ \nabla_x G(\mathbf{x}, \mathbf{y}, w) &= \nabla u(\mathbf{x}), \end{aligned}$$

under the condition that $D_p \bar{\mathbf{m}}$ in (4.64) is invertible [145]. Substituting the generating function (3.65) into these equations gives

$$\begin{aligned} \frac{1}{2w} - \frac{w}{2} |\mathbf{x} - \mathbf{y}|^2 &= u(\mathbf{x}), \\ -(\mathbf{x} - \mathbf{y}) w &= \nabla u. \end{aligned}$$

By solving the first equation for w and substituting this expression into the second equation we can find the mapping $\mathbf{y} = \mathbf{m}(\mathbf{x})$ and function $w(\mathbf{y})$ as

$$\mathbf{y} = \mathbf{x} + \frac{2u \nabla u}{1 - |\nabla u|^2}, \quad (4.89a)$$

$$w(\mathbf{y}) = \frac{1 - |\nabla u|^2}{2u}, \quad (4.89b)$$

under the condition that $|\nabla u|^2 \neq 1$.

4.5 The transport boundary condition and edge-ray principle

In this section, we show that the transport boundary condition in (3.48) follows from the implicit boundary condition $\mathbf{m}(\mathcal{X}) = \mathcal{Y}$, stating that all the light from the source \mathcal{X} must be transferred to the target domain \mathcal{Y} . We present a proof for a single freeform reflector with a point source and a far-field target; see Section 3.4. Similar proofs can be constructed for the other systems with differing complexity; see [124, p. 93] for the equivalence proof for a parallel-to-far-field single surface with mapping $\mathbf{m} = \nabla u$.

The equivalence of the boundary conditions follows from the edge-ray principle [128] and convexity of the optical surface.

As in Section 3.1.3, we define our source domain \mathcal{X} as the closed support of $\tilde{f}(x) = f(\phi(x), \theta(x))$, and our target domain \mathcal{Y} as the image under the mapping \mathbf{m} , i.e., $\mathcal{Y} = \mathbf{m}(\mathcal{X})$. We refer to $\mathbf{m} : \mathcal{X} \rightarrow \mathcal{Y}$ as the optical map $\mathbf{y} = \mathbf{m}(x)$ from the source set of stereographic coordinates \mathcal{X} to the target set of stereographic coordinates \mathcal{Y} . We use the implicit boundary condition $\mathbf{m}(\mathcal{X}) = \mathcal{Y}$.

The equivalence of the boundary conditions $\mathbf{m}(\mathcal{X}) = \mathcal{Y}$ and $\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}$ follows from a basic principle of topology, which is known as the edge-ray principle when applied to optics [128]. The closure of a set \mathcal{A} is denoted by a bar and defined as $\overline{\mathcal{A}} = \text{int}(\mathcal{A}) \cup \partial\mathcal{A}$, where $\text{int}(\mathcal{A})$ denotes the union of all open subsets of \mathcal{A} . Note that our source domain \mathcal{X} is a closure, i.e., $\mathcal{X} = \overline{\mathcal{X}}$ is also defined as the closure of the subset of all points where $\tilde{f}(x)$ is nonzero, but we omit the bar notation.

Theorem 2. *Let \mathcal{X} and \mathcal{Y} be topological spaces. Let $\mathbf{m} : \mathcal{X} \rightarrow \mathcal{Y}$ be a homeomorphism, i.e., \mathbf{m} is a bijection that is continuous and open, so \mathbf{m}^{-1} is also continuous. Then, for any subset $\mathcal{A} \subseteq \mathcal{X}$, we have $\mathbf{m}(\partial\mathcal{A}) = \partial\mathbf{m}(\mathcal{A})$. In particular, $\mathbf{m}(\partial\mathcal{X}) = \partial\mathbf{m}(\mathcal{X}) = \partial\mathcal{Y}$.*

Proof. Let $\mathcal{A} \subseteq \mathcal{X}$ and let $\mathbf{y} \in \mathbf{m}(\partial\mathcal{A})$. Then $x = \mathbf{m}^{-1}(\mathbf{y}) \in \partial\mathcal{A}$ and so $x \in \overline{\mathcal{A}} \setminus \text{int}(\mathcal{A})$. Thus, the point x is not an element of the interior, which implies that the point \mathbf{y} is not an element of $\mathbf{m}(\text{int}(\mathcal{A}))$ since \mathbf{m} is injective. We obtain $\mathbf{y} \in \mathbf{m}(\overline{\mathcal{A}}) \setminus \mathbf{m}(\text{int}(\mathcal{A}))$. Furthermore, we have that since \mathbf{m} is a homeomorphism, the image of the closure of \mathcal{A} is equal to the closure of the image of \mathcal{A} , i.e., $\mathbf{m}(\overline{\mathcal{A}}) = \overline{\mathbf{m}(\mathcal{A})}$. We also have that the image of the interior of \mathcal{A} is equal to the interior of the image of \mathcal{A} , i.e., $\mathbf{m}(\text{int}(\mathcal{A})) = \text{int}(\mathbf{m}(\mathcal{A}))$. Therefore, $\mathbf{y} \in \overline{\mathbf{m}(\mathcal{A})} \setminus \text{int}(\mathbf{m}(\mathcal{A})) = \partial\mathbf{m}(\mathcal{A})$, and we have $\mathbf{m}(\partial\mathcal{A}) \subseteq \partial\mathbf{m}(\mathcal{A})$. Reversing our line of thought above we can also show $\partial\mathbf{m}(\mathcal{A}) \subseteq \mathbf{m}(\partial\mathcal{A})$. We conclude that $\mathbf{m}(\partial\mathcal{A}) = \partial\mathbf{m}(\mathcal{A})$. \square

To apply the theorem above it is essential that the optical mapping m is a homeomorphism. In the remainder of this section, we write the mapping m as $m = (\gamma_+ \circ \hat{t} \circ \gamma_-^{-1})(x)$, where we denote γ_{\pm} as the stereographic maps ($-$ for south pole in Equation (3.7), $+$ for north pole in Equation (3.8)) and \hat{t} the vectorial law of reflection in Equation (3.84). We prove that m is a homeomorphism by showing in Lemma 4.5.1 that γ_{\pm} is a bijection and in Lemma 4.5.2 that \hat{t} is a bijection. Subsequently, we can conclude that $m = (\gamma_+ \circ \hat{t} \circ \gamma_-^{-1})(x)$ is a bijection using the fact that a composition of bijective functions is bijective.

The stereographic projections in (3.7) and (3.8) can be rewritten as

$$\gamma_{\pm}(w) = \frac{1}{1 - w \cdot p_{\pm}} w^*, \quad (4.90)$$

where the vector $\hat{p}_{\pm} = (0, 0, \pm 1)$ denotes the pole, and we consider the vector $w \in S^2 \setminus \hat{p}_{\pm} = \{(w_1, w_2, w_3) \in \mathbb{R}^3 \mid |w|^2 = 1\} \setminus \hat{p}_{\pm}$ and the vector $w^* \in \mathbb{R}^2 = \{(w_1, w_2, 0) \in \mathbb{R}^3 \mid (w_1, w_2) \in \mathbb{R}^2\}$.

Lemma 4.5.1. γ_{\pm} is a bijection from the unit sphere $S^2 \setminus \hat{p}_{\pm}$ to \mathbb{R}^2 , i.e., from S^2 without the north/south pole to \mathbb{R}^2 .

Proof. A function is bijective if and only if it has an inverse. It is sufficient to show that $\gamma_{\pm}(w)$ has an inverse. The inverse is given in Equation (3.9) and can be written as

$$\gamma_{\pm}^{-1}(v) = \hat{p}_{\pm} + \frac{2(v - \hat{p}_{\pm})}{|v - \hat{p}_{\pm}|^2}, \quad (4.91)$$

where $v \in \mathbb{R}^2 = \{(v_1, v_2, 0) \in \mathbb{R}^3 \mid (v_1, v_2) \in \mathbb{R}^2\}$ and $\hat{p}_{\pm} = (0, 0, \pm 1)$ is a unit vector perpendicular to the plane of v . Note that $|v - \hat{p}_{\pm}|^2$ is nowhere 0 in \mathbb{R}^3 since the third component of $v - \hat{p}_{\pm}$ is never 0, and that we retrieve a vector $w \neq \hat{p}_{\pm}$ for v finite. In fact, $\lim_{|v| \rightarrow \infty} \gamma_{\pm}^{-1} = \hat{p}_{\pm}$. The length of w indeed becomes

$$\begin{aligned} |\gamma_{\pm}^{-1}(v)| &= \sqrt{\left(\frac{2v_1}{1+v_1^2+v_2^2}\right)^2 + \left(\frac{2v_2}{1+v_1^2+v_2^2}\right)^2 + \left(\frac{\pm(-1+v_1^2+v_2^2)}{1+v_1^2+v_2^2}\right)^2} \\ &= \sqrt{\frac{1+2v_1^2+v_1^4+2v_2^2+2v_1^2v_2^2+v_2^4}{(1+v_1^2+v_2^2)^2}} \\ &= \sqrt{\frac{(1+v_1^2+v_2^2)^2}{(1+v_1^2+v_2^2)^2}} = 1. \end{aligned} \quad (4.92)$$

□

Lemma 4.5.2. *In general, the vectorial law of reflection $\hat{\mathbf{t}} = \hat{\mathbf{s}} - 2(\hat{\mathbf{s}} \cdot \hat{\mathbf{n}})\hat{\mathbf{n}}$ is a bijection for strictly convex optical surfaces. In our case, if the reflector surface $z(x, y) = \sqrt{u(\phi, \theta)^2 - (x^2 + y^2)}$ is convex with (x, y, z) denoting the Cartesian coordinate vector, then the vectorial law of reflection is a bijection.*

Proof. The law of reflection is surjective by definition since we require \mathcal{Y} to be the image of \mathcal{X} under the mapping m . Using bijectivity of the stereographic projections and $\mathcal{Y} = m(\mathcal{X}) = (\gamma_+ \circ \hat{\mathbf{t}} \circ \gamma_-^{-1})(\mathcal{X})$, we require the inverse stereographic projection of \mathcal{Y} , written as $\gamma_+^{-1}(\mathcal{Y})$, to be the image of the inverse stereographic projection of \mathcal{X} , written as $\gamma_-^{-1}(\mathcal{X})$, under the law of reflection $\hat{\mathbf{t}}$.

To show injectivity, we assume $\hat{\mathbf{s}}_1 \neq \hat{\mathbf{s}}_2$ and aim to prove $\hat{\mathbf{t}}(\hat{\mathbf{s}}_1) \neq \hat{\mathbf{t}}(\hat{\mathbf{s}}_2)$. We write

$$\hat{\mathbf{t}}_1 = \hat{\mathbf{s}}_1 - 2(\hat{\mathbf{s}}_1 \cdot \hat{\mathbf{n}}_1)\hat{\mathbf{n}}_1, \quad (4.93a)$$

$$\hat{\mathbf{t}}_2 = \hat{\mathbf{s}}_2 - 2(\hat{\mathbf{s}}_2 \cdot \hat{\mathbf{n}}_2)\hat{\mathbf{n}}_2, \quad (4.93b)$$

with $\hat{\mathbf{t}}_1 = \hat{\mathbf{t}}(\hat{\mathbf{s}}_1)$, $\hat{\mathbf{t}}_2 = \hat{\mathbf{t}}(\hat{\mathbf{s}}_2)$, $\hat{\mathbf{n}}_1 = \hat{\mathbf{n}}(\hat{\mathbf{s}}_1)$ and $\hat{\mathbf{n}}_2 = \hat{\mathbf{n}}(\hat{\mathbf{s}}_2)$.

The vectorial law of reflection states that the incident ray $\hat{\mathbf{s}}$, the reflected ray $\hat{\mathbf{t}}$ and the normal $\hat{\mathbf{n}}$ to the surface all lie in the same plane. Hence, if $\text{span}(\hat{\mathbf{s}}_1, \hat{\mathbf{n}}_1) \neq \text{span}(\hat{\mathbf{s}}_2, \hat{\mathbf{n}}_2)$ then it immediately follows that $\hat{\mathbf{t}}_1 \neq \hat{\mathbf{t}}_2$. Therefore, it remains to show that if $\text{span}(\hat{\mathbf{s}}_1, \hat{\mathbf{n}}_1) = \text{span}(\hat{\mathbf{s}}_2, \hat{\mathbf{n}}_2)$, i.e., $\hat{\mathbf{s}}_1$ and $\hat{\mathbf{s}}_2$ lie in the same plane, it follows that $\hat{\mathbf{t}}_1 \neq \hat{\mathbf{t}}_2$. In the remainder of this proof, we will give a geometric argument to show that this holds for a convex optical surface.

Figure 4.3 shows a two-dimensional representation of the plane of incidence, which equals the plane of reflection. We draw the vectors $\hat{\mathbf{s}}_1$ and $\hat{\mathbf{s}}_2$ as vectors on the unit sphere. The vectorial law of reflection states that the vector $\hat{\mathbf{t}} - \hat{\mathbf{s}} = -2(\hat{\mathbf{s}} \cdot \hat{\mathbf{n}})\hat{\mathbf{n}}$ is a multiple of the unit normal vector $\hat{\mathbf{n}}$. If we assume that $\hat{\mathbf{t}}_1 = \hat{\mathbf{t}}_2$ we see that as $\hat{\mathbf{s}}$ moves counterclockwise from $\hat{\mathbf{s}}_1$ to $\hat{\mathbf{s}}_2$, the normal $\hat{\mathbf{n}}$ to the surface must also move counterclockwise from $\hat{\mathbf{n}}_1$ to $\hat{\mathbf{n}}_2$. Here $\hat{\mathbf{n}}_1$ and $\hat{\mathbf{n}}_2$ are taken to be directed towards the point source O .

In Figure 4.4b we see a parabolic reflector with focal point at O and, hence, multiple parallel outgoing rays with $\hat{\mathbf{t}}_1 = \hat{\mathbf{t}}_2$ while $\hat{\mathbf{s}}_1 \neq \hat{\mathbf{s}}_2$. The optical surface is strictly concave and we can verify that as $\hat{\mathbf{s}}$ moves counterclockwise from $\hat{\mathbf{s}}_1$ to $\hat{\mathbf{s}}_2$, the unit normal $\hat{\mathbf{n}}$ moves in the same direction.

If the optical surface is convex, as for instance shown in Figure 4.4a, this situation cannot occur. As $\hat{\mathbf{s}}$ moves counterclockwise from $\hat{\mathbf{s}}_1$ to $\hat{\mathbf{s}}_2$, the downward unit normal $\hat{\mathbf{n}}$ moves in the opposite direction (or keeps pointing in the same direction in case of a flat surface). Using Figure 4.3 we saw earlier that for $\hat{\mathbf{t}}_1 = \hat{\mathbf{t}}_2$ we needed the normal to move in the same direction. This contradicts with the assumption that $\hat{\mathbf{t}}_1 = \hat{\mathbf{t}}_2$ and hence, $\hat{\mathbf{t}}_1 \neq \hat{\mathbf{t}}_2$. \square

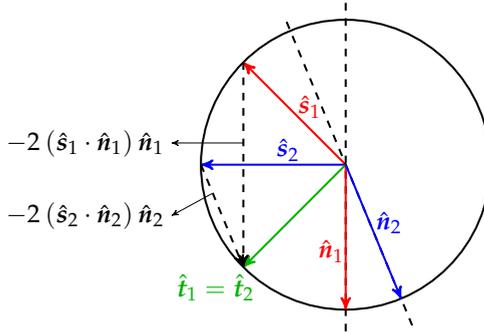


Figure 4.3: As \hat{s} moves counterclockwise from \hat{s}_1 to \hat{s}_2 , the normal \hat{n} to the surface must also move counterclockwise from \hat{n}_1 to \hat{n}_2 if $\hat{t}_1 = \hat{t}_2$.

In summary, we have shown that the stereographic projections γ_{\pm} are bijective functions in Lemma 4.5.1. The law of reflection is bijective for convex optical surfaces as proven in Lemma 4.5.2. Combining Lemma 4.5.1 and 4.5.2, $\mathbf{m} = (\gamma_+ \circ \hat{\mathbf{t}} \circ \gamma_-^{-1})(x)$ is a homeomorphism under the assumption of a convex optical surface by the composition of bijective functions. Subsequently, the equivalence of the boundary conditions $\mathbf{m}(\mathcal{X}) = \mathcal{Y}$ and $\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}$ follows from Theorem 2.

The proof with a parallel source is a lot more straightforward, since injectivity of the law of reflection is easier to prove [4, p. 36]. Injectivity of the law of refraction is yet another story. Moreover, the equivalence proof presented in this section holds for a convex optical surface. In the numerical algorithm presented later in this thesis, we will compute convex and/or concave optical surfaces. However, in order to locally enforce energy conservation and ensure the regularity of \mathbf{m} , we will assume that the optical mapping \mathbf{m} is globally injective and always use the transport boundary condition $\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}$.

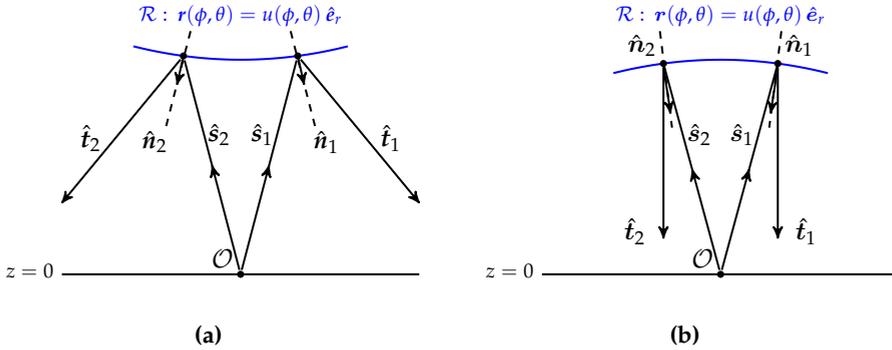


Figure 4.4: (a) A convex parabolic reflector surface gives an injective \hat{t} but (b) a strictly concave parabolic reflector surface may not.

4.6 Summary

In this chapter, we presented the theoretical background on standard Monge-Ampère equations, generalized Monge-Ampère equations and generated Jacobian equations. We consecutively used results from convex analysis, c-convex analysis and G-convex analysis to find the general form of these equations. We started from the Legendre-Fenchel transform, which we subsequently generalized to a c-transform and G-transform.

Using convex pairs, c-convex pairs and G-convex pairs (or concave, c-concave and G-concave pairs) we showed how to find an implicit expression for the mapping $\mathbf{y} = \mathbf{m}(\mathbf{x})$ and presented matrix equations $\mathbf{C} D\mathbf{m} = \mathbf{P}$ which we will use extensively in our numerical algorithm in Chapter 6.

Lastly, we proved for the point-to-far-field reflector (System 6) that the transport boundary condition $\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}$ is equivalent to the implicit boundary condition $\mathbf{m}(\mathcal{X}) = \mathcal{Y}$, stating that all the light from the source domain \mathcal{X} must be transferred to the target domain \mathcal{Y} . For the other optical systems, we also assume we can use the transport boundary condition $\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}$.

Chapter 5

A Literature Review on Numerical Methods

In this chapter, we present an overview of the current literature in freeform illumination optics. There exists a wide range of numerical algorithms to compute freeform optical surfaces. The numerical strategies can be roughly categorized as:

- (1) methods that directly solve the PDE for the optical surface, which could be written as a standard Monge-Ampère equation, generalized Monge-Ampère equation or generated Jacobian equation,
- (2) optimization strategies for the corresponding Monge-Kantorovich mass transportation problem, and
- (3) methods which indirectly compute the surface by using ray-mapping techniques.

In this chapter, we discuss the current research strategies by treating the above categories separately in Section 5.1, 5.2 and 5.3, respectively.

We will also summarize current literature on double freeform surfaces using the above categorization. Optical systems with multiple freeform surfaces can be further categorized according to the shape of the source and target wavefronts, i.e.,

- (A) collimated and/or spherical, i.e., a parallel or point source *and* parallel or point target, and
- (B) at least one wavefront of general shape.

Methods in category (A) are mostly applied to laser beam shaping problems. The input beam is collimated or spherical, and the output beam is also collimated or converging to a point. Laser beam shaping problems involve the construction of two freeform surfaces to control the intensity distribution or phase profile of the beam. For parallel and/or point sources and targets, if the shape of one of the two freeform surfaces is known, the other one directly follows using the Theorem of Malus and Dupin (principle of equal optical path length) [12, p. 130]. Examples in category (A) are System 3, 4, 7, 8, 11, 12, 15 and 16 in Table 3.1 in Chapter 3. An example in category (B) is the double freeform lens that we will discuss later in Chapter 8.

In this chapter, we use the labels (1.A), (1.B), . . . , (3.B) to categorize methods for double freeform systems. For example, methods in category (1.A) are direct PDE solvers for systems with parallel or point sources and targets, while methods in (3.B) use ray-mapping techniques for systems with a source and/or target wavefront of general shape.

We conclude this chapter by presenting relevant literature on generated Jacobian equations in Section 5.4.

5.1 Direct Monge-Ampère solvers

In this section, we give an overview of methods in current literature that directly solve the PDE for the optical surface. First, we discuss methods that directly solve the standard Monge-Ampère equation. Second, we treat solvers for non-standard PDEs, which can be written as generalized Monge-Ampère equations or generated Jacobian equations. Last, we discuss direct methods for the design of double freeform optical systems.

5.1.1 Direct standard Monge-Ampère solvers

Notable authors who directly solve the standard Monge-Ampère equation are:

- **Froese, Oberman, Benamou, Prins** The standard Monge-Ampère equation can be solved using a monotone finite difference scheme to obtain a convex solution [8, 59–61, 111, 112]. The transport boundary condition is treated in [7, 9, 58]. The scheme is derived using the theory of viscosity solutions (of degenerate elliptic equations) and a solution is found by Newton’s method. A convergence proof for the scheme and treatment of the boundary condition is also given. Prins [124, p. 109] uses this approach [9, 61] to develop a wide-stencil algorithm for the interior domain and introduces a signed-distance function for the boundary.

- **Dean, Glowinski, Caboussat, Prins, Beltman** Dean and Glowinski solve the standard Monge-Ampère equation with Dirichlet boundary conditions using a Lagrangian or least-squares approach [35–37]. This method was improved further in [23, 24]. Prins [124, p. 129] was inspired by Caboussat et al. and developed a least-squares algorithm for the standard Monge-Ampère equation [125]. The original least-squares approach only worked for the standard Monge-Ampère equation with quadratic cost function and was extended by my colleagues Beltman et al. [5] to arbitrary orthogonal coordinate systems. Later it was generalized to non-quadratic cost functions and generating functions by ourselves; see Chapter 6.
- **Oliker, Prussner** One of the earliest numerical algorithms for the standard Monge-Ampère equation is described by Oliker and Prussner [115]. They use a finite difference method to solve the equation with Dirichlet boundary conditions on a convex domain. The authors prove the convergence of the algorithm and establish conditions for the existence and uniqueness of a solution.
- **Feng, Neilan** Feng and Neilan [45–48] use a ‘vanishing moment method’ to solve second-order nonlinear PDEs using a new notion of weak solutions, called *moment solutions* as a generalization to viscosity solutions, which is applied to the standard Monge-Ampère equation as one of the examples.
- **Loeper, Rapetti** Loeper and Rapetti use the Newton algorithm to solve the standard Monge-Ampère equation with periodic boundary condition in 2005 [97].
- **Lakkis, Pryer** A Galerkin finite element method for nonlinear elliptic equations is introduced in [84, 89], which can be used for the standard Monge-Ampère equation with Dirichlet boundary conditions [90].

5.1.2 Direct solvers

Numerical methods to solve the generalized Monge-Ampère equation or generated Jacobian equation associated with a particular optical system are developed by:

- **Ries et al.** Ries et al. [127] derive a set of PDEs for a point source using curvatures of wave fronts but do not present details of a numerical solution.

- **Wu et al.** Wu et al. [159, 160] derive a second-order nonlinear PDE for a lens surface with a point source and near-field target, and solve the equation using standard finite differences and Newton iteration. The surface is constructed using B-splines. Recently, Wu et al. developed an iterative approach to include extended light sources [157, 158] and the illumination of ‘hard-to-reach’ areas through a light-guiding system [165] into their framework.
- **Brix et al.** Brix et al. [19, 20] derive a PDE for a point source with a near-field target and use a collocation method with a tensor-product B-spline basis to calculate reflectors and lenses capable of producing a detailed image on a near-field projection screen.

The least-squares method presented in my own papers [131, 133] and in this thesis also falls within the category of direct solvers. The least-squares method for the standard Monge-Ampère equation by Prins et al. [124, 125] was first extended by Yadav et al. to double freeform systems with a parallel source and a parallel target; see the next section. I added to this framework by considering systems involving point sources. The point-to-far-field reflector system in Section 3.4 is treated in [133] and the point-to-far-field lens system in Section 3.5 is considered in [131]. The *generalized least-squares* (GLS) algorithm taking the cost function as input is detailed in Chapter 6, Section 6.1.

Subsequently, the least-squares procedure is further generalized in [130] by taking generating functions as input, which we call the *generated Jacobian least-squares* (GJLS) algorithm in Chapter 6, Section 6.3, of this thesis. Two systems are considered in [130]: (System 14 in Table 3.1) a single freeform lens with a point source and far-field target, and (System 1 in Table 3.1) a single freeform reflector with a parallel source beam and near-field target. System 14 has an associated cost function in optimal transport theory, and we compare the performance of the least-squares algorithm to the previous optimal-transport-based version. System 1 cannot be formulated as an optimal-transport problem, which demonstrates the wider applicability of the new version of the algorithm to any optical system that can be described by a smooth generating function.

5.1.3 Direct solvers for double freeform systems

Using the categories of inverse methods from the introduction to this chapter, we mention a few laser beam shaping methods in category (1.A).

- **Zhang et al.** Zhang et al. [31, 166] derive a second-order nonlinear PDE and solve a nonlinear boundary value problem using Newton’s method.

- **Yadav et al.** Yadav et al. [161, 162, 164] extend the least-squares approach by Prins et al. [124, 125] for the standard Monge-Ampère equation to double freeform systems with a parallel source and a parallel target. A two-reflector system has a quadratic cost function and a two-lens system has a non-quadratic cost function. The freeform surfaces are computed with a least-squares method using the cost function as input. This method is outlined in this thesis in Section 6.1.

Numerical methods in category (1.B) are a little more scarce, which compute double freeform surfaces for a source and/or target wavefront which is non-collimated and not a singular point. To the best of our knowledge, using generating functions we are the first to present a method in category (1.B) in [129]. For the double freeform lens that we will introduce in Chapter 8, we can use the GJLS algorithm twice to successively compute the first and second optical surface. We have an extra degree of freedom in the design by choosing an intermediate target intensity to compute the shape of the first optical surface. In [129] and Chapter 9 of this thesis, an example problem is shown where we progressively translate and scale the intermediate target distribution from the source to the final target distribution, showing the effect of distributing the refractive power over two optical surfaces. By dividing the refractive power we showed that we can compute multiple solutions which differ in design.

5.2 Optimization strategies for the Monge-Kantorovich problem

In this section, we list the current literature on optimization strategies used to solve Monge-Kantorovich mass transportation problems that correspond to single and double freeform optical systems.

5.2.1 Optimization strategies for single freeform systems

Authors who solve Monge-Kantorovich mass transportation problems for optical systems with a single freeform surface are:

- **Benamou, Brenier** solve the optimal transport problem for the quadratic cost function corresponding to the standard Monge-Ampère equation (with periodic boundary conditions) using a fluid mechanics formulation where the source density is transformed continuously into the target density. Extensions to this method can be found in [3, 74]. Another

method by Benamou et al. uses projection methods (such as Bregman and Sinkhorn) which are based on an entropic regularization of the Kantorovich problem [6].

- **Oliker, Caffarelli, Glimm et al.** Oliker [26, 27, 67, 85–87, 115] introduced the *method of supporting paraboloids* and *method of supporting ellipsoids*, which later became known as the *supporting quadric method* (SQM). Oliker considers point light sources, and the freeform optical surfaces are constructed by taking a union or intersection of a set of quadric surfaces with shared foci. The target light distribution is discretized and for each target point a quadric surface can be constructed with the point light source and target point as foci. For a near-field target the quadric surfaces are ellipsoids, and as the target is placed further away to become a far-field target the ellipsoids of revolution converge to paraboloids of revolution whose focus is located at the point source.

The theoretical basis of the SQM method is directly related to the observation that an optical surface u or its related geometrical variable u_1 can be supported from below by tangent planes, the cost function or the generating function, as we saw in the previous chapter. For instance, considering supporting paraboloids for the point-to-far-field reflector and taking the logarithm results in a relation of the form $u_2(\mathbf{y}) - u_1(\mathbf{x}) = c(\mathbf{x}, \mathbf{y})$, where u is supported by paraboloids while u_1 is supported by graphs of the function $G(\cdot, \mathbf{y}, u_2(\mathbf{y})) = -c(\cdot, \mathbf{y}) + u_2(\mathbf{y})$ [67].

Fournier et al. [57] extend Oliker's method of supporting ellipsoids [87] and construct 3D reflectors that produce continuous illuminance distributions using Monte-Carlo ray tracing. Canavesi et al. [28] replace Monte-Carlo ray tracing in this algorithm by a flux estimation method which calculates the intersection points between triplets of ellipsoids. De Castro et al. [29] uses a similar approach for the point-to-far-field reflector problem.

Recently, Oliker et al. [117, 120] used the supporting quadric method on a parallel source light beam which involves a pixelation of the target domain and an iterative adjustment of the parameters of tangent quadrics to the optical surface.

- **Gutierrez, Wang** The generalized Monge-Ampère equation for a point source in [71, 72, 152] is reduced to finding a minimizer or a maximizer of a linear functional subject to a linear constraint, which allows for the use of linear programming algorithms. Moreover, Gutierrez [71] presents a model for refraction taking into account Fresnel losses.

- **Doskolovich et al.** In Doskolovich et al. [22, 41] a point light source and both the far field and near field are considered. In its discrete version, the optimal transport problem is reduced to a linear assignment problem (LAP) for the mapping, based on the construction of an equal-flux grid in the source and target domains. Methods such as the Hungarian algorithm and auction algorithm can be used to solve the LAP. The surface is computed from the mapping using B-splines. Using the LAP-based approach, Doskolovich et al. [41] consider the design of two refractive surfaces for generating irradiance distributions with small angular dimensions as well as off-axis irradiance distributions, designing piecewise-smooth continuous optical surfaces.

5.2.2 Optimization strategies for double freeform systems

Using the categories of inverse methods from the introduction to this chapter, we mention a few laser beam shaping methods in category (2.A).

- **Oliker et al.** Oliker et al. [116, 118] use the Monge-Kantorovich formulation of the problem and compute a unique solution using the supporting quadric method (SQM).
- **Doskolovich et al.** Doskolovich et al. [40] reduce the optimal transport problem in its discrete version to a linear assignment problem (LAP) for the mapping, by constructing an equal-flux grid in the source and target domains. The LAP is solved using methods such as the Hungarian algorithm and auction algorithm. B-splines are used to compute the surface from the mapping.

To the best of our knowledge, there are no methods in category (2.B).

5.3 Ray-mapping methods

In ray-mapping techniques, first the standard Monge-Ampère equation is solved, i.e., $\det(D^2u) = F(x, u, \nabla u)$ with D^2u the Hessian matrix and F a positive function, and an optical map is constructed as the gradient of the solution, $m = \nabla u$. Subsequently, the surface is computed from the mapping using the law of reflection or Snell's law and an integrability condition to ensure continuity of the surface. Ray-mapping techniques work well under the paraxial approximation and thin lens approximation, and allow for extensions to multiple freeform surfaces [51]. Examples can be found in [39, 51, 100].

That using ray-mapping techniques may not be a bad idea is illustrated in [154]. For optical systems with cost functions in optimal transport theory, we can approximate the cost function by taking a Taylor expansion of the cost function truncated after the second-order term. Consequently, this Taylor expansion is quadratic. Using our GLS algorithm, explained in Chapter 6, it was shown that if $|x - y|$ is small for a parallel-to-parallel lens system, the surfaces obtained by solving the standard Monge-Ampère equation and the generalized Monge-Ampère equation did not differ by much. Such an analysis can also be performed for the other optical systems with cost functions discussed in this thesis. However, whether we can make such a comparison for generating functions is an item for future research, which we included in Chapter 10.

5.3.1 Ray-mapping methods for single freeform systems

Examples of authors using ray-mapping techniques to compute single freeform optical surfaces are:

- **Feng et al.** Feng et al. [49] derive a Monge-Ampère equation of a parameterized outgoing wavefront. The ray mapping is obtained with a Newton-Krylov solver. The freeform surface is calculated in a least-squares sense and is iteratively revised by reconstructing the outgoing wavefront. The method is capable of constructing a double freeform lens with two freeform surfaces (the first surface is pre-defined by an analytical formula) and producing complicated images as target irradiance.
- **Bösel et al.** Bösel et al. [15] present a design algorithm for a single freeform lens for general incident wavefronts of zero étendue. The authors first compute the mapping between source and target using a quadratic cost function and subsequently the freeform surfaces using a root-finding algorithm on a system of PDEs.

5.3.2 Ray-mapping methods for double freeform systems

Using the categories of inverse methods from the introduction to this chapter, we mention a few methods in category (3.A) and (3.B).

- **Feng, Froese et al.** (3.A) Feng et al. [50, 52, 53] first compute a ray mapping by using an adaptive mesh method to numerically solve the standard Monge-Ampère equation. A ray mapping $m = \nabla u$ is computed. Subsequently, the two freeform optical surfaces are constructed simultaneously. The first surface is constructed point by point using

the ray mapping and Snell's law, and the second surface follows by equaling the optical path lengths (OPLs) between the input and output wavefronts. The construction of the first freeform surface is analogous to Euler's method for solving ODEs and the simultaneous multiple surfaces (SMS) method in three dimensions to connect the wavefronts [65]. The initial mapping and surfaces are improved in an iterative procedure to find an integrable ray mapping and to improve the initial ray mapping, since it is calculated using the standard Monge-Ampère equation and not the generalized Monge-Ampère equation or generated Jacobian equation. Ray-mapping techniques work well under the paraxial and thin lens approximations.

- **Bösel et al.** (3.A) Bösel et al. [14, 16] extend their ray-mapping approach to the calculation of double freeform surfaces for irradiance and phase control. The model was derived by expressing the coordinates at the target plane in terms of both freeform surfaces and their surface gradients, by formulating three coupled nonlinear PDEs for the first surface and eliminating the second surface from the system by the principle of optical path length.
- **Wei et al.** (3.B) Wei et al. [153] consider source distributions with wavefronts of general shape (i.e., not a parallel or point source). The authors use the least-squares algorithm in [124] to compute an initial mapping and a point-by-point procedure to compute an initial approximation to the first freeform surface. The second freeform surface is constructed according to the principle of equal optical path length. The initial mapping and surfaces are improved in an iterative procedure to find integrable normal vectors to the surfaces, i.e., a surface can be constructed with unit tangent vectors perpendicular to the unit normal vectors derived from the mapping.
- **Bruneton et al.** (3.B) Bruneton et al. [21] consider a point source distribution, and using an initial ray mapping, the two freeform surfaces are constructed with a least-squares optimization algorithm, but the details have been omitted. Similar to our approach, the authors distribute the refraction of the rays over two surfaces. However, Bruneton et al. [21] use an angular redirection of the rays, while minimizing the difference between the light flux that arrives on the target within triangular tubes of rays and the desired target distribution. In our numerical algorithm in Section 8.2, we use an intensity-based approach: we use an intermediate target intensity to compute the first surface, and subsequently the second surface with the final far-field target intensity.

- **Gimenez et al.** (3.B) Gimenez et al. [65] present an extension to the two-dimensional SMS method first described in [105]. The method builds upon the edge-ray principle [127] and, unlike the previously described methods, two input and two output light distributions are required for the construction of the surfaces.

5.4 Generated Jacobian equations

To the best of our knowledge, there is one numerical procedure which attempts to find the solution to general GJEs. **Abedin and Gutiérrez** [1] solve general GJEs using the method of de Leo et al. [91]. De Leo et al. [91] present the point-to-far-field lens problem with a point source and discrete target and use Olikier's supporting quadric method, originally developed in [115]. A freeform reflector or lens for a point source is constructed from a union or intersection of a set of supporting ellipsoids, each having one focus located at the point source and the other one at a discrete target point \mathbf{y}_i in the near field. For a far-field target, the ellipsoids of revolution converge to paraboloids of revolution whose focus is located at the origin. The iterative scheme optimizes the polar radius of each supporting quadric surface and is shown to converge within a finite number of iterations. An iterative optimization algorithm was introduced in [26, 86] and further developed, extended and applied in [28, 29, 57, 104].

Abedin and Gutiérrez extend the algorithm proposed in [91] to consider general GJEs by defining the supporting surfaces as graphs of the generating function $G(\mathbf{x}, \mathbf{y}_i, w_i)$ for each discrete target point \mathbf{y}_i . The algorithm iteratively optimizes w_i for each \mathbf{y}_i by minimizing the discrepancy between the integral of the source intensity over all points \mathbf{x} on the inverse mapping of \mathbf{y}_i and the discrete target intensity at the point \mathbf{y}_i . The resulting solution of the GJE is formed by taking the intersection of the supporting graphs $G(\mathbf{x}, \mathbf{y}_i, w_i)$. While the authors present an application of the algorithm to the parallel-to-near-field reflector problem, they only check that the generating function satisfies the assumptions required to establish convergence of the algorithm. The authors do not yet show numerical results.

	Direct solvers	Optimization	Ray mapping
1. Parallel-to-near-field reflector	[130]		[15]
2. Parallel-to-far-field reflector	[5, 124, 125]		
3. Parallel-to-point reflector	[135]		[16, 52]
4. Parallel-to-parallel reflector	[161, 162]	[116]	[16, 52]
5. Point-to-near-field reflector	[19, 20, 127]	[57, 85, 86]	[15]
6. Point-to-far-field reflector	[127] [133]	[26, 27, 67, 87, 152]	
7. Point-to-point reflector			[16, 52]
8. Point-to-parallel reflector	[135]		[16, 52]
9. Parallel-to-near-field lens		[117, 120]	[15, 39]
10. Parallel-to-far-field lens	[124]		
11. Parallel-to-point lens	[31]		[16, 52]
12. Parallel-to-parallel lens	[161, 164] [166]	[40, 118]	[14, 16, 50, 52, 53]
13. Point-to-near-field lens	[19, 20, 127, 159, 160]	[71]	[15, 49, 100]
14. Point-to-far-field lens	[127] [130, 131]	[22, 41, 71, 72, 91]	[39, 51]
15. Point-to-point lens			[16, 52]
16. Point-to-parallel lens			[16, 52]

Table 5.1: An overview of numerical methods for the 16 base cases discussed in Chapter 5. The numbering of the base cases is the same as in Chapter 3, Table 3.1. The journal articles and PhD theses of our Computational Illumination Optics group at Eindhoven University of Technology are marked in red.

5.5 Summary

In this chapter, we gave an overview of the literature on numerical methods in freeform illumination optics. We classified the methods into three main categories: (1) direct PDE solvers, (2) optimization strategies and (3) ray-mapping techniques. For double freeform systems we added two subcategories: (A) systems with parallel or point sources and parallel or point targets, and (B) systems with at least one wavefront of general shape. In Table 5.1, we present a lookup table citing articles with numerical results for the 16 base-case optical systems. Our own articles are marked in red. Numerical methods in category (1.B), (2.B) and (3.B) are not included since double freeform systems in subcategory (B) are not considered base cases.

We have only briefly introduced our least-squares procedures up until now. In the next chapter, we will first explain the GLS algorithm in detail, which takes an optimal-transport cost function as input. Subsequently, we will describe the GJLS algorithm, which takes a generating function as input. The GJLS algorithm can be used to compute the optical surfaces for all 16 base cases.

So far, my colleagues and I have tested the GLS approach on System 2, 3, 4, 6, 8, 10, 12 and 14, and published our results in the journal papers:

- **Prins, Beltman et al.** System 2 [5, 125].
- **Van Roosmalen et al.** System 3 and 8 [135].
- **Yadav et al.** System 4 [162] and System 12 [164].
- **Romijn et al.** System 6 [133] and System 14 [130, 131].

Results on System 2 and 10 are presented in the PhD thesis of Corien Prins [124] and results on System 4 and 12 in the PhD thesis of Nitin Yadav [161].

We tested the GJLS approach on System 1 and 14 and on a double freeform lens, which is not a base case. We will introduce the double freeform lens in Chapter 8. We published our results in the journal papers:

- **Romijn et al.** System 1 and 14 [130] and a double freeform lens [129].

Chapter 6

The Least-Squares Numerical Algorithm

We can categorize the base-case optical systems presented in Chapter 3. All systems can be formulated by a generated Jacobian equation, while a subset has an associated cost function in optimal transport theory and a generalized Monge-Ampère equation. In Figure 6.1, the base cases are categorized according to type, where $F(x, u, \nabla u)$ denotes the right-hand side of the equations. The innermost circle encapsulates the standard Monge-Ampère equation. The middle circle incorporates systems that have an optimal-transport cost function and generalized Monge-Ampère equation, which includes the standard Monge-Ampère equation for the cost function $c(x, \mathbf{y}) = -x \cdot \mathbf{y}$. The outer circle includes all 16 systems, i.e., all 16 base cases can be described by a generating function.

The generated Jacobian equations can be written in terms of the generating function and solved numerically by using a least-squares algorithm. Originally, this method was developed for the standard Monge-Ampère equation [124], considering the parallel-to-far-field problem. Also the parallel-to-parallel double reflector problem belongs to this category, which has a corresponding quadratic cost function [162]. However, for many optical systems the cost function is more complicated. The numerical procedure was extended to non-quadratic cost functions in [164] (parallel-to-parallel double freeform lens) and [131, 133] (point-to-far-field single reflector/lens). In this chapter, this version of the algorithm will be presented first, and we call it the *generalized least-squares* (GLS) algorithm. We use c -convexity to compute $u_1(x)$ and $\mathbf{y} = \mathbf{m}(x)$ by assuming $u_1(x)$ and $u_2(\mathbf{y})$ in $u_2(\mathbf{y}) - u_1(x) = c(x, \mathbf{y})$ are c -convex or c -concave functions. We will present the full details of the algorithm, and include an extension to polar source coordinates. Subsequently, we further generalize

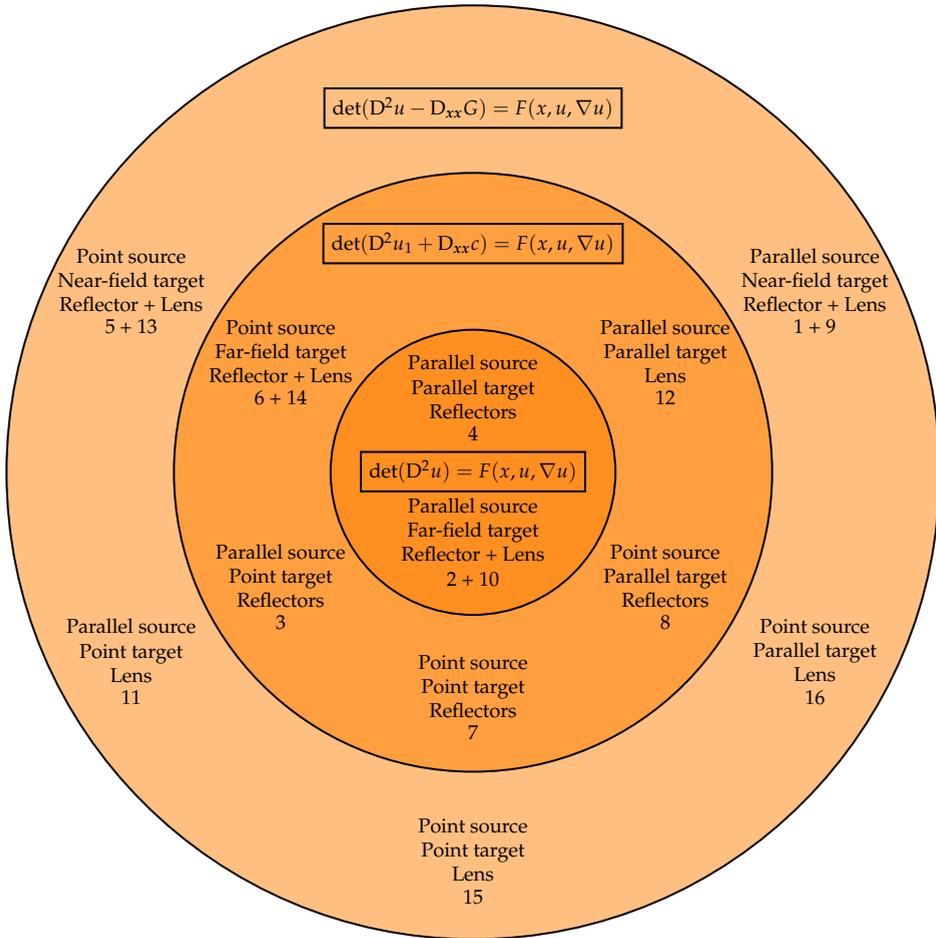


Figure 6.1: Categorization of all base-case optical systems. The numbers refer to the systems presented in the overview in Chapter 3, Table 3.1.

the numerical procedure to a generating-function framework, which allows us to consider a new range of optical systems that cannot be formulated as optimal-transport problems. We use G-convexity theory to compute $u(\mathbf{x})$ and $\mathbf{y} = \mathbf{m}(\mathbf{x})$ assuming $u(\mathbf{x})$ and $w(\mathbf{y})$ in $u(\mathbf{x}) = G(\mathbf{x}, \mathbf{y}, w(\mathbf{y}))$ are *G-convex* or *G-concave functions*. We will call this version the *generated Jacobian least-squares* (GJLS) algorithm and present the required additions to the GLS algorithm.

The numerical procedure works by computing the optical map and optical surface in an iterative procedure which minimizes the global defect in the energy balance.

6.1 The GLS algorithm

The generalized Monge-Ampère equations for optical systems with an optimal-transport cost function are all equations of the form

$$\det(D\mathbf{m}(\mathbf{x})) = \frac{\det(\mathbf{P}(\mathbf{x}))}{\det(\mathbf{C}(\mathbf{x}, \mathbf{m}(\mathbf{x})))} = \frac{f(\mathbf{x})}{g(\mathbf{m}(\mathbf{x}))} = F(\mathbf{x}, \mathbf{m}(\mathbf{x})), \quad (6.1)$$

cf. (4.35), where we take $f(\mathbf{x})$ and $g(\mathbf{m}(\mathbf{x}))$ to incorporate the Jacobians of a coordinate transformation to stereographic coordinates in case of a nonparallel source beam and a nonparallel target beam, respectively, and $F(\mathbf{x}, \mathbf{m}(\mathbf{x})) > 0$ to denote the total right-hand side. As motivated in Section 4.5, we use the transport boundary condition

$$\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}, \quad (6.2)$$

where \mathcal{X} and \mathcal{Y} are the source and target domain, as defined in Section 3.1.3.

We first compute the mapping \mathbf{m} from the generalized Monge-Ampère equation $\det(D\mathbf{m}(\mathbf{x})) = F(\mathbf{x}, \mathbf{m}(\mathbf{x}))$. The mapping \mathbf{m} can be calculated efficiently by an iterative procedure that involves finding the numerical solution of a constrained minimization problem, imposing the transport boundary condition using orthogonal or skew projections on line segments, and computing the numerical solution of a linear boundary value problem. Upon convergence the location of the optical surface u , related to u_1 by a change of variables, is calculated from the mapping using (4.30), i.e.,

$$\nabla u_1(\mathbf{x}) = -\nabla_x c(\mathbf{x}, \mathbf{m}(\mathbf{x})), \quad (6.3)$$

also in a least-squares sense.

To compute a c-convex or c-concave solution, we consider the matrix equation

$$\mathbf{C}(\mathbf{x}, \mathbf{m}(\mathbf{x})) D\mathbf{m}(\mathbf{x}) = \mathbf{P}(\mathbf{x}), \quad (6.4)$$

as introduced in (4.34), with $\mathbf{P}(x)$ an SND or SPD matrix, respectively, satisfying $\det(\mathbf{P}(x)) = F(x, \mathbf{m}(x)) \det(\mathbf{C}(x, \mathbf{m}(x)))$. From Section 4.3.1 we know that we need an SND matrix \mathbf{P} for a c-convex u_1 and an SPD matrix \mathbf{P} for a c-concave u_1 .

We write $\mathbf{m} = \mathbf{m}(x)$ and enforce the matrix equation (6.4) by minimizing the functional

$$J_I[\mathbf{m}, \mathbf{P}] = \frac{1}{2} \iint_{\mathcal{X}} \|\mathbf{C} \mathbf{D} \mathbf{m} - \mathbf{P}\|^2 dx, \quad (6.5)$$

under the constraint $\det(\mathbf{P}) = F \det(\mathbf{C})$. The norm used is the Frobenius norm.

To impose the transport boundary condition (6.2) we minimize the functional

$$J_B[\mathbf{m}, \mathbf{b}] = \frac{1}{2} \oint_{\partial \mathcal{X}} |\mathbf{m} - \mathbf{b}|^2 ds \quad (6.6)$$

over \mathbf{b} , where $|\cdot|$ denotes the L_2 -norm and \mathbf{b} is a function from the source boundary to the target boundary, i.e., $\mathbf{b} : \partial \mathcal{X} \rightarrow \partial \mathcal{Y}$. By minimizing this functional we aim to impose $\mathbf{m}(\partial \mathcal{X}) = \partial \mathcal{Y}$, which holds if $J_B[\mathbf{m}, \mathbf{b}] = 0$. We combine the functionals J_I and J_B by a weighted average as

$$J[\mathbf{m}, \mathbf{P}, \mathbf{b}] = \alpha J_I[\mathbf{m}, \mathbf{P}] + (1 - \alpha) J_B[\mathbf{m}, \mathbf{b}], \quad (6.7)$$

with $0 < \alpha < 1$.

Starting from an initial guess \mathbf{m}^0 and cost function matrix $\mathbf{C}(\cdot, \mathbf{m}^0)$ we perform the iteration:

$$\mathbf{b}^{n+1} = \operatorname{argmin}_{\mathbf{b} \in \mathcal{B}} J_B[\mathbf{m}^n, \mathbf{b}], \quad (6.8a)$$

$$\mathbf{P}^{n+1} = \operatorname{argmin}_{\mathbf{P} \in \mathcal{P}(\mathbf{m}^n)} J_I[\mathbf{m}^n, \mathbf{P}], \quad (6.8b)$$

$$\mathbf{m}^{n+1} = \operatorname{argmin}_{\mathbf{m} \in \mathcal{M}} J[\mathbf{m}, \mathbf{P}^{n+1}, \mathbf{b}^{n+1}], \quad (6.8c)$$

where the minimization steps are performed over the spaces

$$\mathcal{B} = \{\mathbf{b} \in C^1(\partial \mathcal{X})^2 \mid \mathbf{b}(x) \in \partial \mathcal{Y}\}, \quad (6.9a)$$

$$\mathcal{P}(\mathbf{m}) = \{\mathbf{P} \in C^1(\mathcal{X})^{2 \times 2} \mid \mathbf{P} \text{ SND/SPD}, \det(\mathbf{P}) = F(\cdot, \mathbf{m}) \det(\mathbf{C}(\cdot, \mathbf{m}))\}, \quad (6.9b)$$

$$\mathcal{M} = C^2(\mathcal{X})^2, \quad (6.9c)$$

where the smoothness of the spaces is the required smoothness for our numerical algorithm; see Section 6.1.3. After each iteration we update the matrix $\mathbf{C}(\cdot, \mathbf{m}^n)$.

As initial guess \mathbf{m}^0 we map the smallest bounding box enclosing \mathcal{X} to the smallest bounding box enclosing \mathcal{Y} . The bounding box of the source \mathcal{X} has

rectangular shape $[a_{\min}, a_{\max}] \times [b_{\min}, b_{\max}]$ and the bounding box of the target \mathcal{Y} has rectangular shape $[c_{\min}, c_{\max}] \times [d_{\min}, d_{\max}]$. We have two options for an initial guess \mathbf{m}^0 . We can specify the initial guess $\mathbf{m}^0 = (m_1^0, m_2^0)$ as

$$m_1^0 = \frac{x_1 - a_{\min}}{a_{\max} - a_{\min}} c_{\min} + \frac{a_{\max} - x_1}{a_{\max} - a_{\min}} c_{\max}, \quad (6.10a)$$

$$m_2^0 = \frac{x_2 - b_{\min}}{b_{\max} - b_{\min}} d_{\min} + \frac{b_{\max} - x_2}{b_{\max} - b_{\min}} d_{\max}. \quad (6.10b)$$

The corresponding Jacobi matrix $D\mathbf{m}^0$ is diagonal and negative definite. Or, we choose a slightly different initial guess $\mathbf{m}^0 = (m_1^0, m_2^0)$ given by

$$m_1^0 = \frac{x_1 - a_{\min}}{a_{\max} - a_{\min}} c_{\max} + \frac{a_{\max} - x_1}{a_{\max} - a_{\min}} c_{\min}, \quad (6.11a)$$

$$m_2^0 = \frac{x_2 - b_{\min}}{b_{\max} - b_{\min}} d_{\max} + \frac{b_{\max} - x_2}{b_{\max} - b_{\min}} d_{\min}. \quad (6.11b)$$

The corresponding Jacobi matrix $D\mathbf{m}^0$ is diagonal and positive definite.

Below we show by an example how we determine by substitution which initial mapping is appropriate for the required c-convexity or c-concavity.

Example 6.1.1. *We can find the cost function matrix \mathbf{C} in terms of the stereographic coordinates $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ of the point-to-far-field reflector in Section 3.4 as*

$$\begin{aligned} \mathbf{C} = \mathbf{C}(\mathbf{x}, \mathbf{m}(\mathbf{x})) &= D_{xy}c = \begin{pmatrix} \frac{\partial^2 c}{\partial x_1 \partial y_1} & \frac{\partial^2 c}{\partial x_1 \partial y_2} \\ \frac{\partial^2 c}{\partial x_2 \partial y_1} & \frac{\partial^2 c}{\partial x_2 \partial y_2} \end{pmatrix} \\ &= \frac{4}{N(\mathbf{x}, \mathbf{y})^2} (-\mathbf{y} + |\mathbf{y}|^2 \mathbf{x}) (-\mathbf{x} + |\mathbf{x}|^2 \mathbf{y})^T + \frac{2}{N(\mathbf{x}, \mathbf{y})} (\mathbf{I} - 2\mathbf{x} \mathbf{y}^T), \end{aligned}$$

where N is defined in (3.77b). We can compute that

$$\det(D_{xy}c(\mathbf{x}, \mathbf{y})) = \det(\mathbf{C}) = \frac{4}{N(\mathbf{x}, \mathbf{y})^2} > 0.$$

This inequality holds everywhere since we can write $N(\mathbf{x}, \mathbf{y})$ as

$$N(\mathbf{x}(\hat{\mathbf{s}}), \mathbf{y}(\hat{\mathbf{t}})) = \frac{1}{2} (1 + |\mathbf{x}(\hat{\mathbf{s}})|^2) (1 + |\mathbf{y}(\hat{\mathbf{t}})|^2) (1 - \hat{\mathbf{s}} \cdot \hat{\mathbf{t}}),$$

cf. (3.75), (3.76) and (3.96), and we assume that $\hat{\mathbf{s}} \cdot \hat{\mathbf{t}} \neq 1$, which means that the reflector changes the direction of the light rays. We also have

$$\text{tr}(\mathbf{C}) = \frac{4 (-1 + (\mathbf{m}(\mathbf{x}) - \mathbf{J} \mathbf{m}(\mathbf{x})) \cdot \mathbf{x}) (-1 + (\mathbf{m}(\mathbf{x}) + \mathbf{J} \mathbf{m}(\mathbf{x})) \cdot \mathbf{x})}{N(\mathbf{x}, \mathbf{m}(\mathbf{x}))^2},$$

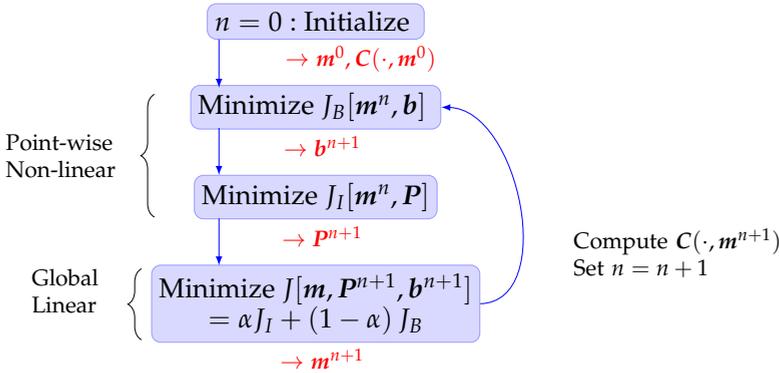


Figure 6.2: Flow chart of the GLS algorithm.

where \mathbf{J} is the symplectic matrix

$$\mathbf{J} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

which acts upon a vector by rotating it over $\pi/2$ in the clockwise direction.

Using the initial mapping in (6.10) or (6.11), we can show that $\det(\mathbf{D}\mathbf{m}^0) > 0$, so that $\det(\mathbf{P}^0) = \det(\mathbf{C}(\cdot, \mathbf{m}^0)) \det(\mathbf{D}\mathbf{m}^0) > 0$. We also know that $\text{tr}(\mathbf{P}^0) = 1/2 \text{tr}(\mathbf{C}(\cdot, \mathbf{m}^0)) \text{tr}(\mathbf{D}\mathbf{m}^0)$, since the diagonal elements of \mathbf{C} are equal and $\mathbf{D}\mathbf{m}^0$ is a diagonal matrix. Substituting \mathbf{m}^0 from (6.10) or (6.11) into the expression for $\text{tr}(\mathbf{C})$ above we find that $\text{tr}(\mathbf{C}(\cdot, \mathbf{m}^0)) \leq 0$ or $\text{tr}(\mathbf{C}(\cdot, \mathbf{m}^0)) \geq 0$, respectively. Hence, using the initial mapping in (6.10) or (6.11), $\mathbf{P}^0 = \mathbf{C}(\cdot, \mathbf{m}^0) \mathbf{D}\mathbf{m}^0$ is negative or positive semi-definite, respectively.

We discretize the source domain \mathcal{X} using a standard rectangular $N_1 \times N_2$ grid for some $N_1, N_2 \in \mathbb{N}$ and introduce $\mathbf{x}_{ij} = (x_{1,i}, x_{2,j})$ with

$$x_{1,i} = a_{\min} + (i - 1) h_1, \quad h_1 = \frac{a_{\max} - a_{\min}}{N_1 - 1}, \quad i = 1, \dots, N_1, \quad (6.12a)$$

$$x_{2,j} = b_{\min} + (j - 1) h_2, \quad h_2 = \frac{b_{\max} - b_{\min}}{N_2 - 1}, \quad j = 1, \dots, N_2. \quad (6.12b)$$

After setting the initial guess \mathbf{m}^0 we perform the minimization steps in (6.8) and subsequently update \mathbf{C} in every iteration. The minimization steps (6.8a), (6.8b), and (6.8c) are explained in detail in Section 6.1.1, 6.1.2, 6.1.3, respectively. Finally, we compute the location of the surface u_1 as described in Section 6.1.4. Figure 6.2 shows a flow chart of the steps in the numerical procedure. The stopping criterion for the iterative procedure is explained in Chapter 7.

6.1.1 Minimization procedure for \mathbf{b}

In this section, we introduce a novel method to impose the transport boundary equation. It is a modification to the method described in [124, 125], which uses orthogonal projections on line segments.

We assume $\mathbf{m} = \mathbf{m}^n$ is given and we need to minimize $J_B[\mathbf{m}, \mathbf{b}]$ over $\mathbf{b} \in \mathcal{B}$. The minimization can be performed point-wise because the integrand does not depend on derivatives of \mathbf{b} . We drop the indices n and $n + 1$ for ease of notation. We denote $\mathbf{m}_{ij} = \mathbf{m}(x_{ij})$, $\mathbf{b}_{ij} = \mathbf{b}(x_{ij})$ and perform the minimization

$$\min_{\mathbf{b}_{ij} \in \mathcal{B}} \frac{1}{2} |\mathbf{m}_{ij} - \mathbf{b}_{ij}|^2. \quad (6.13)$$

We discretize the boundary of \mathcal{Y} using points $\mathbf{z}_k \in \partial\mathcal{Y}$, ($k = 1, 2, \dots, N_b$) with increasing index clockwise along the boundary, and we define $\mathbf{z}_{N_b+1} = \mathbf{z}_1$. We connect adjacent points by line segments $(\mathbf{z}_k, \mathbf{z}_{k+1})$ and determine the “closest” line segment to each \mathbf{m}_{ij} .

First, we define the outward normals \mathbf{n}_k associated with each boundary point \mathbf{z}_k as

$$\mathbf{n}_k = \frac{1}{2} \left(\frac{1}{|\mathbf{z}_{k+1} - \mathbf{z}_k|} \mathbf{J}(\mathbf{z}_{k+1} - \mathbf{z}_k) + \frac{1}{|\mathbf{z}_k - \mathbf{z}_{k-1}|} \mathbf{J}(\mathbf{z}_k - \mathbf{z}_{k-1}) \right), \quad (6.14)$$

where we introduce the symplectic matrix

$$\mathbf{J} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad (6.15)$$

as in Example 6.1.1. Thus, \mathbf{n}_k is the vector pointing in the average direction of the normals to the two adjacent line segments $\mathbf{z}_{k+1} - \mathbf{z}_k$ and $\mathbf{z}_k - \mathbf{z}_{k-1}$. It bisects the angle between the adjacent segments.

We define $l_k : \mathbf{y} = \mathbf{z}_k + \lambda \mathbf{n}_k$, with $\lambda \in \mathbb{R}$ and $k = 1, 2, \dots, N_b$, as the points $\mathbf{y} \in \mathbb{R}^2$ which are on the bisector through \mathbf{z}_k , and let ℓ be the line through \mathbf{m}_{ij} , parallel to the segment $(\mathbf{z}_k, \mathbf{z}_{k+1})$. We let \mathbf{p}_k and \mathbf{p}_{k+1} be the intersection points of ℓ with l_k and l_{k+1} , respectively, as shown in Figure 6.3.

The intersection points \mathbf{p}_k and \mathbf{p}_{k+1} of ℓ with l_k and l_{k+1} , respectively, can be written as

$$\mathbf{p}_k = \mathbf{z}_k + k_1 \mathbf{n}_k, \quad (6.16a)$$

$$\mathbf{p}_{k+1} = \mathbf{z}_{k+1} + k_2 \mathbf{n}_{k+1}, \quad (6.16b)$$

with k_1 and k_2 constants. We can solve for k_1 and k_2 by letting $\mathbf{m}_{ij} - \mathbf{p}_k$ be parallel to $\mathbf{z}_{k+1} - \mathbf{z}_k$ and $\mathbf{m}_{ij} - \mathbf{p}_{k+1}$ be parallel to $\mathbf{z}_{k+1} - \mathbf{z}_k$. Equating the

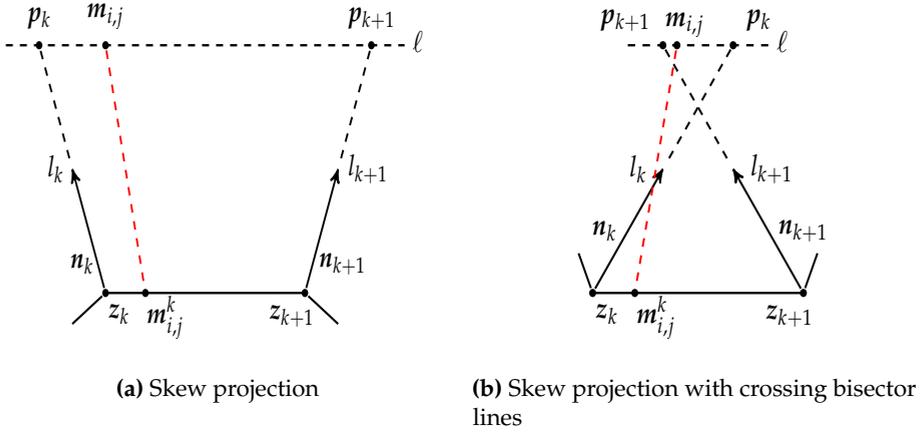


Figure 6.3: Calculation of the distance of m_{ij} to a line segment (z_k, z_{k+1}) with the point m_{ij}^k as the projection of m_{ij} on the line through z_k and z_{k+1} .

slopes of $m_{ij} - p_k = m_{ij} - z_k - k_1 n_k$ and $z_{k+1} - z_k$ gives

$$k_1 = \frac{\det(m_{ij} - z_k, z_{k+1} - z_k)}{\det(n_k, z_{k+1} - z_k)}, \quad (6.16c)$$

where

$$\det(v, w) = \begin{pmatrix} v_1 & w_1 \\ v_2 & w_2 \end{pmatrix}$$

for vectors $v = (v_1, v_2)$ and $w = (w_1, w_2)$. Similarly, equating the slopes of $m_{ij} - p_{k+1} = m_{ij} - z_{k+1} - k_2 n_{k+1}$ and $z_{k+1} - z_k$ gives

$$k_2 = \frac{\det(m_{ij} - z_{k+1}, z_{k+1} - z_k)}{\det(n_{k+1}, z_{k+1} - z_k)}. \quad (6.16d)$$

The bisector lines l_k and l_{k+1} may cross as illustrated in Figure 6.3b. We determine whether this occurs by evaluating if both of the following two conditions hold for each line segment

$$d_k(p_{k+1}) d_k(z_{k+1}) < 0, \quad (6.17a)$$

$$d_{k+1}(p_k) d_{k+1}(z_k) < 0, \quad (6.17b)$$

where

$$d_k(\mathbf{y}) = \det(\mathbf{p}_k - \mathbf{z}_k, \mathbf{y} - \mathbf{z}_k), \quad (6.17c)$$

$$d_{k+1}(\mathbf{y}) = \det(\mathbf{p}_{k+1} - \mathbf{z}_{k+1}, \mathbf{y} - \mathbf{z}_{k+1}). \quad (6.17d)$$

Note that the line l_k through \mathbf{z}_k and \mathbf{p}_k consists of the points \mathbf{y} for which $d_k(\mathbf{y}) = 0$. Likewise, the line l_{k+1} through \mathbf{z}_{k+1} and \mathbf{p}_{k+1} consists of the points \mathbf{y} for which $d_{k+1}(\mathbf{y}) = 0$. The first condition (6.17a) checks whether \mathbf{p}_{k+1} and \mathbf{z}_{k+1} are located on opposite sides of the line segment $\mathbf{p}_k - \mathbf{z}_k$. Likewise, the second condition (6.17b) checks whether \mathbf{p}_k and \mathbf{z}_k are located on opposite sides of $\mathbf{p}_{k+1} - \mathbf{z}_{k+1}$. Together they determine whether the segment $\mathbf{p}_k - \mathbf{z}_k$ crosses the segment $\mathbf{p}_{k+1} - \mathbf{z}_{k+1}$.

The projection \mathbf{m}_{ij}^k of \mathbf{m}_{ij} on the line segment $(\mathbf{z}_k, \mathbf{z}_{k+1})$ is given by

$$\mathbf{m}_{ij}^k = \mathbf{z}_k + t_k (\mathbf{z}_{k+1} - \mathbf{z}_k), \quad (6.18a)$$

$$t_k = \frac{|\mathbf{m}_{ij} - \tilde{\mathbf{p}}|}{|\mathbf{p}_{k+1} - \mathbf{p}_k|}, \quad (6.18b)$$

where, if at least one of the inequalities (6.17a) and (6.17b) does not hold, i.e., the bisector lines do not cross as in Figure 6.3a, we define $\tilde{\mathbf{p}} = \mathbf{p}_k$. On the other hand, if (6.17a) and (6.17b) are both true, i.e., the bisector lines do cross as in Figure 6.3b, we use $\tilde{\mathbf{p}} = \mathbf{p}_{k+1}$.

For a given line segment $(\mathbf{z}_k, \mathbf{z}_{k+1})$, we check if \mathbf{m}_{ij} is located in between \mathbf{p}_k and \mathbf{p}_{k+1} by evaluating

$$0 \leq (\mathbf{m}_{ij} - \mathbf{p}_k) \cdot (\mathbf{p}_{k+1} - \mathbf{p}_k) \leq |\mathbf{p}_{k+1} - \mathbf{p}_k|^2. \quad (6.19)$$

If this does not hold, we set the parameter t_k in (6.18b) to an arbitrarily large number to exclude \mathbf{m}_{ij}^k in the minimization procedure below. The skew projection (6.18) ensures that the ratio between the distances $|\mathbf{m}_{ij} - \tilde{\mathbf{p}}|$ and $|\mathbf{p}_{k+1} - \mathbf{p}_k|$ is the same as the ratio between the distances $|\mathbf{m}_{ij}^k - \mathbf{z}_k|$ and $|\mathbf{z}_{k+1} - \mathbf{z}_k|$.

Finally, we select the points \mathbf{b}_{ij} as

$$\mathbf{b}_{ij} = \operatorname{argmin}_{\mathbf{m}_{ij}^k} \frac{1}{2} |\mathbf{m}_{ij} - \mathbf{m}_{ij}^k|^2. \quad (6.20)$$

This procedure is repeated for all $\mathbf{x}_{ij} \in \partial\mathcal{X}$.

6.1.2 Minimization procedure for P

We assume $\mathbf{m} = \mathbf{m}''$ is fixed. The minimization of $J_I[\mathbf{m}, P]$ can be performed point-wise because the integrand does not depend on derivatives of P . The

minimization can be solved by an analytic procedure, which was first introduced by [124, p. 133]. We minimize $\|\mathbf{Q} - \mathbf{P}\|$ for each grid point $x_{ij} \in \mathcal{X}$, where $\mathbf{Q} = \mathbf{CD} = (q_{ij})$, and $\mathbf{D} = (d_{ij})$ is the central difference approximation of $D\mathbf{m}$. We define

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{pmatrix}. \quad (6.21)$$

Note that \mathbf{P} is symmetric while \mathbf{Q} may not be symmetric. This gives rise to the minimization problem

$$\text{minimize} \quad H_S(p_{11}, p_{22}, p_{12}) = \frac{1}{2} \|\mathbf{Q}_S - \mathbf{P}\|^2, \quad (6.22a)$$

$$\text{subject to} \quad \det(\mathbf{P}) = p_{11} p_{22} - p_{12}^2 = F(\cdot, \mathbf{m}) \det(\mathbf{C}(\cdot, \mathbf{m})), \quad (6.22b)$$

where $\mathbf{Q}_S = \frac{1}{2}(\mathbf{Q} + \mathbf{Q}^T)$. We have replaced \mathbf{Q} by its symmetric part \mathbf{Q}_S , since this gives the same minimizer (p_{11}, p_{22}, p_{12}) . For a c-convex solution we impose the additional constraint

$$\text{tr}(\mathbf{P}) = p_{11} + p_{22} \leq 0, \quad (6.22c)$$

while for a c-concave solution we require

$$\text{tr}(\mathbf{P}) = p_{11} + p_{22} \geq 0. \quad (6.22d)$$

Possible solutions are stationary points of the Lagrangian Λ defined as

$$\Lambda(p_{11}, p_{22}, p_{12}; \mu) = \frac{1}{2} \|\mathbf{Q}_S - \mathbf{P}\|^2 + \mu (\det(\mathbf{P}) - F(\cdot, \mathbf{m}) \det(\mathbf{C}(\cdot, \mathbf{m}))). \quad (6.23)$$

Setting all partial derivatives of Λ to 0 results in the algebraic system

$$p_{11} + \lambda p_{22} = q_{11}, \quad (6.24a)$$

$$\lambda p_{11} + p_{22} = q_{22}, \quad (6.24b)$$

$$(1 - \lambda) p_{12} = \frac{1}{2} (q_{12} + q_{21}) =: \tilde{q}_{12}, \quad (6.24c)$$

$$p_{11} p_{22} - p_{12}^2 = F(\cdot, \mathbf{m}) \det(\mathbf{C}(\cdot, \mathbf{m})), \quad (6.24d)$$

where $\lambda = \mu / \det(\mathbf{C}(\cdot, \mathbf{m}))$. We can always select the possible minimizers that satisfy the nonlinear constraint (6.22b), and the inequality constraint (6.22c) or (6.22d). The system (6.24a) – (6.24c) is linear in p_{11} , p_{12} and p_{22} and invertible if $\lambda \neq \pm 1$. Assuming $\lambda \neq \pm 1$ the calculation of the stationary points is described in Section 6.1.2.1. The case $\lambda = 1$ can only occur if $q_{11} = q_{22}$ and $\tilde{q}_{12} = 0$.

We calculate the possible minimizers differently in this case in Section 6.1.2.4. Similarly, $\lambda = -1$ can only occur if $q_{11} = -q_{22}$ with possible minimizers derived in Section 6.1.2.5. All possible minimizers are displayed in Table 6.1. In [5], my colleague René Beltman has given a geometric interpretation of the minimizers of the Lagrangian Λ , by showing that the minimizers correspond to the points where an ellipsoid is tangent to a hyperboloid.

Table 6.1: All stationary points of (6.24).

Cases	λ	p_{11}	p_{12}	p_{22}
Case A: Regular, $a_4 > 0$	(6.37a)	$\frac{\lambda q_{22} - q_{11}}{\lambda^2 - 1}$	$\frac{\tilde{q}_{12}}{1 - \lambda}$	$\frac{\lambda q_{11} - q_{22}}{\lambda^2 - 1}$
	(6.37b)	"	"	"
	(6.37c)	"	"	"
	(6.37d)	"	"	"
Case B: Regular, $a_4 = 0$	(6.47a)	"	"	"
	(6.47b)	"	"	"
Case C: Regular, $a_4 = 0, a_2 = 0$ and $\det(\mathbf{Q}_5) = 0$	0	q_{11}	\tilde{q}_{12}	q_{22}
Case D: $q_{11} = q_{22}$ and $\tilde{q}_{12} = 0$	1	$\frac{q_{11}}{2}$	$\sqrt{\frac{q_{11}^2}{4} - F \det(\mathbf{C})}$	$\frac{q_{11}}{2}$
	1	$\frac{q_{11}}{2}$	$-\sqrt{\frac{q_{11}^2}{4} - F \det(\mathbf{C})}$	$\frac{q_{11}}{2}$
	1	$\sqrt{F \det(\mathbf{C})}$	0	$\sqrt{F \det(\mathbf{C})}$
	1	$-\sqrt{F \det(\mathbf{C})}$	0	$-\sqrt{F \det(\mathbf{C})}$
Case E: $q_{11} = -q_{22}$	-1	(6.68a)	$\frac{\tilde{q}_{12}}{2}$	(6.68c)
	-1	(6.69a)	$\frac{\tilde{q}_{12}}{2}$	(6.69c)

6.1.2.1 Case A: Regular minimizers

If $\lambda \neq \pm 1$ we can rewrite the system (6.24a) – (6.24c) to

$$p_{11} = \frac{\lambda q_{22} - q_{11}}{\lambda^2 - 1}, \quad (6.25a)$$

$$p_{12} = \frac{\tilde{q}_{12}}{1 - \lambda}, \quad (6.25b)$$

$$p_{22} = \frac{\lambda q_{11} - q_{22}}{\lambda^2 - 1}. \quad (6.25c)$$

We use a small tolerance parameter of 10^{-8} to determine whether $\lambda - 1 \ll 1$ or $\lambda + 1 \ll 1$ and move to the case $\lambda = 1$ in Section 6.1.2.4 if the first condition holds or to $\lambda = -1$ in Section 6.1.2.5 if the latter holds. Substituting (6.25) into the constraint (6.24d) we obtain the quartic equation

$$p(\lambda) = a_4 \lambda^4 + a_2 \lambda^2 + a_1 \lambda + a_0 = 0, \quad (6.26)$$

with the coefficients

$$a_4 = F(\cdot, \mathbf{m}) \det(\mathbf{C}(\cdot, \mathbf{m})) \geq 0, \quad (6.27a)$$

$$a_2 = -2 F(\cdot, \mathbf{m}) \det(\mathbf{C}(\cdot, \mathbf{m})) - \det(\mathbf{Q}_S) = -2 a_4 - \det(\mathbf{Q}_S), \quad (6.27b)$$

$$a_1 = \|\mathbf{Q}_S\|^2 \geq 0, \quad (6.27c)$$

$$a_0 = F(\cdot, \mathbf{m}) \det(\mathbf{C}(\cdot, \mathbf{m})) - \det(\mathbf{Q}_S) = a_4 - \det(\mathbf{Q}_S). \quad (6.27d)$$

For an SND/SPD matrix \mathbf{P} we require $\det(\mathbf{P}) = F(\cdot, \mathbf{m}) \det(\mathbf{C}(\cdot, \mathbf{m})) = a_4 \geq 0$. We have $F(\cdot, \mathbf{m}) > 0$ in (6.1), so we require $\det(\mathbf{C}(\cdot, \mathbf{m})) \geq 0$ for an SND/SPD matrix \mathbf{P} . We note that this condition should be checked for each system individually, e.g., for a point-to-far-field reflector we showed that this condition is satisfied in Example 6.1.1. Hence, from now on we assume $\det(\mathbf{C}) \geq 0$. By (6.24a) and (6.24b), the condition (6.22c) becomes

$$\text{tr}(\mathbf{P}) = \frac{\text{tr}(\mathbf{Q}_S)}{1 + \lambda} \leq 0, \quad (6.28)$$

and the condition (6.22d) becomes

$$\text{tr}(\mathbf{P}) = \frac{\text{tr}(\mathbf{Q}_S)}{1 + \lambda} \geq 0. \quad (6.29)$$

We will show that the quartic equation (6.26) has at least two real roots, and at least one root satisfies $\lambda < -1$ and another one $\lambda > -1$. Thus, the convexity or concavity condition (6.28) or (6.29) can be satisfied by choosing the appropriate

value of λ for a given $\text{tr}(\mathbf{Q}_S)$. Afterwards, we can compute the minimizers from (6.25).

We solve (6.26) using Ferrari's method [144, p. 32] and rewrite the quartic equation as two quadratic equations. We will first assume $a_4 > 0$ and discuss the case $a_4 = 0$ later. Note that we use a small tolerance parameter of 10^{-8} to determine if a_4 is close to zero and move to the case $a_4 = 0$ if $|a_4| < 10^{-8}$. We rewrite (6.26) to

$$\left(\lambda^2 + \frac{a_2}{2a_4}\right)^2 = -\frac{a_1}{a_4}\lambda - \frac{a_0}{a_4} + \left(\frac{a_2}{2a_4}\right)^2. \quad (6.30)$$

Adding an arbitrary value y to the left-hand side under the square, results in extra terms on the right-hand side as

$$\left(\lambda^2 + \frac{a_2}{2a_4} + y\right)^2 = 2y\lambda^2 - \frac{a_1}{a_4}\lambda - \frac{a_0}{a_4} + \left(\frac{a_2}{2a_4}\right)^2 + \frac{a_2}{a_4}y + y^2, \quad (6.31)$$

which simplifies to

$$\left(\lambda^2 + \frac{a_2}{2a_4} + y\right)^2 = \left(\sqrt{2y}\lambda - \frac{a_1}{2a_4\sqrt{2y}}\right)^2, \quad (6.32)$$

if y is a solution to the cubic equation

$$y^3 + b_2y^2 + b_1y + b_0 = 0, \quad (6.33)$$

with coefficients

$$b_2 = \frac{a_2}{a_4} = -2 - \frac{\det(\mathbf{Q}_S)}{F \det(\mathbf{C})}, \quad (6.34a)$$

$$b_1 = \frac{1}{4} \left(\frac{a_2}{a_4}\right)^2 - \frac{a_0}{a_4} = \frac{\det(\mathbf{Q}_S) (8F \det(\mathbf{C}) + \det(\mathbf{Q}_S))}{4F^2 \det(\mathbf{C})^2}, \quad (6.34b)$$

$$b_0 = -\frac{1}{8} \left(\frac{a_1}{a_4}\right)^2 = -\frac{1}{8} \frac{\|\mathbf{Q}_S\|^4}{F^2 \det(\mathbf{C})^2}. \quad (6.34c)$$

One solution for y is given by [123, p. 179]

$$Q = \frac{b_2^2 - 3b_1}{9}, \quad R = \frac{2b_2^3 - 9b_1b_2 + 27b_0}{54}, \quad (6.35a)$$

$$A = -\text{sgn}(R) \left(|R| + \sqrt{R^2 - Q^3}\right)^{1/3}, \quad (6.35b)$$

$$y = A + \frac{Q}{A} - \frac{b_2}{3}. \quad (6.35c)$$

If $Q = R = 0$ then $A = 0$ and y is undefined. In that case, the cubic equation has a triple root given by $y = -\frac{b_2}{3}$. Using y in (6.35c) or $y = -\frac{b_2}{3}$, we find that (6.32) becomes

$$\lambda^2 + \frac{a_2}{2 a_4} + y = \pm \left(\sqrt{2 y} \lambda - \frac{a_1}{2 a_4 \sqrt{2 y}} \right). \quad (6.36)$$

These are two quadratic equations for λ , which results in the roots

$$\lambda_1 = -\sqrt{\frac{y}{2}} + \sqrt{-\frac{y}{2} - \frac{a_2}{2 a_4} + \frac{a_1}{2 a_4 \sqrt{2 y}}}, \quad (6.37a)$$

$$\lambda_2 = -\sqrt{\frac{y}{2}} - \sqrt{-\frac{y}{2} - \frac{a_2}{2 a_4} + \frac{a_1}{2 a_4 \sqrt{2 y}}}, \quad (6.37b)$$

$$\lambda_3 = \sqrt{\frac{y}{2}} + \sqrt{-\frac{y}{2} - \frac{a_2}{2 a_4} - \frac{a_1}{2 a_4 \sqrt{2 y}}}, \quad (6.37c)$$

$$\lambda_4 = \sqrt{\frac{y}{2}} - \sqrt{-\frac{y}{2} - \frac{a_2}{2 a_4} - \frac{a_1}{2 a_4 \sqrt{2 y}}}. \quad (6.37d)$$

Note that we could have division by 0 in (6.37) if $y = 0$. By substituting $y = 0$ into the cubic equation (6.33) we see that this happens when $a_1 = 0$, which is the case if $q_{11} = q_{22} = \tilde{q}_{12} = 0$. Again, we use a small tolerance parameter of 10^{-8} to determine if y is close to zero and the algorithm throws an error if $|y| < 10^{-8}$. Another important question is whether we may only find complex roots. The number of real roots to the quartic equation in (6.26) may be zero, two or four. A necessary condition for the quartic equation to have no real roots [63] is for the discriminant

$$D = 256 a_4^3 a_0^3 - 128 a_4^2 a_2^2 a_0^2 + 144 a_4^2 a_2 a_1 a_0 - 27 a_4^2 a_1^4 + 16 a_4 a_2^4 a_0 - 4 a_4 a_2^3 a_1^2 \geq 0, \quad (6.38)$$

and

$$\left(\frac{a_2}{a_4} \geq 0 \right) \quad \text{or} \quad \left(\left(\frac{a_2}{a_4} \right)^2 - 4 a_0 \geq 0 \right). \quad (6.39)$$

Substituting (6.27) into (6.39) and defining

$$V = \frac{\det(\mathbf{Q}_S)}{F(\cdot, \mathbf{m}) \det(\mathbf{C}(\cdot, \mathbf{m}))}, \quad r = \frac{\|\mathbf{Q}_S\|^2}{F(\cdot, \mathbf{m}) \det(\mathbf{C}(\cdot, \mathbf{m}))} \geq 0, \quad (6.40)$$

results in the two inequalities

$$(V \leq -2) \quad \text{or} \quad (-8 \leq V \leq 0), \quad (6.41)$$

which gives

$$V \leq 0. \quad (6.42)$$

The discriminant in (6.38) can be rewritten using (6.40) as

$$D = (2V - r)(2V + r)(27r^2 + 256 - 192V - 60V^2 - 4V^3). \quad (6.43)$$

The quartic equation has no real roots for $V \leq 0$. If this is the case, then

$$256 - 192V - 60V^2 - 4V^3 \geq 0. \quad (6.44)$$

Moreover, $(2V - r)(2V + r) \geq 0$ only if $2V \leq -r$. Substituting (6.40) into $2V \leq -r$ gives

$$0 \leq -(q_{11} + q_{22})^2, \quad (6.45)$$

which only holds if $q_{11} = -q_{22}$, which is the special case treated separately in Section 6.1.2.5. Hence, if we assume $q_{11} \neq -q_{22}$ the quartic equation has at least two real roots, so that we can always find a minimizer.

Assuming $\lambda \neq \pm 1$ is real, we have a stationary point of the Lagrangian Λ in (6.22). We calculate p_{11} , p_{12} , and p_{22} from (6.25) for all real solutions λ . This results in at most four vectors (p_{11}, p_{22}, p_{12}) . We select the vectors satisfying the c-convexity condition (6.28) or c-concavity condition (6.29). For these vectors, we calculate and compare the values of H_5 in (6.22a) to find the global minimum.

Note that we can always find a λ for which either the c-convexity condition (6.28) or the c-concavity condition (6.29) is satisfied, since we can show that one root will satisfy $\lambda < -1$ and another $\lambda > -1$. For the matrix \mathbf{Q}_5 it holds that

$$\text{tr}(\mathbf{Q}_5)^2 - 4 \det(\mathbf{Q}_5) = (q_{11} - q_{22})^2 + 4 \tilde{q}_{12}^2 \geq 0. \quad (6.46)$$

Using this inequality and substituting $\lambda = -1$ into (6.26) gives $p(-1) = -\text{tr}(\mathbf{Q}_5)^2 < 0$. Similarly, substituting $\lambda = 1$ into (6.26) gives the relation $p(1) = \text{tr}(\mathbf{Q}_5)^2 - 4 \det(\mathbf{Q}_5) \geq 0$. Since $a_4 > 0$ we have $\lim_{\lambda \rightarrow \pm\infty} p(\lambda) = \infty$.

Hence, by the intermediate value theorem $p(\lambda)$ has at least two real roots, a root $-1 < \lambda \leq 1$ and another root $\lambda < -1$, to match the familiar shape of a quartic function. Hence, there is a root $\lambda < -1$ and a root $\lambda > -1$.

6.1.2.2 Case B: Regular minimizers with $a_4 = 0$

We will now discuss the case $a_4 = 0$. We arrive at this case if $\det(\mathbf{C}(\cdot, \mathbf{m})) = 0$ or if $|a_4| < 10^{-8}$, which can occur if $\det(\mathbf{C}(\cdot, \mathbf{m})) \ll 1$ or $f \ll 1$ at a point in

the domain. In this case, the quartic equation becomes a quadratic equation with roots

$$\lambda_1 = \frac{-a_1 + \sqrt{a_1^2 - 4a_2 a_0}}{2 a_2}, \quad (6.47a)$$

$$\lambda_2 = \frac{-a_1 - \sqrt{a_1^2 - 4a_2 a_0}}{2 a_2}, \quad (6.47b)$$

if $a_2 \neq 0$. Below we will discuss the case $a_2 = 0$ and if $|a_2| < 10^{-8}$ using a small tolerance of 10^{-8} we will also move to this case. The discriminant (6.38) with $a_4 = 0$ and substituting (6.27) becomes

$$a_1^2 - 4 a_2 a_0 = (q_{11}^2 - q_{22}^2)^2 + 4 \tilde{q}_{12}^2 (q_{11} + q_{22})^2 \geq 0. \quad (6.48)$$

Hence, solutions to the quadratic equation are always real. Furthermore, also in this case substituting $\lambda = -1$ into (6.26) gives $p(-1) = -\text{tr}(\mathbf{Q}_S)^2 < 0$ and substituting $\lambda = 1$ into (6.26) gives $p(1) = \text{tr}(\mathbf{Q}_S)^2 - 4 \det(\mathbf{Q}_S) \geq 0$, and consequently (6.26) has at least one solution $\lambda > -1$. As a result, one of the conditions (6.28) or (6.29) is satisfied. However, whether there is another $\lambda < -1$ is unclear since $a_2 < 0$ and (6.26) is a concave downward parabola. Thus, suppose you would like to impose c-convexity or c-concavity of the solution u_1 and $\lambda > -1$ is the only solution; if this solution matches either one of (6.28) or (6.29) but not the one you wanted, we let the algorithm throw an error.

6.1.2.3 Case C: Regular minimizers with $a_4 = 0$ and $a_2 = 0$

If $a_4 = 0$ and $a_2 = 0$, then we deduce from (6.27) that $a_0 = 0$ as well, which gives $\lambda = 0$ as the only solution to (6.26) with corresponding values of $\mathbf{P} = \mathbf{Q}_S$ using (6.24). Again, this value of λ can only satisfy one of the constraints (6.28) or (6.29), and we should check whether it matches with the constraint that we imposed.

In the next two sections, we will treat the cases $\lambda = \pm 1$. Although the system (6.24a) – (6.24c) can only be inverted for $\lambda \neq \pm 1$, what happens if we substitute these values in the quartic equation? If we substitute $\lambda = 1$ in (6.26) using (6.27), we obtain

$$4 \tilde{q}_{12}^2 + (q_{11} - q_{22})^2 = 0, \quad (6.49)$$

i.e., $\lambda = 1$ can only occur if $\tilde{q}_{12} = 0$ and $q_{11} = q_{22}$, which is exactly the case treated in Section 6.1.2.4. Similarly, substituting $\lambda = -1$ in (6.26) using (6.27) gives

$$(q_{11} + q_{22})^2 = 0, \quad (6.50)$$

i.e., $q_{11} = -q_{22}$ as in Section 6.1.2.5.

6.1.2.4 Case D: Minimizers if $q_{11} = q_{22}$ and $\tilde{q}_{12} = 0$

If $q_{11} = q_{22}$ and $\tilde{q}_{12} = 0$, then we cannot use the calculations in the previous section. We substitute $\tilde{q}_{12} = \frac{1}{2}(q_{12} + q_{21})$ in (6.22a) and find

$$\begin{aligned} H_S(p_{11}, p_{22}, p_{12}) - \frac{1}{4}(q_{12} - q_{21})^2 \\ = \frac{1}{2}((p_{11} - q_{11})^2 + 2(p_{12} - \tilde{q}_{12})^2 + (p_{22} - q_{22})^2). \end{aligned} \quad (6.51)$$

Since no elements (p_{11}, p_{22}, p_{12}) of \mathbf{P} occur in the subtracted term, we may as well minimize the right-hand side of (6.51), i.e.,

$$\text{minimize} \quad \hat{H}_S(p_{11}, p_{22}, p_{12}) = \frac{1}{2}((p_{11} - q_{11})^2 + 2p_{12}^2 + (p_{22} - q_{11})^2), \quad (6.52a)$$

$$\text{subject to} \quad \det(\mathbf{P}) = p_{11} p_{22} - p_{12}^2 = F(\cdot, \mathbf{m}) \det(\mathbf{C}(\cdot, \mathbf{m})), \quad (6.52b)$$

using $\tilde{q}_{12} = 0$ and $q_{11} = q_{22}$. We can simplify the minimization further by rewriting the constraint to $p_{12}^2 = p_{11} p_{22} - F \det(\mathbf{C})$ and restricting the minimization over the domain where $p_{12} = \pm \sqrt{p_{11} p_{22} - F \det(\mathbf{C})}$ is real, i.e.,

$$\operatorname{argmin}_{(p_{11}, p_{22}) \in \mathbb{R}^2} \frac{1}{2}((p_{11} - q_{11})^2 + 2(p_{11} p_{22} - F \det(\mathbf{C})) + (p_{22} - q_{11})^2). \quad (6.53)$$

The minimizer could lie in the interior of this domain or on the boundary. The minimizer in the interior can be found by setting the derivatives with respect to p_{11} and p_{22} to 0. This results in the minimizing line

$$p_{11} + p_{22} = q_{11}. \quad (6.54)$$

Real values of p_{12} only occur on part of this line. Substituting this equation into $p_{11} p_{22} - F \det(\mathbf{C}) \geq 0$ we obtain

$$p_{11}^2 - q_{11} p_{11} + F \det(\mathbf{C}) \leq 0. \quad (6.55)$$

This equation has real solutions p_{11} if the discriminant $q_{11}^2 - 4F \det(\mathbf{C}) \geq 0$. Consequently, the part of the minimizing line which gives real values of p_{11} is given by the line segment

$$\frac{q_{11} - \sqrt{q_{11}^2 - 4F \det(\mathbf{C})}}{2} \leq p_{11} \leq \frac{q_{11} + \sqrt{q_{11}^2 - 4F \det(\mathbf{C})}}{2}. \quad (6.56)$$

Values of p_{11} on the whole segment are minimizers. We pick a unique p_{11} as the midpoint of the segment, such that

$$p_{11} = \frac{q_{11}}{2}, \quad (6.57a)$$

$$p_{12} = \pm \sqrt{\frac{q_{11}^2}{4} - F \det(\mathbf{C})}, \quad (6.57b)$$

$$p_{22} = \frac{q_{11}}{2}, \quad (6.57c)$$

where p_{22} follows from (6.54) and subsequently p_{12} is computed using the constraint (6.52b). The minimizers may also lie on the boundary of the domain where p_{12} is real. On the boundary we have $p_{12} = 0$, such that the constraint (6.52b) becomes

$$p_{11} p_{22} = F \det(\mathbf{C}). \quad (6.58)$$

Substituting $p_{12} = 0$ and $p_{22} = F \det(\mathbf{C}) / p_{11}$ in (6.53) gives

$$\operatorname{argmin}_{p_{11} \in \mathbb{R}} \frac{1}{2} \left((p_{11} - q_{11})^2 + \left(\frac{F \det(\mathbf{C})}{p_{11}} - q_{11} \right)^2 \right). \quad (6.59)$$

Differentiating with respect to p_{11} and subsequent multiplication with p_{11}^3 gives

$$p_{11}^4 - q_{11} p_{11}^3 + F \det(\mathbf{C}) q_{11} p_{11} - F^2 \det(\mathbf{C})^2 = 0. \quad (6.60)$$

This quartic equation can be factored as

$$(p_{11}^2 - F \det(\mathbf{C})) (p_{11}^2 - q_{11} p_{11} + F \det(\mathbf{C})) = 0, \quad (6.61)$$

which has four solutions given by

$$p_{11} = \pm \sqrt{F \det(\mathbf{C})}, \quad (6.62a)$$

$$p_{11} = \frac{q_{11} \pm \sqrt{q_{11}^2 - 4 F \det(\mathbf{C})}}{2}, \quad (6.62b)$$

where the first two solutions are always real and the other two solutions are either the endpoints of the line segments in (6.57) or complex-valued. Hence, they may be ignored since they lead to the same minimizing value of H_S as in (6.57). The first two solutions result in

$$p_{11} = \pm \sqrt{F \det(\mathbf{C})}, \quad (6.63)$$

$$p_{12} = 0, \quad (6.64)$$

$$p_{22} = \pm \sqrt{F \det(\mathbf{C})}. \quad (6.65)$$

From the above equations we can see that on the boundary we can satisfy either $\text{tr}(\mathbf{P}) \leq 0$ or $\text{tr}(\mathbf{P}) \geq 0$. Hence, we can always find a point on the line segment which satisfies either the c-convexity condition (6.22c) or c-concavity condition (6.22d).

6.1.2.5 Case E: Minimizers if $q_{11} = -q_{22}$

If $q_{11} = -q_{22}$ we have seen that $\lambda = -1$. Again, we cannot use the method described in Section 6.1.2.1. From (6.24a) and (6.24b) we know that

$$p_{22} = p_{11} - q_{11}, \quad (6.66a)$$

$$p_{12} = \frac{\tilde{q}_{12}}{2}. \quad (6.66b)$$

Substitution into (6.24d) gives

$$p_{11}^2 - q_{11} p_{11} - \frac{\tilde{q}_{12}^2}{4} - F \det(\mathbf{C}) = 0. \quad (6.67)$$

Solving for p_{11} gives the two solutions

$$p_{11} = \frac{q_{11}}{2} + \frac{\sqrt{q_{11}^2 + 4F \det(\mathbf{C}) + \tilde{q}_{12}^2}}{2}, \quad (6.68a)$$

$$p_{12} = \frac{\tilde{q}_{12}}{2}, \quad (6.68b)$$

$$p_{22} = -\frac{q_{11}}{2} + \frac{\sqrt{q_{11}^2 + 4F \det(\mathbf{C}) + \tilde{q}_{12}^2}}{2}, \quad (6.68c)$$

and

$$p_{11} = \frac{q_{11}}{2} - \frac{\sqrt{q_{11}^2 + 4F \det(\mathbf{C}) + \tilde{q}_{12}^2}}{2}, \quad (6.69a)$$

$$p_{12} = \frac{\tilde{q}_{12}}{2}, \quad (6.69b)$$

$$p_{22} = -\frac{q_{11}}{2} - \frac{\sqrt{q_{11}^2 + 4F \det(\mathbf{C}) + \tilde{q}_{12}^2}}{2}, \quad (6.69c)$$

which are always real. The first solution satisfies $\text{tr}(\mathbf{P}) \geq 0$, while the second solution may or may not satisfy $\text{tr}(\mathbf{P}) \leq 0$. The algorithm throws an error if we impose the c-convexity condition (6.22c) on all points for which $\lambda = -1$ and we cannot find a solution such that $\text{tr}(\mathbf{P}) \leq 0$

A summary of all possible stationary points is given in Table 6.1.

6.1.3 Minimization procedure for m

We minimize the combined functional $J[\mathbf{m}, \mathbf{P}, \mathbf{b}]$ over all $\mathbf{m} \in \mathcal{M}$. This step cannot be performed point-wise and we compute the first variation $\delta J[\mathbf{m}, \mathbf{P}, \mathbf{b}](\boldsymbol{\eta})$ with respect to \mathbf{m} for $\boldsymbol{\eta} \in \mathcal{M}$, i.e.,

$$\begin{aligned} \delta J[\mathbf{m}, \mathbf{P}, \mathbf{b}](\boldsymbol{\eta}) &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \left(J[\mathbf{m} + \epsilon \boldsymbol{\eta}, \mathbf{P}, \mathbf{b}] - J[\mathbf{m}, \mathbf{P}, \mathbf{b}] \right) \\ &= \lim_{\epsilon \rightarrow 0} \left[\frac{\alpha}{2} \int_{\mathcal{X}} 2 (\mathbf{C D m} - \mathbf{P}) : \mathbf{C D \eta} + \epsilon \|\mathbf{C D \eta}\|^2 dx \right. \\ &\quad \left. + \frac{1-\alpha}{2} \oint_{\partial \mathcal{X}} 2 (\mathbf{m} - \mathbf{b}) \cdot \boldsymbol{\eta} + \epsilon |\boldsymbol{\eta}|^2 ds \right] \\ &= \alpha \int_{\mathcal{X}} (\mathbf{C D m} - \mathbf{P}) : \mathbf{C D \eta} dx + (1-\alpha) \oint_{\partial \mathcal{X}} (\mathbf{m} - \mathbf{b}) \cdot \boldsymbol{\eta} ds. \end{aligned} \quad (6.70)$$

The minimizer is given by $\delta J[\mathbf{m}, \mathbf{P}, \mathbf{b}](\boldsymbol{\eta}) = 0$ for all $\boldsymbol{\eta} \in \mathcal{M}$. We rewrite the Frobenius inner product in (6.70) to

$$(\mathbf{C D m} - \mathbf{P}) : \mathbf{C D \eta} = \mathbf{C}^T (\mathbf{C D m} - \mathbf{P}) : \mathbf{D \eta} = \mathbf{V} : \mathbf{D \eta}, \quad (6.71)$$

where we introduced the matrix $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2] = \mathbf{C}^T (\mathbf{C D m} - \mathbf{P})$, with column vectors \mathbf{v}_1 and \mathbf{v}_2 . Defining $\mathbf{W} = \mathbf{V}^T = [\mathbf{w}_1, \mathbf{w}_2]$ we can derive

$$\begin{aligned} (\mathbf{C D m} - \mathbf{P}) : \mathbf{C D \eta} &= \mathbf{W} : (\mathbf{D \eta})^T = \mathbf{w}_1 \cdot \nabla \eta_1 + \mathbf{w}_2 \cdot \nabla \eta_2 \\ &= \nabla \cdot (\eta_1 \mathbf{w}_1 + \eta_2 \mathbf{w}_2) - (\eta_1 \nabla \cdot \mathbf{w}_1 + \eta_2 \nabla \cdot \mathbf{w}_2) \\ &= \nabla \cdot (\mathbf{W} \boldsymbol{\eta}) - \boldsymbol{\eta} \cdot \begin{pmatrix} \nabla \cdot \mathbf{w}_1 \\ \nabla \cdot \mathbf{w}_2 \end{pmatrix}. \end{aligned} \quad (6.72)$$

The divergence of \mathbf{V} is defined as

$$\nabla \cdot \mathbf{V} = \begin{pmatrix} \frac{\partial v_{11}}{\partial x_1} + \frac{\partial v_{12}}{\partial x_2} \\ \frac{\partial v_{21}}{\partial x_1} + \frac{\partial v_{22}}{\partial x_2} \end{pmatrix} = \begin{pmatrix} \nabla \cdot \mathbf{w}_1 \\ \nabla \cdot \mathbf{w}_2 \end{pmatrix}. \quad (6.73)$$

Hence, we can rewrite (6.72) to

$$(\mathbf{C D m} - \mathbf{P}) : \mathbf{C D \eta} = \nabla \cdot (\mathbf{V}^T \boldsymbol{\eta}) - \boldsymbol{\eta} \cdot (\nabla \cdot \mathbf{V}). \quad (6.74)$$

Using the latter equation and Gauss's theorem the first integral in (6.70) becomes

$$\begin{aligned} \int_{\mathcal{X}} (\mathbf{C D m} - \mathbf{P}) : \mathbf{C D \eta} dx &= \int_{\mathcal{X}} \nabla \cdot (\mathbf{V}^T \boldsymbol{\eta}) dx - \int_{\mathcal{X}} \boldsymbol{\eta} \cdot (\nabla \cdot \mathbf{V}) dx \\ &= \oint_{\partial \mathcal{X}} \mathbf{V}^T \hat{\mathbf{n}} \cdot \boldsymbol{\eta} ds - \int_{\mathcal{X}} \boldsymbol{\eta} \cdot (\nabla \cdot \mathbf{V}) dx, \end{aligned} \quad (6.75)$$

where $\hat{\mathbf{n}}$ is the outward unit normal on the boundary of the source domain $\partial\mathcal{X}$.

The minimizer is given by $\delta J[\mathbf{m}, \mathbf{P}, \mathbf{b}](\boldsymbol{\eta}) = 0$ for all $\boldsymbol{\eta} \in \mathcal{M}$. Setting (6.70) to 0 and using (6.75) gives

$$\alpha \int_{\mathcal{X}} \nabla \cdot (\mathbf{V}^T \boldsymbol{\eta}) \, dx - \alpha \int_{\mathcal{X}} \boldsymbol{\eta} \cdot (\nabla \cdot \mathbf{V}) \, dx + (1 - \alpha) \oint_{\partial\mathcal{X}} (\mathbf{m} - \mathbf{b}) \, ds = 0,$$

which is equal to

$$\oint_{\partial\mathcal{X}} \left(\alpha \mathbf{V} \hat{\mathbf{n}} + (1 - \alpha) (\mathbf{m} - \mathbf{b}) \right) \cdot \boldsymbol{\eta} \, ds - \alpha \int_{\mathcal{X}} \boldsymbol{\eta} \cdot (\nabla \cdot \mathbf{V}) \, dx = 0. \quad (6.76)$$

Setting $\eta_1 = 0$ or $\eta_2 = 0$ and applying the fundamental lemma of calculus of variations [32] twice, we obtain $\nabla \cdot \mathbf{V} = 0$ for $\mathbf{x} \in \mathcal{X}$, i.e.,

$$\nabla \cdot (\mathbf{C}^T \mathbf{C} \mathbf{D} \mathbf{m}) = \nabla \cdot (\mathbf{C}^T \mathbf{P}). \quad (6.77)$$

For $\mathbf{x} \in \partial\mathcal{X}$ we have

$$\begin{aligned} \alpha \mathbf{V} \hat{\mathbf{n}} + (1 - \alpha) (\mathbf{m} - \mathbf{b}) &= 0 && \iff \\ \alpha (\mathbf{C}^T (\mathbf{C} \mathbf{D} \mathbf{m} - \mathbf{P})) \hat{\mathbf{n}} + (1 - \alpha) (\mathbf{m} - \mathbf{b}) &= 0 && \iff \\ (1 - \alpha) \mathbf{m} + \alpha (\mathbf{C}^T \mathbf{C} \mathbf{D} \mathbf{m}) \hat{\mathbf{n}} &= (1 - \alpha) \mathbf{b} + \alpha \mathbf{C}^T \mathbf{P} \hat{\mathbf{n}}. \end{aligned} \quad (6.78)$$

Combining (6.77) and (6.78) gives the coupled boundary value problem

$$\nabla \cdot (\mathbf{C}^T \mathbf{C} \mathbf{D} \mathbf{m}) = \nabla \cdot (\mathbf{C}^T \mathbf{P}), \quad \mathbf{x} \in \mathcal{X}, \quad (6.79a)$$

$$(1 - \alpha) \mathbf{m} + \alpha (\mathbf{C}^T \mathbf{C} \mathbf{D} \mathbf{m}) \hat{\mathbf{n}} = (1 - \alpha) \mathbf{b} + \alpha \mathbf{C}^T \mathbf{P} \hat{\mathbf{n}}, \quad \mathbf{x} \in \partial\mathcal{X}. \quad (6.79b)$$

We solve this system for \mathbf{m} using the finite volume method, which is explained in Appendix B.1. We first solve a linear system to compute the first component m_1 of \mathbf{m} . In these equations we substitute the value of m_2 from the previous iteration. Subsequently, using the new m_1 we solve a linear system for the second component m_2 . In this way, we compute \mathbf{m}^{n+1} in one step, i.e., we do not perform multiple iterations where we substitute the newly found m_2 into the equation for m_1 . We experimented with multiple iterations until this resubstitution barely changes m_1 and m_2 (with a tolerance of 10^{-8}), but we found that performing one iteration is sufficient for most example problems.

The required smoothness for the minimization spaces in (6.8) can be derived from (6.79). In (6.79a), we need \mathbf{m} to be at least twice differentiable and \mathbf{P} to be at least once differentiable. From Equation (6.79b), we require \mathbf{b} to be at least once differentiable since first-order partial derivatives of \mathbf{m} appear in the Robin boundary condition and \mathbf{m} is at least twice differentiable. Rigorous regularity results of solutions to the basic PDE of optimal transport in (6.3) and (6.1) are beyond the scope of this thesis and we refer to [148, Ch. 12] for a detailed overview.

6.1.4 Computation of u_1

Upon convergence of the iterative procedure for the mapping \mathbf{m} , we calculate the location of the optical surface from Equation (6.3). In the ideal situation, $\nabla_x c(\mathbf{x}, \mathbf{m}(\mathbf{x}))$ is a conservative vector field, so that there exists a u_1 that satisfies (6.3). Then we also have that $C D\mathbf{m} = \mathbf{P}$ with \mathbf{P} symmetric, which is most likely not satisfied after running the iterative procedure. In the numerical algorithm $\nabla_x c(\mathbf{x}, \mathbf{m}(\mathbf{x}))$ is not conservative due to numerical errors. Direct integration methods do not give a unique solution u_1 . The solution will depend on the order of integration. Hence, we compute the generalized least-squares solution by minimizing the functional

$$I[u_1] = \frac{1}{2} \int_{\mathcal{X}} |\nabla u_1 + \nabla_x c(\cdot, \mathbf{m})|^2 dx. \quad (6.80)$$

We cannot perform this step point-wise and analogously to the minimization procedure for \mathbf{m} , we compute the first variation of $\delta I[u_1](v)$ for $v \in C^2(\mathcal{X})$ as

$$\begin{aligned} \delta I[u_1](v) &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \left(I[u_1 + \epsilon v] - I[u_1] \right) \\ &= \lim_{\epsilon \rightarrow 0} \frac{1}{2} \int_{\mathcal{X}} 2 (\nabla u_1 + \nabla_x c(\cdot, \mathbf{m})) \cdot \nabla v + \epsilon |v|^2 dx \\ &= \int_{\mathcal{X}} (\nabla u_1 + \nabla_x c(\cdot, \mathbf{m})) \cdot \nabla v dx. \end{aligned} \quad (6.81)$$

The minimizer is given by $\delta I[u_1](v) = 0$ for all $v \in C^2(\mathcal{X})$. Once more, using Gauss's theorem and the fundamental lemma of calculus of variations [32], we obtain the boundary value problem

$$\Delta u_1 = -\nabla \cdot \nabla_x c(\cdot, \mathbf{m}), \quad \mathbf{x} \in \mathcal{X}, \quad (6.82a)$$

$$\nabla u_1 \cdot \hat{\mathbf{n}} = -\nabla_x c(\cdot, \mathbf{m}) \cdot \hat{\mathbf{n}}, \quad \mathbf{x} \in \partial \mathcal{X}. \quad (6.82b)$$

The Neumann problem (6.82) for u_1 has multiple solutions and a corresponding discretization matrix with incomplete rank. Given the mapping \mathbf{m}^{n+1} during iteration n , the average height of the surface u_1 can be fixed to calculate a unique solution. In particular, for a parallel-to-far-field reflector or lens we can show that the solution is unique up to an additive constant, and unique up to a multiplicative constant for a point-to-far-field reflector or lens. This can be verified by substituting the respective cost functions (3.49) and (3.96) into the boundary value problem (6.82).

However, we calculate a unique solution by prescribing the average value of u_1 as a constraint which adds an extra row and column to the discretization

matrix. We let $A \mathbf{u}_1 = \mathbf{b}$ be the linear system corresponding to the finite volume approximation, where A is the discrete Laplacian and $\mathbf{u}_1 = (u_{1,ij})$ is the discretized surface u_1 , which is reshaped from an $N \times N$ array to an $N^2 \times 1$ vector. Since A is symmetric and negative semi-definite, any solution of $A \mathbf{u}_1 = \mathbf{b}$ also maximizes $\frac{1}{2} \mathbf{u}_1^T A \mathbf{u}_1 - \mathbf{u}_1^T \mathbf{b}$. If we prescribe the average value of $u_1 = l$, with $l > 0$ a constant, as a constraint we obtain the constrained maximization problem

$$\max_{\mathbf{u}_1} \left\{ \frac{1}{2} \mathbf{u}_1^T A \mathbf{u}_1 - \mathbf{u}_1^T \mathbf{b} \quad \middle| \quad \mathbf{e}^T \mathbf{u}_1 = N^2 l \right\}, \quad (6.83)$$

where $\mathbf{e} = (1, 1, \dots, 1) \in N(\mathbf{A})$ is a vector of ones of size $N^2 \times 1$ in the kernel of A . We introduce the Lagrangian Λ as

$$\Lambda(\mathbf{u}_1; \lambda) = \frac{1}{2} \mathbf{u}_1^T A \mathbf{u}_1 - \mathbf{u}_1^T \mathbf{b} + \lambda (\mathbf{e}^T \mathbf{u}_1 - N^2 l). \quad (6.84)$$

Setting all partial derivatives to zero gives the numerical scheme

$$\begin{pmatrix} A & \mathbf{e} \\ \mathbf{e}^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ N^2 l \end{pmatrix}. \quad (6.85)$$

Using this system, we compute \mathbf{u}_1 in one step.

The compatibility condition of the Neumann problem is

$$\int_{\mathcal{X}} \nabla \cdot \nabla_x c(\mathbf{x}, \mathbf{m}(\mathbf{x})) \, d\mathbf{x} = \oint_{\partial \mathcal{X}} \nabla_x c(\mathbf{x}, \mathbf{m}(\mathbf{x})) \cdot \hat{\mathbf{n}} \, ds, \quad (6.86)$$

which is automatically satisfied by Gauss's theorem.

We compute the location of the optical surface u as a function of \mathbf{x}_{ij} using the relation of u to u_1 for the optical system in particular. An example is given below.

Example 6.1.2. For instance, for a point-to-far-field reflector, we have the relation $u_1(\mathbf{x}) = -\log(u(\mathbf{x})/(1 + |\mathbf{x}|^2))$, cf. (3.76). Hence, $u(\mathbf{x}) = e^{-u_1(\mathbf{x})}(1 + |\mathbf{x}|^2)$. Since we consider a point source the coordinates $\mathbf{x} \in \mathcal{X}$ are stereographic coordinates. We transform the stereographic source coordinates defined in (3.7) to Cartesian coordinates $(x, y, z) = u(\hat{\mathbf{s}}) \hat{\mathbf{s}}$ and plot the surface using

$$x = \frac{2 u(\mathbf{x}) x_1}{1 + x_1^2 + x_2^2}, \quad y = \frac{2 u(\mathbf{x}) x_2}{1 + x_1^2 + x_2^2}, \quad z = \frac{u(\mathbf{x}) (1 - x_1^2 - x_2^2)}{1 + x_1^2 + x_2^2},$$

cf. (3.9), for all $\mathbf{x}_{ij} = (x_{1,i}, x_{2,j}) \in \mathcal{X}$.

6.2 Extension to polar source coordinates

In practical problems, the incoming beam emitted from a point source is often cone-shaped and has a corresponding circular domain in \mathcal{X} . In a similar way, we can imagine a parallel source beam emitting a cylindrical beam. Hence, we perform a change of coordinates for the source variables $x \in \mathcal{X}$ on the source domain to polar stereographic coordinates $\omega = (\rho, \zeta) \in \tilde{\mathcal{X}}$ with the transformation

$$x_1 = \rho \cos(\zeta), \quad x_2 = \rho \sin(\zeta), \quad (6.87)$$

where $\rho \geq 0$ is the radial coordinate and $0 \leq \zeta < 2\pi$ the azimuth (angle with respect to positive x_1 -axis). We maintain Cartesian or Cartesian stereographic coordinates for the target domain.

We let $\tilde{u}_1(\omega) = u_1(x)$ and denote the optical map by $\tilde{\mathbf{y}} = \tilde{\mathbf{m}}(\omega)$ from source domain $\tilde{\mathcal{X}}$ to target domain $\tilde{\mathcal{Y}}$ such that $\tilde{\mathcal{Y}} = \tilde{\mathbf{m}}(\tilde{\mathcal{X}})$. For ease of notation, we drop all tildes and in the following we use the notation $u_1(\omega)$ for the surface and $\mathbf{y} = \mathbf{m}(\omega)$ for the mapping $\mathbf{m} : \mathcal{X} \rightarrow \mathcal{Y}$ from the polar source domain to the target domain, etc.

The mapping is implicitly given by

$$\nabla u_1(\omega) = -\nabla_{\omega} c(\omega, \mathbf{m}(\omega)), \quad (6.88)$$

where $\nabla_{\omega} c$ is the gradient of $c(\omega, \mathbf{m}(\omega)) = c(x, \mathbf{m}(x))$ (the cost function rewritten in polar stereographic coordinates) with respect to ω , defined as

$$\nabla_{\omega} = \hat{\mathbf{e}}_{\rho} \frac{\partial}{\partial \rho} + \frac{1}{\rho} \hat{\mathbf{e}}_{\zeta} \frac{\partial}{\partial \zeta}, \quad (6.89)$$

with $\hat{\mathbf{e}}_{\rho} = \cos(\zeta) \hat{\mathbf{e}}_{x_1} + \sin(\zeta) \hat{\mathbf{e}}_{x_2}$ and $\hat{\mathbf{e}}_{\zeta} = -\sin(\zeta) \hat{\mathbf{e}}_{x_1} + \cos(\zeta) \hat{\mathbf{e}}_{x_2}$. We require the Hessian w.r.t. ω of $-u_1(\omega) - c(\omega, \mathbf{y})$ to be SND/SPD for a c-convex/c-concave solution, respectively, as explained in Section 4.3.1. The Hessian matrix of a function f in polar coordinates is given by [4]

$$H[f] = \begin{pmatrix} \frac{\partial^2 f}{\partial \rho^2} & \frac{1}{\rho} \frac{\partial^2 f}{\partial \rho \partial \zeta} - \frac{1}{\rho^2} \frac{\partial f}{\partial \zeta} \\ \frac{1}{\rho} \frac{\partial^2 f}{\partial \rho \partial \zeta} - \frac{1}{\rho^2} \frac{\partial f}{\partial \zeta} & \frac{1}{\rho^2} \frac{\partial^2 f}{\partial \zeta^2} + \frac{1}{\rho} \frac{\partial f}{\partial \zeta} \end{pmatrix}. \quad (6.90)$$

For the Hessian of $-u_1(\omega) - c(\omega, \mathbf{m}(\omega))$ all first derivative terms in the above matrix cancel because of (6.88). Hence, the SND/SPD Hessian is given by

$$H[-u_1 - c(\omega, \mathbf{y})] = -D^2 u_1(\omega) - D_{\omega \omega} c(\omega, \mathbf{y}) = \mathbf{P}, \quad (6.91)$$

where

$$D^2 u_1(\boldsymbol{\omega}) = \begin{pmatrix} \frac{\partial^2 u_1}{\partial \rho^2} & \frac{1}{\rho} \frac{\partial^2 u_1}{\partial \rho \partial \zeta} \\ \frac{1}{\rho} \frac{\partial^2 u_1}{\partial \rho \partial \zeta} & \frac{1}{\rho^2} \frac{\partial^2 u_1}{\partial \zeta^2} \end{pmatrix}, \quad (6.92a)$$

$$D_{\boldsymbol{\omega}\boldsymbol{\omega}} c(\boldsymbol{\omega}, \mathbf{y}) = \begin{pmatrix} \frac{\partial^2 c}{\partial \rho^2} & \frac{1}{\rho} \frac{\partial^2 c}{\partial \rho \partial \zeta} \\ \frac{1}{\rho} \frac{\partial^2 c}{\partial \rho \partial \zeta} & \frac{1}{\rho^2} \frac{\partial^2 c}{\partial \zeta^2} \end{pmatrix}. \quad (6.92b)$$

Note that $D^2 u_1(\boldsymbol{\omega})$ and $D_{\boldsymbol{\omega}\boldsymbol{\omega}} c(\boldsymbol{\omega}, \mathbf{y})$ are *not* Hessian matrices.

Differentiating (6.88) again w.r.t. $\boldsymbol{\omega}$ gives

$$D_{\boldsymbol{\omega}\boldsymbol{\omega}} c(\boldsymbol{\omega}, \mathbf{m}(\boldsymbol{\omega})) + C D\mathbf{m}(\boldsymbol{\omega}) + D^2 u_1(\boldsymbol{\omega}) = \mathbf{O}, \quad (6.93)$$

where

$$C = C(\boldsymbol{\omega}, \mathbf{m}(\boldsymbol{\omega})) = D_{\boldsymbol{\omega}\mathbf{y}} c = \begin{pmatrix} \frac{\partial^2 c}{\partial \rho \partial y_1} & \frac{\partial^2 c}{\partial \rho \partial y_2} \\ \frac{1}{\rho} \frac{\partial^2 c}{\partial \zeta \partial y_1} & \frac{1}{\rho} \frac{\partial^2 c}{\partial \zeta \partial y_2} \end{pmatrix}, \quad (6.94a)$$

$$D\mathbf{m}(\boldsymbol{\omega}) = \begin{pmatrix} \frac{\partial m_1}{\partial \rho} & \frac{1}{\rho} \frac{\partial m_1}{\partial \zeta} \\ \frac{\partial m_2}{\partial \rho} & \frac{1}{\rho} \frac{\partial m_2}{\partial \zeta} \end{pmatrix}. \quad (6.94b)$$

Substituting $\mathbf{y} = \mathbf{m}(\boldsymbol{\omega})$ in (6.91) and combining (6.91) and (6.93) gives the matrix equation

$$C D\mathbf{m}(\boldsymbol{\omega}) = \mathbf{P}. \quad (6.95)$$

We also need to rewrite the energy conservation relation, which is different for every optical system. An example for a point-to-far-field single surface is given below.

Example 6.2.1. *We consider the systems in Section 3.4 and 3.5. For ease of notation we let $\tilde{f}(\boldsymbol{\omega}) = \tilde{f}(\mathbf{x})$ and $\tilde{g}(\mathbf{m}(\boldsymbol{\omega})) = \tilde{g}(\mathbf{m}(\mathbf{x}))$. Changing coordinates in (3.94) gives*

$$\begin{aligned} & \int_{\boldsymbol{\omega}(\mathcal{A})} \tilde{f}(\boldsymbol{\omega}) \frac{4\rho}{(1+\rho^2)^2} d\boldsymbol{\omega} \\ &= \int_{\boldsymbol{\omega}(\mathcal{A})} \tilde{g}(\mathbf{m}(\boldsymbol{\omega})) \frac{4\rho}{(1+|\mathbf{m}(\boldsymbol{\omega})|^2)^2} \det(D\mathbf{m}(\boldsymbol{\omega})) d\boldsymbol{\omega}, \end{aligned}$$

using that $\det(D\mathbf{m}(\boldsymbol{\omega})) = \det(D\mathbf{m}(\mathbf{x}))$. We derive the generalized Monge-Ampère equation

$$\det(D\mathbf{m}(\boldsymbol{\omega})) = \frac{\tilde{f}(\boldsymbol{\omega}) (1 + |\mathbf{m}(\boldsymbol{\omega})|^2)^2}{\tilde{g}(\mathbf{m}(\boldsymbol{\omega})) (1 + \rho^2)^2} = \frac{\det(\mathbf{P}(\boldsymbol{\omega}))}{\det(C(\boldsymbol{\omega}, \mathbf{m}(\boldsymbol{\omega})))} = \tilde{F}(\boldsymbol{\omega}, \mathbf{m}(\boldsymbol{\omega})),$$

where we introduce $\tilde{F}(\boldsymbol{\omega}, \mathbf{m}(\boldsymbol{\omega}))$ to denote the total right hand side. Again, we define the corresponding transport boundary condition as

$$\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}.$$

The introduction of polar coordinates requires a couple of changes to the numerical algorithm. In this section, I will only present the modifications with respect to the numerical algorithm described in Section 6.1.

First, the initial mappings in (6.10) and (6.11) need to be modified. As initial guess \mathbf{m}^0 we map the circular source domain \mathcal{X} centered around the origin O enclosing \mathcal{X} to a disc on top of \mathcal{Y} . We assume that the source \mathcal{X} has radius ρ_{\max} and that the target \mathcal{Y} can also be enclosed by a bounding box of rectangular shape $[a_{\min}, a_{\max}] \times [b_{\min}, b_{\max}]$. We will use this bounding box to compute the radius R of the disc. We consider two options for an initial guess \mathbf{m}^0 . We can specify the initial guess $\mathbf{m}^0 = (m_1^0, m_2^0)$ with Cartesian components (m_1^0, m_2^0) for $(\rho, \zeta) \in \mathcal{X}$, with $0 \leq \rho \leq \rho_{\max}$ and $0 \leq \zeta < 2\pi$, as

$$m_1^0 = \rho R \cos(\zeta) + \frac{a_{\min} + a_{\max}}{2}, \quad (6.96a)$$

$$m_2^0 = \rho R \sin(\zeta) + \frac{b_{\min} + b_{\max}}{2}, \quad (6.96b)$$

where

$$R = \frac{\max(a_{\max} - a_{\min}, b_{\max} - b_{\min})}{2\rho_{\max}}. \quad (6.96c)$$

In other words, the initial guess maps \mathcal{X} onto a disc of radius $\rho_{\max} R$ centered around $((a_{\min} + a_{\max})/2, (b_{\min} + b_{\max})/2)$. For example, for a square target domain \mathcal{Y} this disc is located inside the rectangle and just fits inside the domain. The corresponding Jacobi matrix $D\mathbf{m}^0$ is SPD.

Alternatively, we specify the initial guess $\mathbf{m}^0 = (m_1^0, m_2^0)$ for $(\rho, \zeta) \in \mathcal{X}$, with $0 \leq \rho \leq \rho_{\max}$ and $0 \leq \zeta < 2\pi$, as

$$m_1^0 = -\rho R \cos(\zeta) + \frac{a_{\min} + a_{\max}}{2}, \quad (6.97a)$$

$$m_2^0 = -\rho R \sin(\zeta) + \frac{b_{\min} + b_{\max}}{2}, \quad (6.97b)$$

with R as in (6.96c). This mapping has a Jacobi matrix that is SND. For each optical system we determine which initial mapping is suitable using similar reasoning as in Example 6.1.1.

After setting the initial guess \mathbf{m}^0 we perform the minimization steps in (6.8) and subsequently update \mathbf{C} in every iteration. The minimization steps (6.8a)

and (6.8b) are performed point-wise and described in detail in Section 6.1.1 and 6.1.2. The operations in these point-wise minimization steps are completely analogous to the Cartesian case. In contrast, the minimization step (6.8c) and the subsequent calculation of the lens surface cannot be performed point-wise and require more alterations when using polar stereographic coordinates. We will describe this in Section 6.2.1 and 6.2.2.

6.2.1 Minimization procedure for m

We minimize the combined functional $J[m, \mathbf{P}, \mathbf{b}]$ over all $m \in \mathcal{M}$. This step cannot be performed point-wise and we compute the first variation $\delta J[m, \mathbf{P}, \mathbf{b}](\eta)$ with respect to m for $\eta \in \mathcal{M}$, i.e.,

$$\begin{aligned} \delta J[m, \mathbf{P}, \mathbf{b}](\eta) &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \left(J[m + \epsilon \eta, \mathbf{P}, \mathbf{b}] - J[m, \mathbf{P}, \mathbf{b}] \right) \\ &= \lim_{\epsilon \rightarrow 0} \left[\frac{\alpha}{2} \int_{\mathcal{X}} 2\rho (\mathbf{C} \mathbf{D}m - \mathbf{P}) : \mathbf{C} \mathbf{D}\eta + \epsilon \rho \|\mathbf{C} \mathbf{D}\eta\|^2 d\omega \right. \\ &\quad \left. + \frac{1-\alpha}{2} \oint_{\partial\mathcal{X}} 2(m - \mathbf{b}) \cdot \eta + \epsilon |\eta|^2 ds \right] \\ &= \alpha \int_{\mathcal{X}} (\mathbf{C} \mathbf{D}m - \mathbf{P}) : \mathbf{C} \mathbf{D}\eta \rho d\omega + (1-\alpha) \oint_{\partial\mathcal{X}} (m - \mathbf{b}) \cdot \eta ds. \end{aligned} \quad (6.98)$$

The minimizer is given by $\delta J[m, \mathbf{P}, \mathbf{b}](\eta) = 0$ for all $\eta \in \mathcal{M}$. Using Gauss's law and the fundamental lemma of calculus of variations [32], as we did in Section 6.1.3, we obtain the boundary value problem

$$\nabla \cdot (\mathbf{C}^T \mathbf{C} \mathbf{D}m) = \nabla \cdot (\mathbf{C}^T \mathbf{P}), \quad \omega \in \mathcal{X}, \quad (6.99a)$$

$$(1-\alpha) m + \alpha (\mathbf{C}^T \mathbf{C} \mathbf{D}m) \hat{n} = (1-\alpha) \mathbf{b} + \alpha \mathbf{C}^T \mathbf{P} \hat{n}, \quad \omega \in \partial\mathcal{X}, \quad (6.99b)$$

where the divergence operator in polar coordinates works on the matrices $\mathbf{A} = (a_{ij}) = \mathbf{C}^T \mathbf{C} \mathbf{D}m$ and $\mathbf{A} = (a_{ij}) = \mathbf{C}^T \mathbf{P}$ as follows:

$$\nabla \cdot \mathbf{A} = \begin{pmatrix} \frac{1}{\rho} \frac{\partial}{\partial \rho} (\rho a_{11}) + \frac{1}{\rho} \frac{\partial}{\partial \zeta} a_{12} \\ \frac{1}{\rho} \frac{\partial}{\partial \rho} (\rho a_{21}) + \frac{1}{\rho} \frac{\partial}{\partial \zeta} a_{22} \end{pmatrix}. \quad (6.100)$$

Hence, we obtain two coupled equations for the components m_1 and m_2 of m .

We solve for m using the finite volume method on a polar grid and eliminate the singularity at the origin by considering a control area around the origin, as explained in Appendix B.2. As in Section 6.1.3, we first compute the component m_1 , substituting m_2 from the previous iteration, and then compute the component m_2 using the new-found m_1 . We compute m^{n+1} in one step, i.e., we do not perform multiple iterations.

6.2.2 Computation of u_1

Upon convergence of the iterative procedure for the mapping \mathbf{m} , we can calculate the location of the optical surface from (6.88). As in Section 6.1.4, we compute the generalized least-squares solution by minimizing the functional

$$I[u_1] = \frac{1}{2} \int_{\mathcal{X}} |\nabla u_1 + \nabla_{\omega} c(\cdot, \mathbf{m})|^2 \rho \, d\omega, \quad (6.101)$$

where $u_1(\omega) = u_1(\mathbf{x})$ and we use the short-hand notation $\nabla_{\omega} c(\cdot, \mathbf{m}) = \nabla_{\omega} c(\omega, \mathbf{m}(\omega)) = \nabla c(\omega, \mathbf{y})|_{\mathbf{y}=\mathbf{m}(\omega)}$.

We cannot perform this step point-wise and analogous to the minimization procedure for \mathbf{m} , we compute the first variation $\delta I[u_1](v)$ with respect to u_1 for $v \in C^2(\mathcal{X})$ as

$$\begin{aligned} \delta I[u_1](v) &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \left(I[u_1 + \epsilon v] - I[u_1] \right) \\ &= \lim_{\epsilon \rightarrow 0} \frac{1}{2} \int_{\mathcal{X}} 2\rho (\nabla u_1 + \nabla_{\omega} c(\cdot, \mathbf{m})) \cdot \nabla v + \epsilon \rho |v|^2 \, d\omega \end{aligned} \quad (6.102)$$

$$= \int_{\mathcal{X}} (\nabla u_1 + \nabla_{\omega} c(\cdot, \mathbf{m})) \cdot \nabla v \rho \, d\omega. \quad (6.103)$$

The minimizer is given by $\delta I[u_1](v) = 0$ for all $v \in C^2(\mathcal{X})$. Once more, using Gauss's law and the fundamental lemma of calculus of variations [32], we obtain the boundary value problem

$$\nabla \cdot \nabla u_1 = -\nabla \cdot \nabla_{\omega} c(\cdot, \mathbf{m}), \quad \omega \in \mathcal{X}, \quad (6.104a)$$

$$\frac{\partial u_1}{\partial \rho} = -\frac{\partial c(\cdot, \mathbf{y})}{\partial \rho} \Big|_{\mathbf{y}=\mathbf{m}(\omega)}, \quad \omega \in \partial \mathcal{X}, \quad (6.104b)$$

where the divergence operator works on the vector $\mathbf{w} = (w_1, w_2) = \nabla u_1$ and $\mathbf{w} = (w_1, w_2) = \nabla_{\omega} c(\cdot, \mathbf{m})$ as follows:

$$\nabla \cdot \mathbf{w} = \frac{1}{\rho} \frac{\partial}{\partial \rho} (\rho w_1) + \frac{1}{\rho} \frac{\partial}{\partial \bar{\zeta}} w_2.$$

This is a Neumann problem which has a unique solution up to a constant, and consequently a corresponding finite volume matrix with incomplete rank. We calculate a unique least-squares solution using the method described in Section 6.1.4.

Finally, we compute the location of the optical surface u using its relation to u_1 , as illustrated by the example below.

Example 6.2.2. For example, for a point-to-far-field lens, we have the relation $u_1(\omega) = -\log(u(\omega))$, cf. the change of variables introduced for (3.106), rewritten in polar coordinates. Hence, we calculate $u(\omega) = e^{-u_1(\omega)}$. We transform the polar coordinates, which are polar stereographic coordinates for a point source, defined in (3.7) and (6.87) to Cartesian coordinates $(x, y, z) = u(\omega) \hat{s}$ and plot the reflector surface using

$$x = \frac{2 u(\omega) \rho \cos(\zeta)}{1 + \rho^2}, \quad y = \frac{2 u(\omega) \rho \sin(\zeta)}{1 + \rho^2}, \quad z = \frac{u(\omega) (1 - \rho^2)}{1 + \rho^2},$$

cf. (3.9), for all $\omega \in \mathcal{X}$.

6.3 The GJLS algorithm

In this section, we discuss the GJLS algorithm, which can be used to find solutions for all 16 base cases in Chapter 3. For these optical systems the generated Jacobian equations are all equations of the form

$$\det(D\mathbf{m}(\mathbf{x})) = \frac{\det(\mathbf{P}(\mathbf{x}))}{\det(\mathbf{C}(\mathbf{x}, \mathbf{m}(\mathbf{x}), u(\mathbf{x})))} = \frac{f(\mathbf{x})}{g(\mathbf{m}(\mathbf{x}))} = F(\mathbf{x}, \mathbf{m}(\mathbf{x}), u(\mathbf{x})), \quad (6.105)$$

cf. (4.83), where $\mathbf{y} = \mathbf{m}(\mathbf{x}) = \bar{\mathbf{m}}(\mathbf{x}, u, \nabla u)$ and $u(\mathbf{x}) = G(\mathbf{x}, \mathbf{y}, w)$ is the optical surface we are interested in. Instead of taking the cost function as input, as in the GLS algorithm, the GJLS algorithm takes the generating function as input.

We compute the mapping \mathbf{m} and surface u from (6.105) by using the extension to the GLS algorithm presented in Section 6.1. The extension of the least-squares algorithm to a generating-function approach lies in the additional dependency of the mixed Hessian matrix $\mathbf{C} = \mathbf{C}(\mathbf{x}, \mathbf{m}(\mathbf{x}), u(\mathbf{x})) = D_{xy} \tilde{H}$, cf. (4.82), on the surface $u(\mathbf{x})$.

To compute the mapping \mathbf{m} , we write the Monge-Ampère equation (6.105) as the matrix equation

$$\mathbf{P}(\mathbf{x}) = \mathbf{C}(\mathbf{x}, \mathbf{m}(\mathbf{x}), u(\mathbf{x})) D\mathbf{m}(\mathbf{x}), \quad (6.106)$$

cf. (4.81), with $\mathbf{P}(\mathbf{x})$ an SND/SPD matrix satisfying $\det(\mathbf{P}) = F \det(\mathbf{C})$. From Section 4.4.2 we know that if $G_w > 0$ we need an SND matrix \mathbf{P} for a G-convex u and an SPD matrix \mathbf{P} for a G-concave u . If $G_w < 0$ we need an SPD matrix \mathbf{P} for a G-convex u and an SND matrix \mathbf{P} for a G-concave u .

We write $\mathbf{m} = \mathbf{m}(\mathbf{x})$ and enforce the matrix equation (6.106) by minimizing the functional

$$J_I[\mathbf{m}, \mathbf{P}] = \frac{1}{2} \int_{\mathcal{X}} \|\mathbf{C} D\mathbf{m} - \mathbf{P}\|^2 dx, \quad (6.107)$$

under the constraint $\det(\mathbf{P}) = F \det(\mathbf{C})$. The norm used is the Frobenius norm. To impose the transport boundary condition (6.2) we minimize the functional

$$J_B[\mathbf{m}, \mathbf{b}] = \frac{1}{2} \oint_{\partial\mathcal{X}} |\mathbf{m} - \mathbf{b}|^2 ds \quad (6.108)$$

over \mathbf{b} , where $|\cdot|$ denotes the L_2 -norm and \mathbf{b} is a function from the source boundary to the target boundary, i.e., $\mathbf{b} : \partial\mathcal{X} \rightarrow \partial\mathcal{Y}$. By minimizing this functional we aim to impose $\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}$, which holds if $J_B[\mathbf{m}, \mathbf{b}] = 0$. We combine the functionals J_I and J_B into the weighted average

$$J[\mathbf{m}, \mathbf{P}, \mathbf{b}] = \alpha J_I[\mathbf{m}, \mathbf{P}] + (1 - \alpha) J_B[\mathbf{m}, \mathbf{b}], \quad (6.109)$$

with $0 < \alpha < 1$.

We use $\nabla_x H(\mathbf{x}, \mathbf{y}, u(\mathbf{x})) + H_w(\mathbf{x}, \mathbf{y}, u(\mathbf{x})) \nabla u(\mathbf{x}) = \mathbf{0}$ from (4.77) to compute $u = u(\mathbf{x})$ from \mathbf{m} and minimize the functional

$$I[u, \mathbf{m}] = \frac{1}{2} \int_{\mathcal{X}} |\nabla_x H(\mathbf{x}, \mathbf{m}, u) + H_w(\mathbf{x}, \mathbf{m}, u) \nabla u|^2 dx, \quad (6.110)$$

detailed in Section 6.3.1, where $|\cdot|$ denotes the L_2 -norm.

We use initial guesses \mathbf{m}^0 and u^0 , specified below, and the mixed Hessian matrix $\mathbf{C}(\cdot, \mathbf{m}^0, u^0)$. Let $n = 0$ and compute

$$\mathbf{b}^{n+1} = \operatorname{argmin}_{\mathbf{b} \in \mathcal{B}} J_B[\mathbf{m}^n, \mathbf{b}], \quad (6.111a)$$

$$\mathbf{P}^{n+1} = \operatorname{argmin}_{\mathbf{P} \in \mathcal{P}(\mathbf{m}^n)} J_I[\mathbf{m}^n, \mathbf{P}], \quad (6.111b)$$

$$\mathbf{m}^{n+1} = \operatorname{argmin}_{\mathbf{m} \in \mathcal{M}} J[\mathbf{m}, \mathbf{P}^{n+1}, \mathbf{b}^{n+1}], \quad (6.111c)$$

$$u^{n+1} = \operatorname{argmin}_{u \in \mathcal{U}} I[u, \mathbf{m}^{n+1}], \quad (6.111d)$$

where the minimization steps are performed over the spaces

$$\mathcal{B} = \{\mathbf{b} \in C^1(\partial\mathcal{X})^2 \mid \mathbf{b}(\mathbf{x}) \in \partial\mathcal{Y}\}, \quad (6.112a)$$

$$\begin{aligned} \mathcal{P}(\mathbf{m}, u) &= \{\mathbf{P} \in C^1(\mathcal{X})^{2 \times 2} \mid \mathbf{P} \text{ SND / SPD}, \\ &\quad \det(\mathbf{P}) = F(\cdot, \mathbf{m}, u) \det(\mathbf{C}(\cdot, \mathbf{m}))\}, \end{aligned} \quad (6.112b)$$

$$\mathcal{M} = C^2(\mathcal{X})^2, \quad (6.112c)$$

$$\mathcal{U} = C^2(\mathcal{X}). \quad (6.112d)$$

After each iteration we compute $\mathbf{C}(\cdot, \mathbf{m}^{n+1}, u^{n+1})$.

As initial guess \mathbf{m}^0 we map the smallest bounding box enclosing \mathcal{X} to the smallest bounding box enclosing \mathcal{Y} . We use either (6.10) or (6.11) as initial guess and check whether $\det(\mathbf{P}^0) = \det(\mathbf{C}(\cdot, \mathbf{m}^0, u^0)) \det(\mathbf{D}\mathbf{m}^0) > 0$,

i.e., if $\det(\mathbf{C}(\cdot, \mathbf{m}^0, u^0)) > 0$ since the initial guess satisfies $\det(\mathbf{D}\mathbf{m}^0) > 0$, and $\text{tr}(\mathbf{P}^0) \leq 0$ or $\text{tr}(\mathbf{P}^0) \geq 0$ for a negative or positive semi-definite matrix. Note that to verify this we also need to initialize the surface u . We take $u^0(\mathbf{x}) = c$, with c a constant. The example below briefly illustrates how we determine which initial guess to use.

Example 6.3.1. *For the parallel-to-near-field reflector in Section 3.3 the initial surface $u^0(\mathbf{x}) = c$ is a flat reflector surface. For a point-to-far-field lens in Section 3.5 the initial surface $u^0(\mathbf{x}) = c$ is a spherical lens, since the radial surface parameter u^0 is constant. Using these initial guesses we can show that the initial guess (6.11) gives $\det(\mathbf{P}^0) > 0$ since $\det(\mathbf{C}(\cdot, \mathbf{m}^0, u^0)) > 0$ and $\det(\mathbf{D}\mathbf{m}^0) > 0$. Moreover, $\text{tr}(\mathbf{P}^0) = 1/2 \text{tr}(\mathbf{C}(\cdot, \mathbf{m}^0, u^0)) \text{tr}(\mathbf{D}\mathbf{m}^0)$, since we can show that the diagonal elements of $\mathbf{C} = \mathbf{D}_{xy} \tilde{H}(\mathbf{x}, \mathbf{m}^0, u^0)$ are equal in this case for both systems, and $\mathbf{D}\mathbf{m}^0$ is a diagonal matrix. Substituting \mathbf{m}^0 and u^0 we find that $\text{tr}(\mathbf{C}(\cdot, \mathbf{m}^0, u^0)) \leq 0$ and $\text{tr}(\mathbf{D}\mathbf{m}^0) > 0$. Hence, $\mathbf{P}^0 = \mathbf{C}(\cdot, \mathbf{m}^0) \mathbf{D}\mathbf{m}^0$ is SND. Since $G_w(\mathbf{x}, \mathbf{m}^0, u^0) > 0$ we have a max/min G-conjugate pair and calculate a G-convex u .*

We discretize the source domain \mathcal{X} using a standard rectangular $N_1 \times N_2$ grid for some $N_1, N_2 \in \mathbb{N}$ and introduce $\mathbf{x}_{ij} = (x_{1,j}, x_{2,j})$ as in (6.12).

The minimization steps (6.111a), (6.111b) and (6.111c) are described in detail in Section 6.1.1, 6.1.2 and 6.1.3 (and 6.2.1), respectively. Evidently, in the minimization of \mathbf{P} in Section 6.1.2 we now impose G-convexity/G-concavity instead of c-convexity/c-concavity. Minimization step (6.111d) is the new step required for the generating-function approach. It is explained in detail in the next section. After each iteration we compute the matrix $\mathbf{C}(\cdot, \mathbf{m}^{n+1}, u^{n+1})$. Figure 6.4 shows a flow chart of the steps in the numerical procedure. The stopping criterion for the iterative procedure is explained in Chapter 7.

6.3.1 Minimization procedure for u

To find u we use the implicit relations in (4.77) and (4.78), i.e.,

$$\nabla_{\mathbf{x}} \tilde{H}(\mathbf{x}, \mathbf{y}) = \nabla_{\mathbf{x}} H(\mathbf{x}, \mathbf{y}, u(\mathbf{x})) + H_w(\mathbf{x}, \mathbf{y}, u(\mathbf{x})) \nabla u(\mathbf{x}) = \mathbf{0}. \quad (6.113)$$

We can compute u by minimizing the functional

$$I[u, \mathbf{y}] = \frac{1}{2} \int_{\mathcal{X}} |\nabla_{\mathbf{x}} H(\mathbf{x}, \mathbf{y}, u) + H_w(\mathbf{x}, \mathbf{y}, u) \nabla u|^2 \, d\mathbf{x}. \quad (6.114)$$

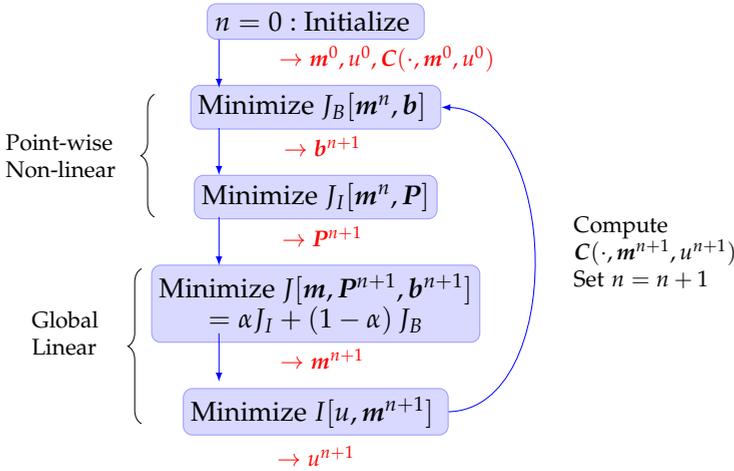


Figure 6.4: Flow chart of the GJLS algorithm.

Analogous to the minimization procedure for \mathbf{m} detailed in Section 6.1.3, we compute the first variation of $\delta I[u, \mathbf{y}](v)$ with respect to u for $v \in C^2(\mathcal{X})$ as

$$\begin{aligned}
 \delta I[u, \mathbf{y}](v) &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \left(I[u + \epsilon v, \mathbf{y}] - I[u, \mathbf{y}] \right) = \frac{d}{d\epsilon} I[u + \epsilon v, \mathbf{y}] \Big|_{\epsilon=0} \\
 &= \frac{d}{d\epsilon} \frac{1}{2} \int_{\mathcal{X}} |\nabla_x H(\mathbf{x}, \mathbf{y}, u + \epsilon v) + H_w(\mathbf{x}, \mathbf{y}, u + \epsilon v) \nabla(u + \epsilon v)|^2 d\mathbf{x} \Big|_{\epsilon=0} \\
 &= \int_{\mathcal{X}} \nabla_x \tilde{H} \cdot \frac{d}{d\epsilon} \left(\nabla_x H(\mathbf{x}, \mathbf{y}, u + \epsilon v) + H_w(\mathbf{x}, \mathbf{y}, u + \epsilon v) \nabla(u + \epsilon v) \right) \Big|_{\epsilon=0} d\mathbf{x},
 \end{aligned} \tag{6.115}$$

where $\tilde{H} = \tilde{H}(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}, \mathbf{y}, u)$. Using Taylor expansions, the short-hand notation $H = H(\mathbf{x}, \mathbf{y}, u)$, and the expression $\mathcal{O}(\epsilon^2)$ for the higher than first-order terms, we obtain

$$\begin{aligned}
 \delta I[u, \mathbf{y}](v) &= \int_{\mathcal{X}} \nabla_x \tilde{H} \cdot \frac{d}{d\epsilon} \left(\nabla_x H + \epsilon v \nabla_x H_w \right. \\
 &\quad \left. + H_w \nabla u + \epsilon v H_{ww} \nabla u + \epsilon H_w \nabla v + \mathcal{O}(\epsilon^2) \right) \Big|_{\epsilon=0} d\mathbf{x} \\
 &= \int_{\mathcal{X}} \nabla_x \tilde{H} \cdot (v \nabla_x H_w + v H_{ww} \nabla u + H_w \nabla v) d\mathbf{x} \\
 &= \int_{\mathcal{X}} \nabla_x \tilde{H} \cdot (\nabla_x H_w + H_{ww} \nabla u) v d\mathbf{x} + \int_{\mathcal{X}} H_w \nabla_x \tilde{H} \cdot \nabla v d\mathbf{x} \\
 &= \int_{\mathcal{X}} \frac{1}{2} \frac{d}{dw} |\nabla_x \tilde{H}|^2 v d\mathbf{x} + \int_{\mathcal{X}} H_w \nabla_x \tilde{H} \cdot \nabla v d\mathbf{x}.
 \end{aligned} \tag{6.116}$$

For the second integral, we use Gauss's theorem and the vector-scalar product rule, i.e.,

$$\int_{\mathcal{X}} \mathbf{F} \cdot \nabla v \, dx = \oint_{\partial\mathcal{X}} \mathbf{F} v \cdot \hat{\mathbf{n}} \, ds - \int_{\mathcal{X}} (\nabla \cdot \mathbf{F}) v \, dx, \quad (6.117)$$

with $\mathbf{F} = H_w \nabla_x \tilde{H}$. We now set $\mathbf{y} = \mathbf{m}(x)$ and let $\nabla \cdot$ denote the divergence operator with respect to \mathbf{x} , taking into account the dependencies $\mathbf{y} = \mathbf{m}(x)$ and $u = u(x)$ via the chain rule. The gradient ∇_x still only works on the first variable of $\tilde{H}(\mathbf{x}, \mathbf{y})$. Hence, we obtain

$$\begin{aligned} \delta I[u, \mathbf{m}](v) &= \int_{\mathcal{X}} \frac{1}{2} \frac{d}{dw} \left| \nabla_x \tilde{H} \right|^2 v - (\nabla \cdot (H_w \nabla_x \tilde{H})) v \, dx \\ &\quad + \oint_{\partial\mathcal{X}} H_w \nabla_x \tilde{H} v \cdot \hat{\mathbf{n}} \, ds. \end{aligned} \quad (6.118)$$

The minimizer is given by $\delta I[u, \mathbf{m}](v) = 0$ for all $v \in C^2(\mathcal{X})$ and results in the boundary value problem

$$\nabla \cdot (H_w \nabla_x \tilde{H}) = \frac{1}{2} \frac{d}{dw} \left| \nabla_x \tilde{H} \right|^2, \quad \mathbf{x} \in \mathcal{X}, \quad (6.119a)$$

$$H_w \nabla_x \tilde{H} \cdot \hat{\mathbf{n}} = 0, \quad \mathbf{x} \in \partial\mathcal{X}. \quad (6.119b)$$

We rewrite (6.119) using $\tilde{H}(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}, \mathbf{y}, u(x))$ to

$$\nabla \cdot (H_w \nabla_x H + H_w^2 \nabla u) = \frac{1}{2} \frac{d}{dw} \left| \nabla_x H + H_w \nabla u \right|^2, \quad \mathbf{x} \in \mathcal{X}, \quad (6.120a)$$

$$H_w (\nabla_x H + H_w \nabla u) \cdot \hat{\mathbf{n}} = 0, \quad \mathbf{x} \in \partial\mathcal{X}. \quad (6.120b)$$

Substituting \mathbf{m}^{n+1} and the function $H(\mathbf{x}, \mathbf{m}^{n+1}, u)$ at iteration n , this is a Neumann problem for u which has a corresponding discretization matrix with incomplete rank. We calculate a unique solution by prescribing the average value of u as a constraint which adds an extra row to the discretization matrix, as in Section 6.1.4. In Appendix B.3 we present the full details of the finite volume method used to solve (6.120).

The Neumann problem only has a solution if the compatibility condition is satisfied. Integrating (6.120b) over $\partial\mathcal{X}$ and using Gauss's theorem gives

$$0 = \oint_{\partial\mathcal{X}} H_w (\nabla_x H + H_w \nabla u) \cdot \hat{\mathbf{n}} \, ds = \int_{\mathcal{X}} \nabla \cdot (H_w (\nabla_x H + H_w \nabla u)) \, dx, \quad (6.121)$$

and integrating (6.120a) over \mathcal{X} this reduces to

$$\int_{\mathcal{X}} \frac{1}{2} \frac{d}{dw} \left| \nabla_x H + H_w \nabla u \right|^2 \, dx = 0. \quad (6.122)$$

After computation of the surface u we check whether the compatibility condition is satisfied using Simpson's rule to approximate the integral in (6.122).

6.4 Summary

In this chapter, we gave a complete description of the GLS algorithm and an extension to polar coordinates in the source domain. This algorithm takes the cost function of an optical system as input and can be used to solve 9 out of the 16 base cases in Chapter 3. Subsequently, we presented the GJLS algorithm, which takes the generating function as input and can be used to compute the optical surfaces for all 16 base cases. In the next chapter, we will present numerical results for a range of base-case optical systems. We will also compare the performance of the GLS and GJLS procedures.

Chapter 7

Numerical Results – Part I

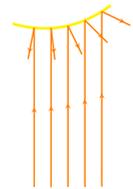
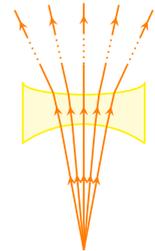
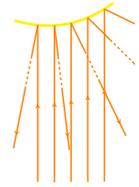
In this chapter, we will show numerical results computed with the GLS and GJLS algorithms. The laptop used for the numerical computations in this thesis has an Intel Core i7-7700HQ CPU 2.80 GHz with 32.0 GB of RAM.

We will cover a subset of the base cases discussed in Chapter 3 and consider the main systems that I have studied over the past four years. The optical systems that we consider in this chapter are presented in a slightly different order than in Table 3.1. We start with some examples to investigate the accuracy and convergence of the algorithms and subsequently increase the complexity of the simulations as follows:

- **Point-to-far-field reflector:** In Section 7.1, we first test the GLS algorithm on a problem for which the exact solution is a tilted flat reflector surface. We show that the numerical algorithm converges to the exact solution with an increasing number of grid points. Second, we compute a reflector surface which converts a square source domain into a circular target domain. For this problem we investigate the convergence of the algorithm in great detail, varying the number of grid points, the number of boundary points and α . The results of the tilted reflector surface and square-to-circle problem are published in [133].



- **Parallel-to-far-field reflector:** In Section 7.2, we use the GLS algorithm to compute a reflector surface that converts the light of a parallel beam with a nonuniform source emittance corresponding to a picture of a frog into a far-field target intensity corresponding to a picture of a prince. This simulation is an adaptation from the results published in [142]. Here we compute a reflector, while a lens system is considered in [142].
- **Point-to-far-field lens:** In Section 7.3.1, we first compute a peanut lens for road lighting applications using the GLS algorithm, as published in [131]. Second, we compare the performance of the GLS and GJLS algorithms for a problem for which the exact solution is an ellipsoidal lens surface. The results on the comparison of the algorithms are adapted from [130].
- **A parallel-to-near-field reflector:** This system does not have a cost function and, consequently, we can only use the GJLS algorithm. In Section 7.4, we compute a reflector surface for a target corresponding to a picture of my supervisor Jan. This simulation is adapted from [130].



7.1 Point-to-far-field reflector

We consider the point-to-far-field reflector problem in Section 3.4 with a point source at O .

7.1.1 Exact solution: tilted flat surface

One way to test our algorithm is by pre-computing the target domain corresponding to a known surface, for which we can derive the exact mapping. For example, we consider a tilted flat reflector surface $ax + by + cz = d$, with given constants $a, b, c, d > 0$. By substituting $x = u s_1, y = u s_2, z = u s_3$ and $u = e^{-u_1} (1 + |x|^2)$, and changing to stereographic coordinates using (3.9), we

can derive an expression for the c-concave solution u_1

$$u_1(\mathbf{x}) = -\log(d) + \log(2 \mathbf{a} \cdot \mathbf{x} + c(1 - |\mathbf{x}|^2)), \quad \mathbf{a} = \begin{pmatrix} a \\ b \end{pmatrix}. \quad (7.1)$$

Obviously, we require our source domain \mathcal{X} to lie within the interior of the ellipse $c^2 |\mathbf{x} - \frac{\mathbf{a}}{c}|^2 = |\mathbf{a}|^2 + c^2$, since we require that $2 \mathbf{a} \cdot \mathbf{x} + c(1 - |\mathbf{x}|^2) > 0$. Subsequently calculating $\frac{\partial u_1}{\partial x_1}$ and $\frac{\partial u_1}{\partial x_2}$, and substituting these into Equation (3.99) we obtain the mapping as

$$\mathbf{y} = \mathbf{m}(\mathbf{x}) = \frac{\mathbf{B} \mathbf{x} + \mathbf{a} c (-1 + |\mathbf{x}|^2)}{c^2 + 2 c \mathbf{a} \cdot \mathbf{x} + |\mathbf{a}|^2 |\mathbf{x}|^2}, \quad (7.2a)$$

where

$$\mathbf{B} = \begin{pmatrix} -a^2 + b^2 + c^2 & -2 a b \\ 2 a b & a^2 - b^2 + c^2 \end{pmatrix}. \quad (7.2b)$$

We consider a square source domain $\mathcal{X} = [-0.2, 0.2]^2$ and a 100×100 grid with $N_b = 1000$. We choose $a = 2$, $b = 1$, $c = 3$, and $d = 1$. Using Equation (7.2) we can compute the target domain \mathcal{Y} , the corresponding target boundary, and the correct mapping.

Next, we solve the boundary value problem

$$\det(\mathbf{D}\mathbf{m}(\mathbf{x})) = \frac{(1 + |\mathbf{m}(\mathbf{x})|^2)^2}{(1 + |\mathbf{x}|^2)^2} \frac{\tilde{f}(\mathbf{x})}{\tilde{g}(\mathbf{m}(\mathbf{x}))}, \quad (7.3)$$

cf. (3.100), for a uniform source and target distribution, i.e., $\tilde{f}(\mathbf{x}) = 1$ on \mathcal{X} and $\tilde{g}(\mathbf{y}) = 1$ on \mathcal{Y} . In order to satisfy global energy conservation we calculate

$$\tilde{F}(\mathcal{X}) = \int_{\mathcal{X}} \tilde{f}(\mathbf{x}) \frac{4}{(1 + |\mathbf{x}|^2)^2} d\mathbf{x}, \quad (7.4a)$$

and

$$\tilde{G}(\mathcal{Y}) = \int_{\mathcal{Y}} \tilde{g}(\mathbf{y}) \frac{4}{(1 + |\mathbf{y}|^2)^2} d\mathbf{y}, \quad (7.4b)$$

cf. (3.92) and (3.93) with $\mathcal{A} = \mathcal{X}$ and $\mathcal{Y} = \mathbf{m}(\mathcal{X})$. We normalize $\tilde{f}(\mathbf{x})$ and $\tilde{g}(\mathbf{y})$ by $\tilde{F}(\mathcal{X})$ and $\tilde{G}(\mathcal{Y})$, respectively. In fact, we know that $\tilde{F}(\mathcal{X}) = \tilde{G}(\mathcal{Y})$ is satisfied already by our choice of the target domain \mathcal{Y} corresponding to a flat reflector surface.

In the remaining numerical examples of this thesis, we also normalize the source and target intensities to make sure global energy conservation is satisfied, but we will not mention it explicitly.

We use the initial guess \mathbf{m}^0 given in (6.11) and find that $\text{tr}(\mathbf{P}^0) \geq 0$. Hence, we compute a c-concave u_1 , as explained in Section 4.3.1. Figure 7.1 shows the

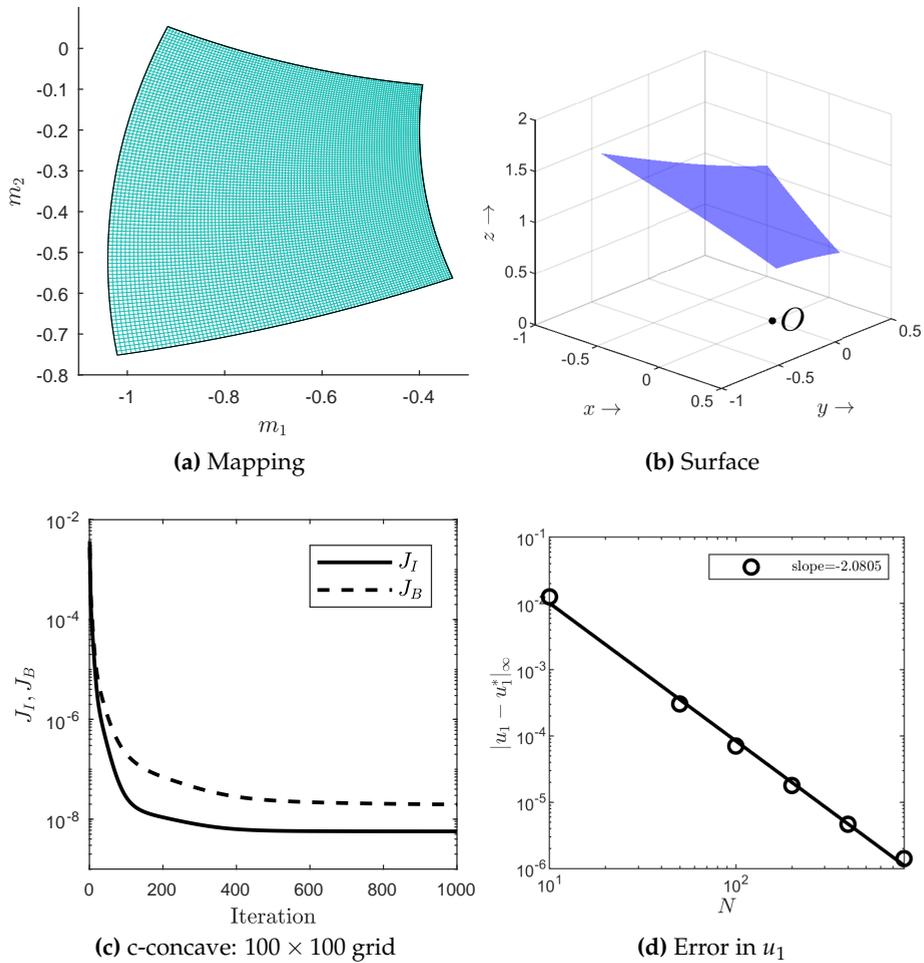


Figure 7.1: “Tilted-flat-surface” problem: (a) the mapping and (b) reflector surface after 1000 iterations on a 100×100 grid with $\alpha = 0.2$ and $N_b = 1000$, $\tilde{f}(x) = 1$ and $\tilde{g}(y) = 1$. Figure (c) shows the convergence history and (d) the maximum absolute difference between the computed reflector surface u_1 and exact solution u_1^* .

results after 1000 iterations. The converged mapping is displayed in Figure 7.1a and the reflector surface upon convergence in Figure 7.1b, calculated as explained in Example 6.1.2. The error convergence is given in Figure 7.1c. The maximum absolute difference between u_1 and the exact u_1 restricted to the grid after 1000 iterations is shown in Figure 7.1d as a function of the grid size $N = N_1 = N_2$, from $N = 10$ to $N = 800$, keeping $N_b = 1000$ constant. The slope of a logarithmic least-squares fit indicates second-order convergence to the exact surface.

7.1.2 Square-to-circle problem

We consider a square source domain $\mathcal{X} = [-0.5, 0.5]^2$ and circular target domain $\mathcal{Y} = \{(y_1, y_2) \in \mathbb{R}^2 \mid |y|^2 \leq 0.5\}$. We solve the boundary value problem (3.100) for a uniform source and target distribution, i.e., $\tilde{f}(x) = 1$ on \mathcal{X} and $\tilde{g}(y) = 1$ on \mathcal{Y} . As in the previous section, we use the initial guess \mathbf{m}^0 given in (6.11) and compute a c-concave u_1 . Figure 7.2a shows the converged mapping on a 50×50 grid.

To investigate convergence of the numerical algorithm, we introduce the norms

$$\|A\|_{2 \times 2} = \left(\iint_{\mathcal{X}} \|A\|^2 dx \right)^{1/2}, \quad \|a\|_2 = \left(\oint_{\partial \mathcal{X}} |a|^2 ds \right)^{1/2}, \quad (7.5)$$

for $A \in [C^1(\mathcal{X})]^{2 \times 2}$ and $a \in [C^1(\partial \mathcal{X})]^2$, as described in [125]. We denote $J_I^n = J_I[\mathbf{m}^n, \mathbf{P}^n]$ and $J_B^n = J_B[\mathbf{m}^n, \mathbf{b}^n]$. Using elementary properties of norms [125] we have

$$\left| \sqrt{J_I^{n+1}} - \sqrt{J_I^n} \right| \leq \frac{1}{\sqrt{2}} \left(\|C^{n+1} D\mathbf{m}^{n+1} - C^n D\mathbf{m}^n\|_{2 \times 2} + \|\mathbf{P}^{n+1} - \mathbf{P}^n\|_{2 \times 2} \right) =: c_I^n, \quad (7.6a)$$

$$\left| \sqrt{J_B^{n+1}} - \sqrt{J_B^n} \right| \leq \frac{1}{\sqrt{2}} \left(\|\mathbf{m}^{n+1} - \mathbf{m}^n\|_2 + \|\mathbf{b}^{n+1} - \mathbf{b}^n\|_2 \right) =: c_B^n. \quad (7.6b)$$

Figure 7.2 shows J_I, J_B for several $N \times N$ grids. It also shows the changes in $C D\mathbf{m}, \mathbf{m}|_{\partial \mathcal{X}}$ (\mathbf{m} on the boundary), \mathbf{P} and \mathbf{b} , with the updates in \mathbf{P} and \mathbf{b} . We used $N_b = 1000$ and $\alpha = 0.2$. The functionals J_I and J_B reach a plateau at a certain iteration number while the individual error terms continue to decrease up to machine precision. For the numerical simulations in the current section we introduce the stopping criterion

$$c_I^n \leq 0.1 \sqrt{J_I^n}, \quad \text{and} \quad c_B^n \leq 0.1 \sqrt{J_B^n}, \quad (7.7)$$

i.e., we require the relative change in both $\sqrt{J_I^m}$ and $\sqrt{J_B^m}$ to be less than 0.1. Figure 7.2 and Table 7.1 show the results using this stopping criterion, along with logarithmic least-squares fits. The number of iterations required increases sublinearly with N . We see that J_I has approximately third-order convergence and J_B second-order convergence, when using the stopping criterion. If we keep running the algorithm for longer, until the changes in $C Dm$ and $m|_{\partial\mathcal{X}}$ reach machine precision, we expect J_I and J_B to have approximately fourth-order and second-order convergence, respectively. This follows from the second-order discretization methods that we use in the numerical algorithm (e.g., for all differential operators and in the finite volume method). In general, we will not use this stopping criterion in the remainder of this thesis, only when indicated.

Grids	100×100	200×200	300×300	400×400	500×500	Fits
Iterations	208	351	466	563	650	$\propto N^{0.7}$
Time	22	115	333	772	1370	$\propto N^{2.6}$
J_I	4.0×10^{-7}	3.9×10^{-8}	1.2×10^{-8}	5.5×10^{-9}	3.3×10^{-9}	$\propto N^{-3.0}$
J_B	1.9×10^{-7}	4.6×10^{-8}	2.4×10^{-8}	1.6×10^{-8}	1.2×10^{-8}	$\propto N^{-1.7}$

Table 7.1: “Square-to-circle” problem: number of iterations, total computation time (in seconds) and residuals in the GLS algorithm.

Figure 7.3 shows the influence of N_b for a 200×200 grid. Choosing a larger N_b decreases J_I and J_B , but this effect becomes smaller as N_b becomes larger than N .

Figure 7.4a shows the calculation times of the minimization procedures for \mathbf{P} , \mathbf{b} , \mathbf{m} , and the computation of u_1 as a function of $N = N_1 = N_2$. As expected, the calculation time for the minimization procedure for \mathbf{P} is quadratic in N , and thus linear in the number of grid points. The minimization procedure for \mathbf{b} is linear in N . The calculation of \mathbf{m} and u_1 should be at least linear in the number of grid points and thus quadratic in N . As shown in Figure 7.4b, the minimization procedure for \mathbf{b} increases approximately linearly with increasing N_b . The slopes of the logarithmic least-squares fits are also displayed. The total calculation time for one iteration is approximately proportional to N^2 , and with the number of iterations growing sublinearly in N , see Table 7.1, the total calculation time scales roughly with N^3 .

The choice of the value of α is examined in Figure 7.5. We plot the functional $J = (1 - \alpha) J_I + \alpha J_B$, introduced in (6.7). We use the stopping criterion in (7.7) with a maximum of 400 iterations and see that the criterion and a lower J is reached sooner for small but not too small α , i.e., $\alpha = 0.1$ and $\alpha = 0.2$.

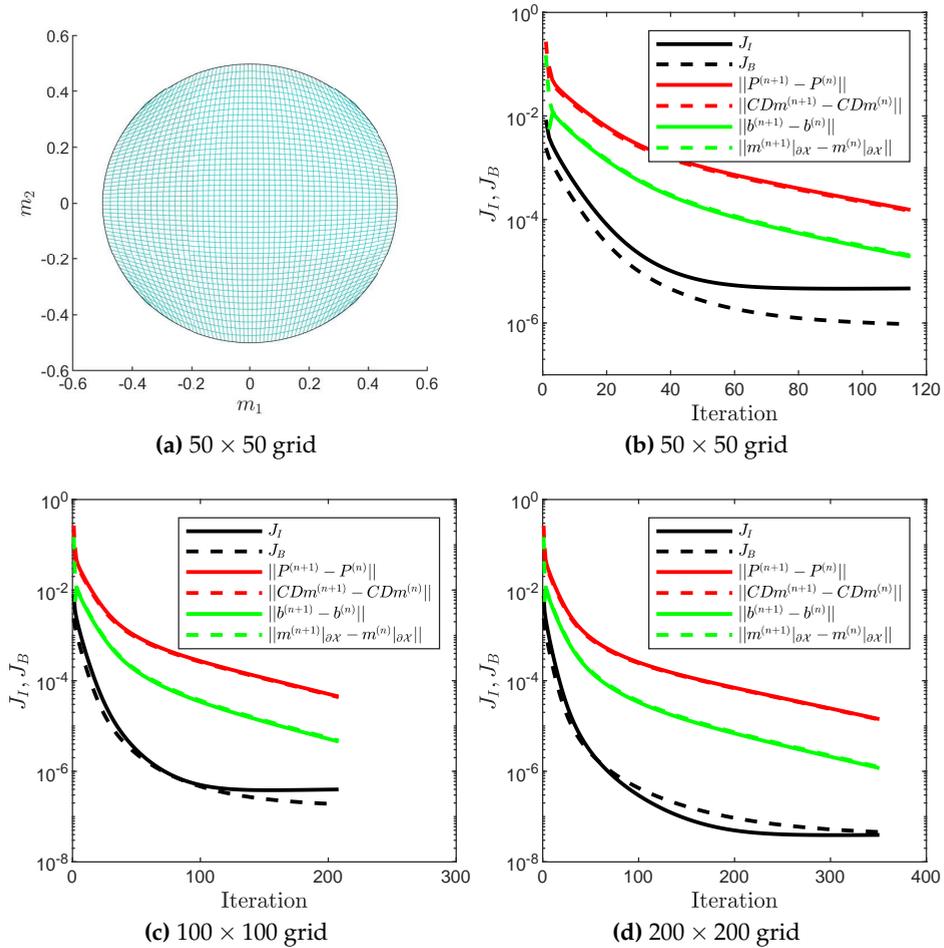


Figure 7.2: “Square-to-circle” problem: convergence history for several grid sizes. We calculate a c-concave solution u_1 and parameter values are $\alpha = 0.2$, $N_b = 1000$.

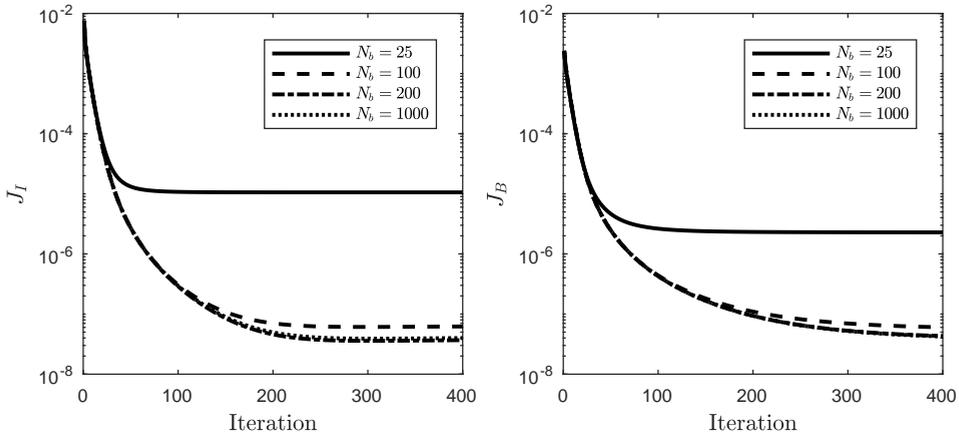
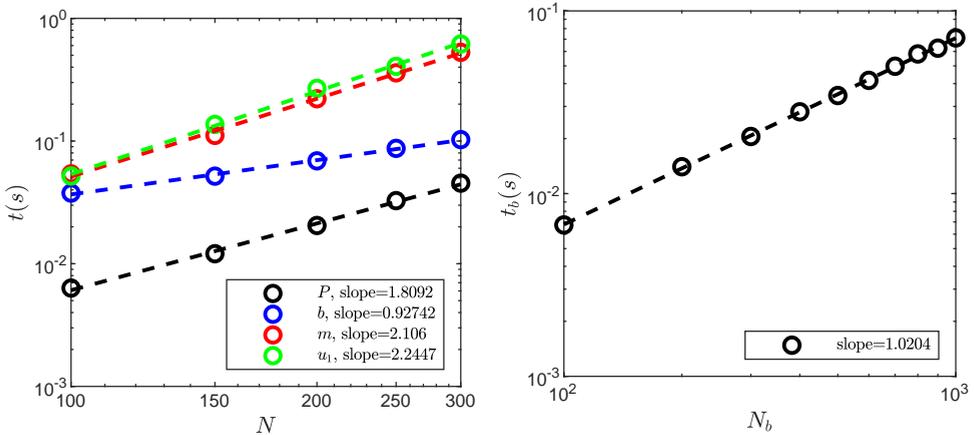


Figure 7.3: “Square-to-circle” problem: values of J_I and J_B as function of the iteration number for different values of N_b on a 200×200 grid, $\alpha = 0.2$.



(a) Calculation times for the minimization steps for P , b , m and the calculation of u_1 where we used $N_b = 1000$

(b) Calculation time for the minimization procedure for b as a function of N_b for a 200×200 grid

Figure 7.4: “Square-to-circle” problem: average calculation time per iteration (50 iterations total) as function of N and N_b . We calculate a c-concave solution u_1 and use parameter value $\alpha = 0.2$. The dashed lines are least-squares fits.

This result is independent of grid size. For large values of α , i.e., $\alpha = 0.8$, and small grid size 50×50 we see that the value of J increases before reaching the stopping criterion, which happens during the minimization procedure for P . When the mapping m has changed, the set $\mathcal{P}(m)$ has changed over which we minimize in (6.8b). Therefore, it may occur that

$$\min_{P \in \mathcal{P}(m^{n+1})} J_I[m^{n+1}, P] > \min_{P \in \mathcal{P}(m^n)} J_I[m^n, P]. \quad (7.8)$$

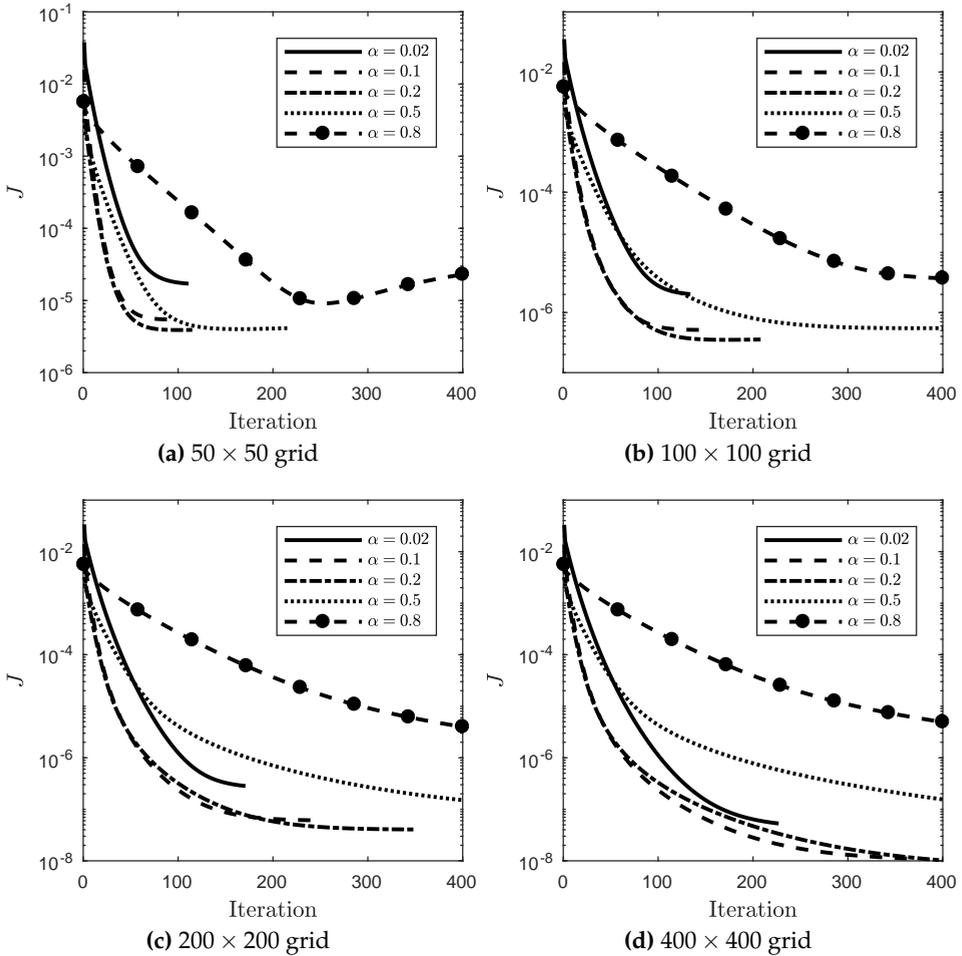


Figure 7.5: “Square-to-circle” problem: values of $J = (1 - \alpha) J_I + \alpha J_B$ as function of the iteration number for several grid sizes and different values of α with $N_b = 1000$. We use the stopping criterion in (7.7).

7.2 Parallel-to-far-field reflector: frog to prince

We now consider an example for the parallel-to-far-field system explained in Section 3.2. Recall that $c(\mathbf{x}, \mathbf{y}) = -\mathbf{x} \cdot \mathbf{y}$ as in (3.49) and consequently, $C = D_{xy}c = -I$. We use the GLS algorithm presented in Section 6.1. Since $C = -I$ the GLS algorithm reduces to the original least-squares algorithm designed for the standard Monge-Ampère equation [124].

We compute a freeform reflector surface that converts a nonuniform source emittance, corresponding to a picture of a frog, into a far-field target intensity corresponding to a picture of a prince. We consider a square source domain $\mathcal{X} = [-1, 1]^2$. For space discretization we cover the source domain \mathcal{X} with a uniform 800×800 grid and use 1000 points to discretize the boundary. The emittance $f(\mathbf{x})$ of the source is given by the grayscale values of the picture of a frog, shown in Figure 7.6a. The source emits a parallel bundle of light in the positive z -direction and the reflected rays are projected on a screen at distance $d = 10$, which we consider to be in the far field, parallel to the plane $x_2 = 0$. The desired illumination $L = L(\xi, \eta)$ [lm/m^2], with (ξ, η) the local Cartesian coordinates on the screen, is prescribed as the grayscale values of a picture of a prince, see Figure 7.6b.

The actual target distribution computed by the GLS algorithm is the far-field intensity $\tilde{g}(\mathbf{y})$, a deformation of the illuminance L ; for more details on the conversion from $L(\xi, \eta)$ to $\tilde{g}(\mathbf{y})$, see Section 3.1.4. Our reflector transforms the frog into a handsome prince, just like in the fairy tale of the Grimm brothers, and for that reason we dub this problem the ‘Frog-to-prince problem’. The pictures of the frog and prince have a colored original and the color-to-grayscale conversion creates black regions on the screen. Consequently, $\tilde{g}(\mathbf{y}) = 0$ for some $\mathbf{y} \in \mathcal{Y}$. To avoid division by 0 in the right-hand side of Equation (3.53) we locally increase the value of \tilde{g} to 15 % of its maximum value, when the local value is below this threshold. We apply the GLS method using the initial mapping in (6.10) to compute the optical map m . We have that $\text{tr}(\mathbf{P}^0) \leq 0$ and consequently compute a convex reflector surface $z = u(\mathbf{x})$, since c -convexity means regular convexity in this example.

The optical map, i.e., the image of the uniform source grid on the target domain, restricted to a coarsened version of the grid, is shown in Figure 7.7a. The contours of both the frog and prince are visible in the mapping. The corresponding reflector surface $z = u(\mathbf{x})$ is shown in Figure 7.7b. Clearly, the reflector surface is convex. The convergence history of J_I and J_B are given in Figure 7.7c.

To validate the solution, we have computed a ray-trace image. We use a self-programmed ray-tracing algorithm in Matlab. The image is created

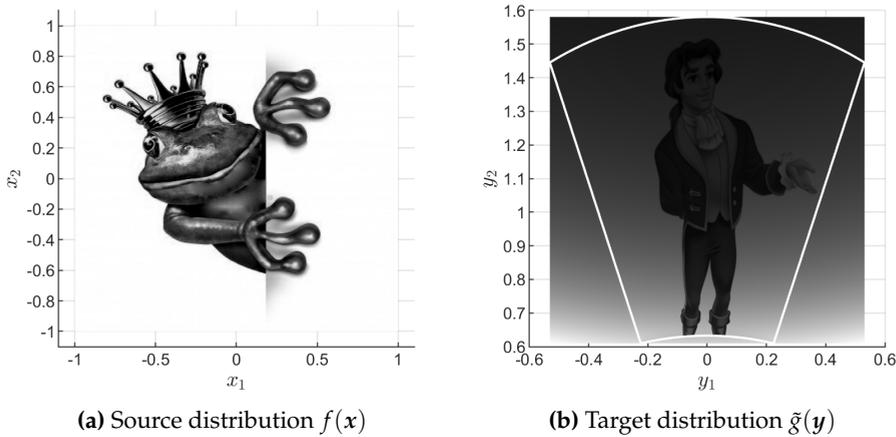
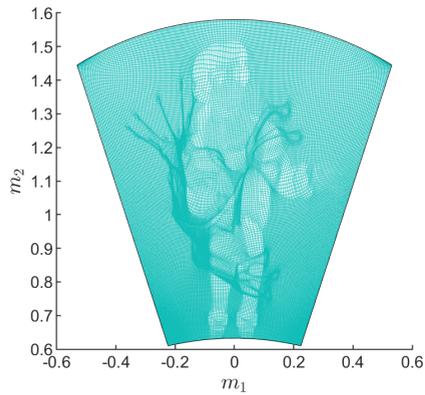
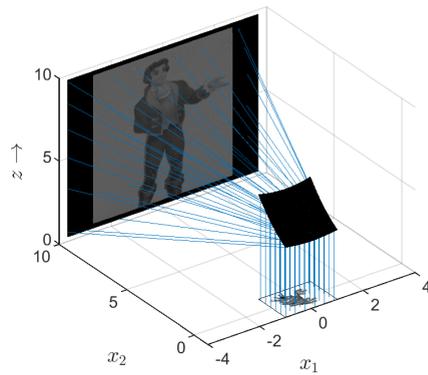


Figure 7.6: “Frog-to-prince” problem: the source and target distributions. Images used for the simulation: Frog ©Fotosearch, Prince ©The Walt Disney Company.

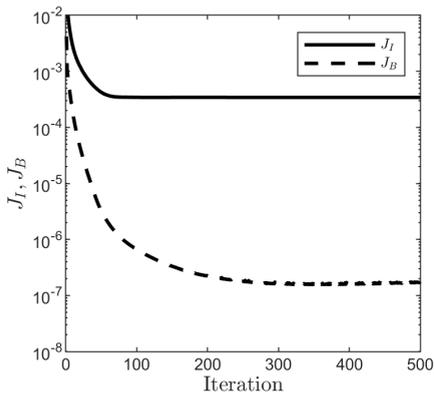
with 2.5 million rays from quasi-random positions (quasi-Monte Carlo) on the source domain. We triangulate the surface u , and for each ray we calculate the intersection with the surface and the corresponding normal vectors to the faces of the triangles using the Möller–Trumbore ray-triangle intersection algorithm [106]. Subsequently, we compute the directions of the light rays $\hat{\mathbf{f}}$ using the vectorial law of reflection and determine the corresponding bin on the target domain, divided in 500×500 bins. In the remainder of this chapter and in Chapter 9, we use this same approach to obtain ray-tracing results. The resulting target intensity is plotted in Figure 7.7d. Although slightly blurred, the ray-trace image closely resembles the original picture, even complex details are reproduced. We have also included the ray-trace result for an ‘intermediate’ reflector surface computed after only two iterations of the least squares algorithm for m in Figure 7.8b. We can still see the frog because the iteration has not yet converged.



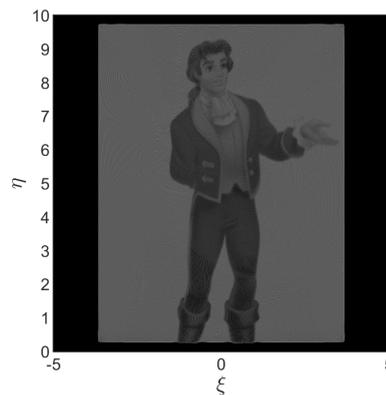
(a) Mapping



(b) Surface



(c) c-convex: 800×800 grid



(d) Ray-traced image

Figure 7.7: “Frog-to-prince” problem: convergence history for $N = 800$. We calculate a c-convex solution u_1 with parameter values $\alpha = 0.2$, $N_b = 1000$. The mapping is shown in (a) and the surface are shown in (b) with a number of rays traced. The error is shown in (c) and the ray-traced result in (d).

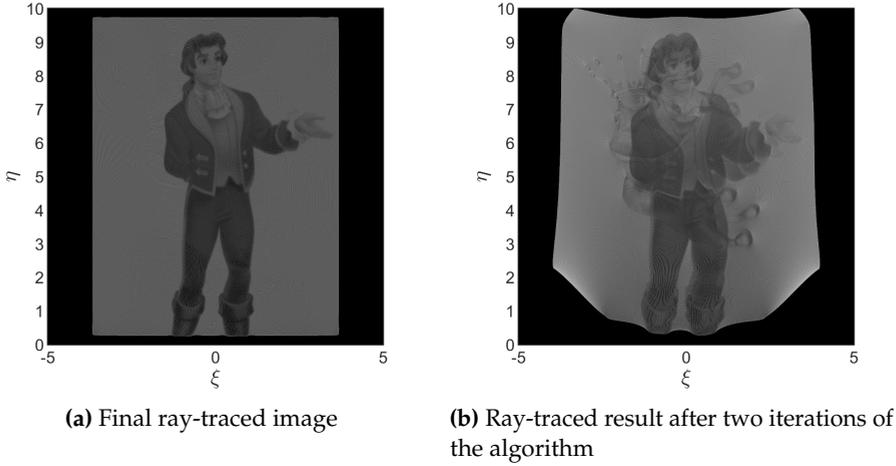


Figure 7.8: “Frog-to-prince” problem: ray-traced images (a) after 500 iterations and (b) after 2 iterations. In (b) the frog is still visible since the algorithm has not converged.

7.3 Point-to-far-field lens

We consider the point-to-far-field lens problem in Section 3.5 with a point source at O .

7.3.1 Peanut lens for road-lighting

An illumination system where LEDs and freeform optical components can be applied is road lighting. The LED light source is suspended high above the street. We approximate the LED light source as a point light source. We compute a lens surface that transforms the light from a point source into a far-field target intensity distribution corresponding to a typical roadlighting profile.

We consider a cone-shaped incoming bundle and a circular stereographic source domain $\mathcal{X} = \{\omega = (\rho, \zeta) \in \mathbb{R}^2 \mid 0 \leq \rho \leq 1, 0 \leq \zeta < 2\pi\}$ with a Gaussian light distribution

$$\tilde{f}(\omega) = \frac{10}{\pi} e^{-10\rho^2}. \quad (7.9)$$

The opening angle of the cone-shaped bundle is π rad = 180° and we take the refractive index of the lens to be $n = 1.5$.

The target intensity $g(\psi, \chi)$ is shown in Figure 7.10 on the left, with zenith $0 \leq \psi \leq \pi/2$ with respect to the negative z -axis and azimuth $0 \leq \chi < 2\pi$

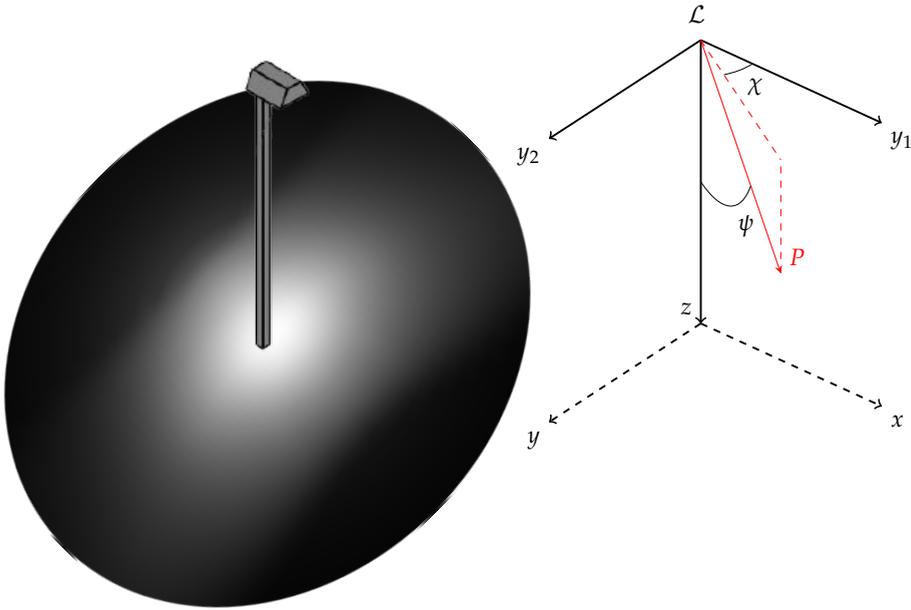


Figure 7.9: “Peanut-lens” problem: schematic representation indicating the position of the lamp, the spherical coordinate system of the target and the stereographic projection y on the plane $z = 0$.

extending into the far field on the street, as illustrated schematically in Figure 7.9.

Changing to stereographic coordinates using (3.8) (using a south pole and plus sign) transforms Figure 7.10a into Figure 7.10b.

We use the GLS algorithm in polar stereographic coordinates from Section 6.2 to compute the optical map m and the lens surface. We discretize the source domain using a 100×100 grid and use the initial mapping in (6.96) and compute a c -concave u_1 . The optical map, plotted using a coarsened version of the source grid, the lens surface, and convergence results are shown in Figure 7.11. The errors J_I and J_B oscillate slightly at the beginning of the simulation but stabilize after 35 iterations. The total computation time of performing 200 iterations to calculate m is 24.9 seconds. The subsequent computation time of u_1 is 0.7 seconds.

The left figure in Figure 7.12 shows the target intensity converted to the local Cartesian coordinates (x, y) on the street $z = 6$ for a lamp a distance 6m above the street at $z = 0$; see Figure 7.9 with the z -axis oriented in the downwards direction. The conversion from the target intensity $g(\psi, \chi)$ to local Cartesian coordinates is explained in detail in Chapter 3, Section 3.1.4. The right figure shows the ray-trace results of our peanut lens, using the algorithm

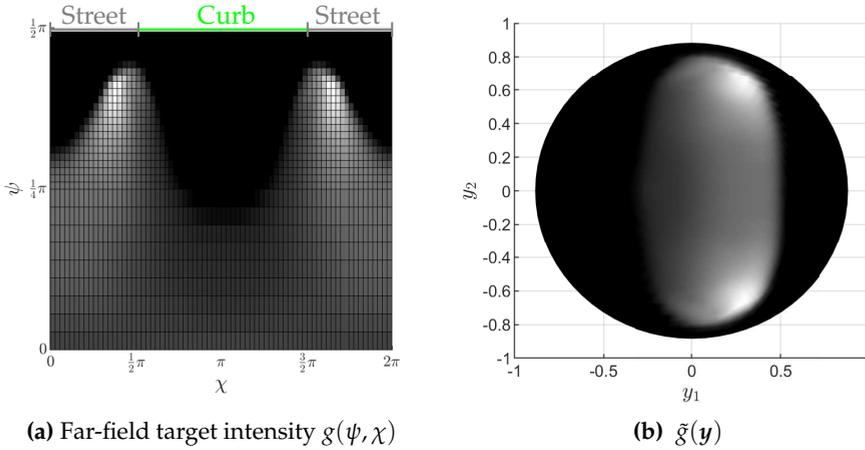


Figure 7.10: “Peanut-lens” problem: a road light on the street.

explained in Section 7.2 with 1 million rays and 200×200 bins. The target on the street, which we plotted up to 15m from the lamp in the x - and y -directions, is divided into a uniformly spaced grid. The white dot indicates the position of the lamp and the black line marks the separation between the curb (left) and the street (right). The ray-trace results match the desired profile on the street, with more light directed onto the street. To measure the deviation between the original image and the ray-trace result we take the difference between the target intensity on the street (interpolated bilinearly onto the ray-tracing grid) and the ray-tracing irradiance. The mean absolute difference in intensity is 7% and the maximum absolute difference is 15% (with the maximum value of the target intensity interpreted as the 100% value). Additionally, we can verify that condition (2.53) is not satisfied for all rays traced, i.e., TIR does not occur.

7.3.2 An ellipsoidal lens comparison

As in the previous section, we consider the point-to-far-field lens problem, explained in Section 3.5. For this system we can formulate a cost function in optimal transport theory, cf. (3.113), and a generating function, cf. (3.114). In this section we compare the performance of the GLS procedure in Section 6.1 and the GJLS procedure in Section 6.3.

We compare our two algorithms using exact solutions for the mapping and lens surface in a similar way as in Section 7.1.1. We consider an ellipsoidal lens surface

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1, \quad (7.10)$$

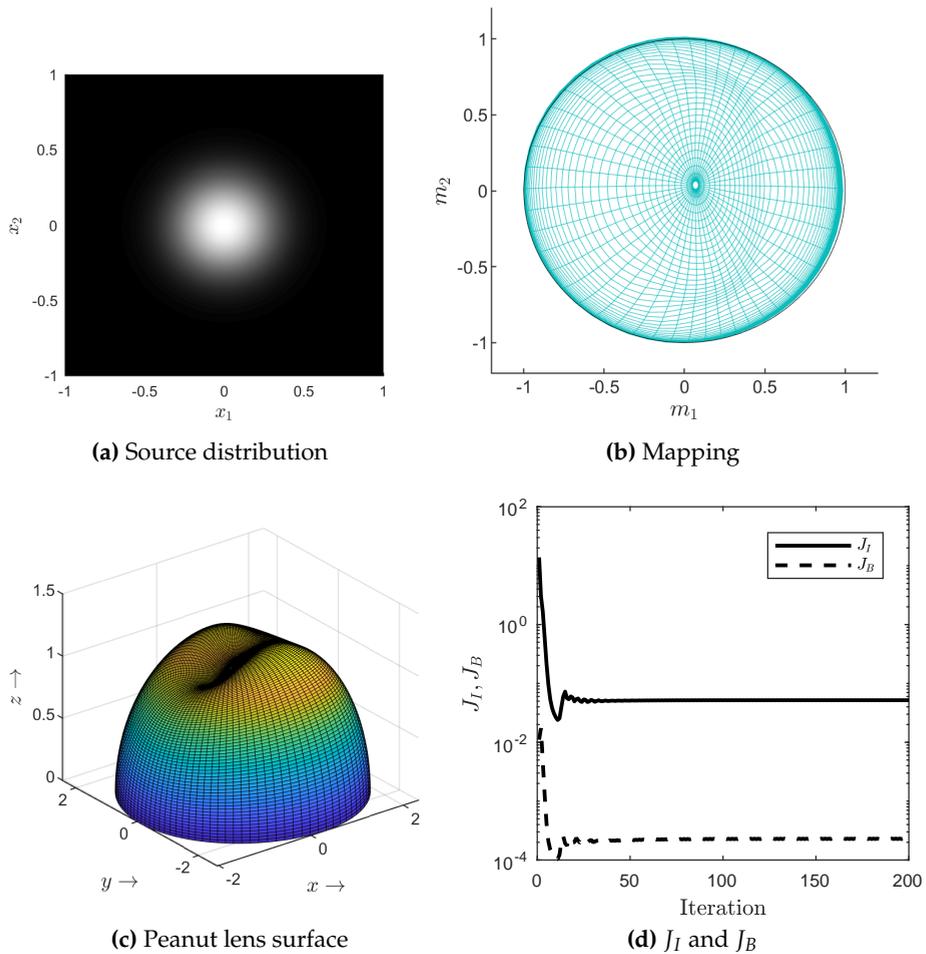


Figure 7.11: “Peanut-lens” problem: (a) the source distribution in stereographic coordinates, (b) the optical mapping, (c) lens surface, and (d) J_I and J_B .

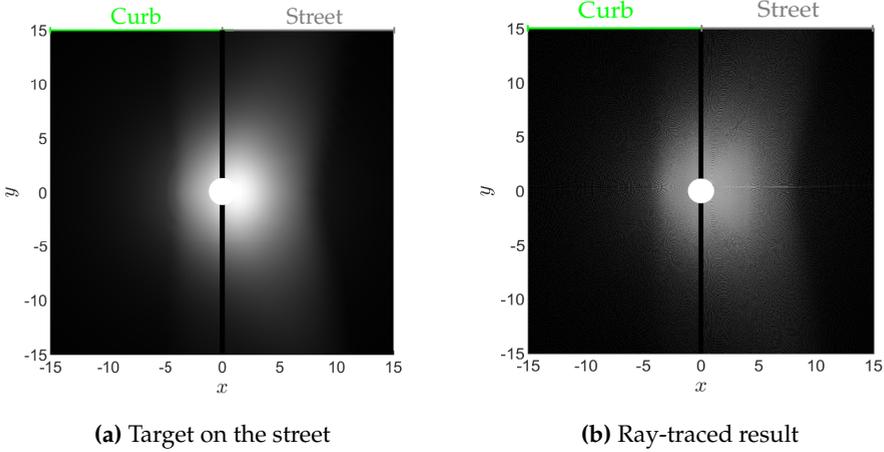


Figure 7.12: “Peanut-lens” problem: the target intensity on the street and ray-traced image of the peanut lens. The white dot in the center indicates the position of the lamp and the black line marks the division between the curb (left) and the street (right).

where $z > 0$ and with given constants $a \neq 0$, $b \neq 0$, $c \neq 0$. By substituting $x = u s_1$, $y = u s_2$, $z = u s_3$, and changing to stereographic coordinates using (3.9) (using a south pole and plus sign), we can derive an expression for the surface u as

$$u(\mathbf{x}) = \frac{1 + |\mathbf{x}|^2}{2 \sqrt{\left(\frac{x_1}{a}\right)^2 + \left(\frac{x_2}{b}\right)^2 + \left(\frac{-1 + |\mathbf{x}|^2}{2c}\right)^2}}. \quad (7.11)$$

Subsequently calculating ∇u , and using the implicit relation for the mapping in (6.3) we can solve for the mapping $\mathbf{m}(\mathbf{x})$. (The solution is a few pages long and not included in this thesis for brevity.) We consider a square source domain $\mathcal{X} = [-0.5, 0.5]^2$ and choose $a = 2$, $b = 1$, and $c = 1$. Using the mapping $\mathbf{m}(\mathbf{x})$ we compute the target domain \mathcal{Y} and the corresponding target boundary. We set the right-hand side $F(\mathbf{x}, \mathbf{m}(\mathbf{x}), u(\mathbf{x})) = F(\mathbf{x}, \mathbf{m}(\mathbf{x}))$ using the expression in (6.105), since the solution for the mapping \mathbf{m} is only an exact solution for a source intensity $\tilde{f}(\mathbf{x})$ and target intensity $\tilde{g}(\mathbf{y})$ such that (6.105) is satisfied. For both methods, we use the initial mapping in (6.11). For the generating-function approach we initialize the surface u to a spherical surface (initializing u_1 is not necessary for the cost-function approach). We have that $G_w = n - \hat{\mathbf{s}} \cdot \hat{\mathbf{t}} > 0$, so we either consider a max/min conjugate pair or min/max conjugate pair, as explained in Section 4.4.2. With the initial mapping we obtain that $\text{tr}(\mathbf{P}^0) \geq 0$ in the GLS algorithm and $\text{tr}(\mathbf{P}^0) \leq 0$ in the GJLS algorithm so we compute a c-concave u_1 and G-convex u , as explained

in Section 4.3.1 and 4.4.2, respectively.

To investigate convergence of the numerical algorithm, we use the norms defined in (7.5). Figure 7.13 shows J_I and J_B for $N = 100$. It also displays the changes in $C D\mathbf{m}$, $\mathbf{m}|_{\partial\mathcal{X}}$, \mathbf{P} and \mathbf{b} . We take the number of boundary points required for minimization step (6.111a) to be $N_b = 100$ and weighting parameter $\alpha = 0.2$. The functionals J_I and J_B reach a plateau at a certain iteration number while the updates in $C D\mathbf{m}$, $\mathbf{m}|_{\partial\mathcal{X}}$ (\mathbf{m} on the boundary), \mathbf{P} and \mathbf{b} continue to decrease up to machine precision. We use the stopping criterion in (7.7). Table 7.2 shows the results for several $N \times N$ grids with logarithmic least-squares fits. The c-concave, G-convex and exact surfaces for $N = 100$ are plotted in Figure 7.13b. The surfaces largely overlap and deviate slightly from the exact solution. The number of iterations required increases sublinearly with N . We see that J_I and J_B have approximately third- to fourth-order convergence. Using Simpson's rule we calculated the integral given in the compatibility condition in (6.122). The last row of Table 7.2 shows nearly second-order convergence of the compatibility condition.

Figure 7.14 shows the maximum absolute differences between the computed mapping and surface with the exact solution (where we transformed u_1 back to u for the c-concave solution). We observe almost second-order convergence for both methods, but the G-convex solution is closer to the exact solution for all grid sizes.

Figure 7.15 shows the calculation times of the minimization procedures for \mathbf{P} , \mathbf{b} , \mathbf{m} , and the computations of u_1 and u as a function of $N = N_1 = N_2$. The slopes of performed logarithmic least-squares fits are also displayed. The calculation time for the minimization procedure for \mathbf{P} (linear in N) is better than expected, since it is sublinear in the number of grid points. The calculation times for the minimization procedures for \mathbf{b} (linear in N), and \mathbf{m} (quadratic in N) are as expected. The calculation time of u_1 or u should be at least linear in the number of grid points and thus quadratic in N . For the c-concave solution, u_1 is only computed once at the end of the iterative procedure. For the G-convex solution, u is recomputed at each iteration by solving the Neumann problem (6.119). The computation time shows approximately a quadratic growth in N .

For a c-concave solution, the total calculation time for one iteration is approximately proportional to $N^{1.7}$, taking the steepest slope in Figure 7.15a. With the number of iterations growing sublinearly in N , see Table 7.2, the total calculation time scales roughly with N^2 , as displayed in the second row for the c-concave results of Table 7.2. For a G-convex solution, the total calculation time for one iteration is approximately proportional to $N^{1.8}$, taking the steepest slope in Figure 7.15b, and with the number of iterations scaling as $N^{0.4}$, see

Table 7.2, the total calculation time scales roughly with N^2 as well, as shown in the second row of the G-convex results of Table 7.2.

In summary, the G-convex solution is more accurate than the c-concave solution, requires fewer iterations, and has approximately the same computation time.

c-concave

Grids	20 × 20	40 × 40	60 × 60	80 × 80	100 × 100	Fits
Iterations	211	430	593	720	824	$\propto N^{0.8}$
Time [s]	2	6	16	35	61	$\propto N^{2.2}$
J_I	7.5×10^{-4}	5.1×10^{-5}	9.9×10^{-6}	3.1×10^{-6}	1.3×10^{-6}	$\propto N^{-4.0}$
J_B	3.0×10^{-4}	4.3×10^{-5}	1.1×10^{-5}	3.7×10^{-6}	1.6×10^{-6}	$\propto N^{-3.3}$

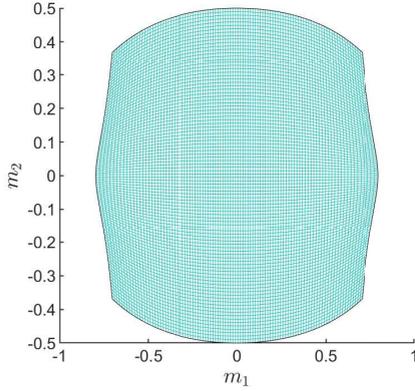
G-convex

Iterations	118	172	200	221	242	$\propto N^{0.4}$
Time [s]	1	3	7	15	28	$\propto N^{1.8}$
J_I	1.7×10^{-4}	1.3×10^{-5}	2.7×10^{-6}	8.9×10^{-7}	3.7×10^{-7}	$\propto N^{-3.8}$
J_B	2.7×10^{-5}	2.8×10^{-6}	6.7×10^{-7}	2.4×10^{-7}	1.1×10^{-7}	$\propto N^{-3.5}$
Compatibility	9.8×10^{-5}	2.8×10^{-5}	1.4×10^{-5}	8.1×10^{-6}	5.4×10^{-6}	$\propto N^{-1.8}$

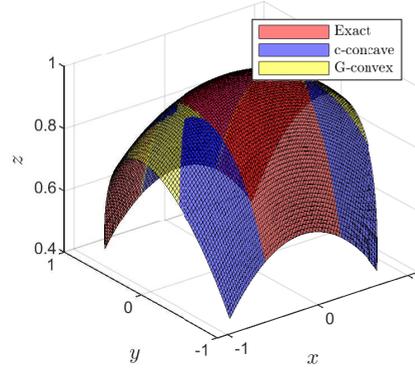
Table 7.2: “Ellipsoidal-lens” problem: number of iterations, total computation time (in seconds) and residuals in the GLS and GJLS algorithms.

7.3.3 Reduction in surface calculations

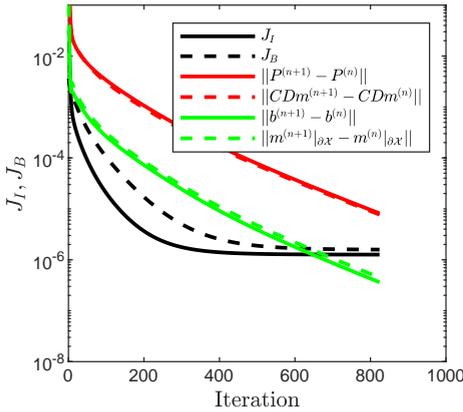
As we increase the number of grid points N the computation time of u , i.e., the new minimization step (6.111d), grows most steeply in Figure 7.15b. Since this step could become problematic with an increasing number of grid points, we investigate whether we can reduce the number of updates of the surface u . We rerun the experiment of the previous section for a G-convex pair considering a grid of $N \times N = 100 \times 100$, but we do not perform minimization step (6.111d) at every iteration. We increase the period T_u of updating u from every one to every 100 iterations. Figure 7.16 shows J_I , J_B and the changes in $\mathbf{C} D\mathbf{m}$, $\mathbf{m}|_{\partial\mathcal{X}}$ (\mathbf{m} on the boundary), \mathbf{P} and \mathbf{b} for two example periods. The functionals J_I and J_B plateau to approximately the same value. The values of J_I and J_B temporarily increase immediately after the iteration when u is updated. In Table 7.3 we see that the final values of J_I and J_B remain roughly constant



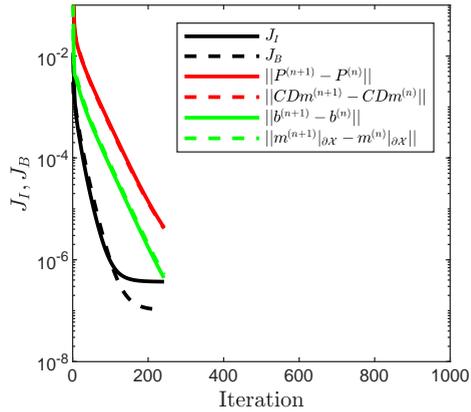
(a) Ellipsoidal mapping



(b) Ellipsoidal surfaces



(c) c-concave: 100×100 grid



(d) G-convex: 100×100 grid

Figure 7.13: “Ellipsoidal-lens” problem: convergence history for $N = 100$ for both methods. We calculate a c-concave solution u_1 and G-convex solution u , respectively, with parameter values $\alpha = 0.2$, $N_b = 100$. The mapping of the G-convex solution is shown in (a) and the c-concave and G-convex surfaces are shown in (b) together with the exact solution. (c) and (d) show J_I and J_B for the c-concave and G-convex solution, respectively, for $N = 100$ with the updates in CDm , $m|_{\partial\chi}$ (m on the boundary), P and b .

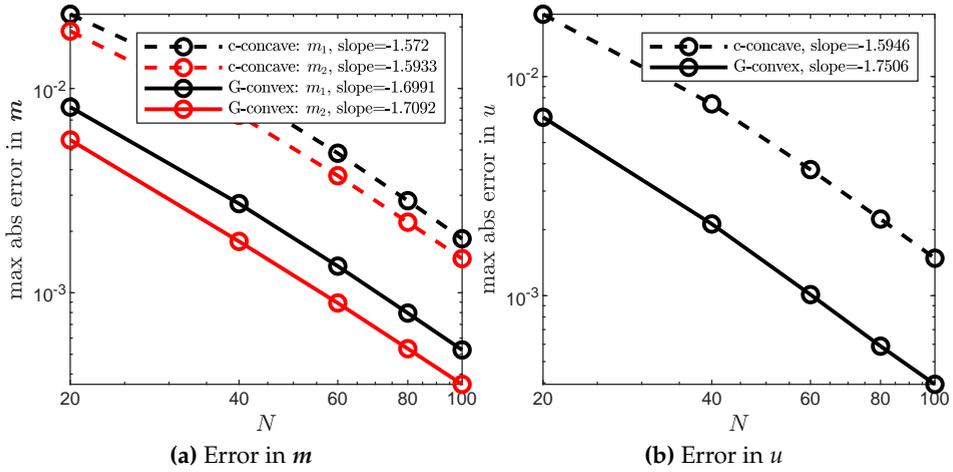


Figure 7.14: “Ellipsoidal-lens” problem: (a) maximum absolute differences between the components of the final mapping $m = (m_1, m_2)$ and the exact mapping. (b) Maximum absolute difference between the final surface u and the exact solution.

and the number of iterations increases slightly. The total computation time decreases significantly when increasing T_u from 1 to 20, but further extending the period to 50 and 100 iterations increases the total computation time again. This is due to an increase in the number of iterations in order to reach the stopping criterion. For all periods, the maximum absolute difference between the computed mapping and the exact solution is approximately 5.2×10^{-4} for m_1 and 3.6×10^{-4} for m_2 . The maximum absolute difference between u with the exact solution is approximately 4.0×10^{-4} for all periods. The lowest computation time of 18 seconds is reached at $T_u = 20$, lower than for the c-concave and G-convex solutions in Table 7.2 (61 and 28 seconds), while maintaining lower values for J_I and J_B and a higher solution accuracy than the c-concave solution. Hence, we can reduce the computation time by updating the surface u less frequently.

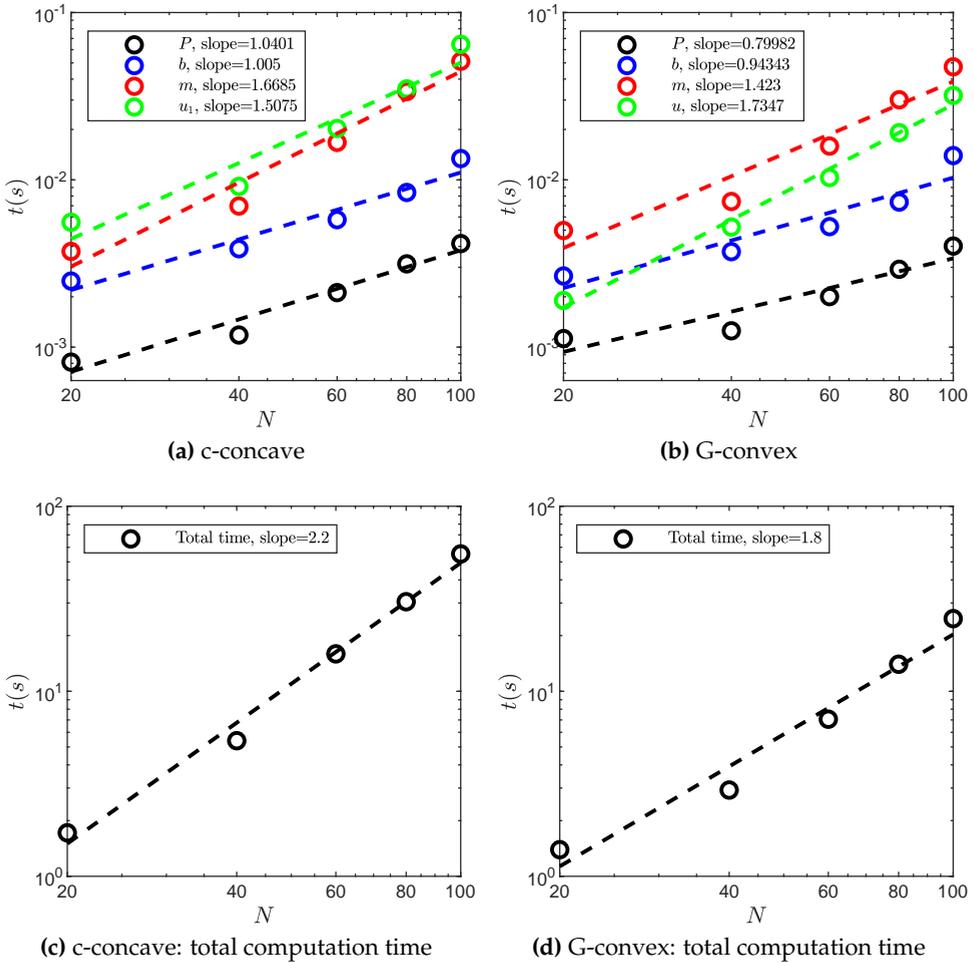


Figure 7.15: “Ellipsoidal-lens” problem: (a,b) average calculation time per iteration as a function of N for the minimization steps for P , b , m and the calculation of u_1 for the c-concave method and u for the G-convex method. (c,d) The total computation time, cf. Table 7.2. The dashed lines are least-squares fits.

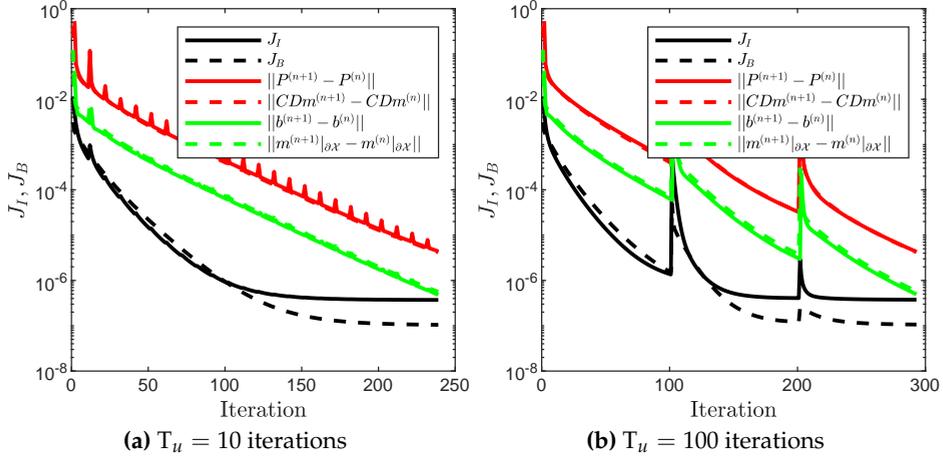


Figure 7.16: “Ellipsoidal-lens” problem: convergence history for $N = 100$. We calculate a G-convex solution u with parameter values $\alpha = 0.2$, $N_b = 100$. The surface calculation period T_u is increased from 1 to 100 iterations.

G-convex

T_u	1	5	10	20	50	100
Iterations	242	240	239	239	257	293
Time [s]	27	20	19	18	21	22
J_I	3.7×10^{-7}	3.8×10^{-7}				
J_B	1.1×10^{-7}					
Compatibility	5.4×10^{-6}	5.2×10^{-6}	5.0×10^{-6}	4.2×10^{-6}	5.1×10^{-6}	1.6×10^{-4}

Table 7.3: “Ellipsoidal-lens” problem: number of iterations, total computation time (in seconds) and residuals in the GJLS algorithm, with a grid size of $N \times N = 100 \times 100$ and increasing the period T_u .

7.4 Parallel-to-near-field reflector

We consider the system in Section 3.3. We compute a freeform reflector surface that converts the light from a parallel incoming beam into a near-field target illuminance distribution corresponding to a picture.

We consider the square source domain $\mathcal{X} = [-1, 1]^2$ and a uniform light distribution $f(\mathbf{x}) = 1/4$. The reflected rays are projected on a screen in the near field, parallel to the source plane. The required illumination $g(\mathbf{y})$ is derived from the grayscale values of a photo of my supervisor Jan ten Thije Boonkkamp.

The grayscale values of the picture prescribe the illuminance. Analogously to Section 7.2, we increase values of $g(\mathbf{y})$ which are below a threshold of 15% of its maximum value to this threshold. We use the GJLS algorithm to compute the optical map \mathbf{m} and the reflector surface. We calculate u every 20th iteration ($T_u = 20$) and use a 500×500 grid. We use the initial guess \mathbf{m}^0 given in (6.11) and find $\text{tr}(\mathbf{P}^0) \leq 0$, so we compute a G-convex u . The optical map, plotted using a coarsened version of the source grid, the reflector surface, and convergence results are shown in Figure 7.17.

Subsequently, we validated the resulting reflector image using our ray-tracing algorithm explained in Section 7.2. We traced 2.5 million rays with quasi-random positions (quasi-Monte Carlo) from source to near field and 500×500 bins. The resulting target illuminance $g(\mathbf{y})$ is plotted in Figure 7.17c.

The average computation time per iteration is 2.4 seconds (2.2 seconds without computation of u and 5.0 seconds with computation of u). The total computation time is 12 minutes. Using Simpson's rule, the compatibility integral in (6.122) evaluates to -9.3×10^{-6} . The ray-trace image closely resembles Jan.

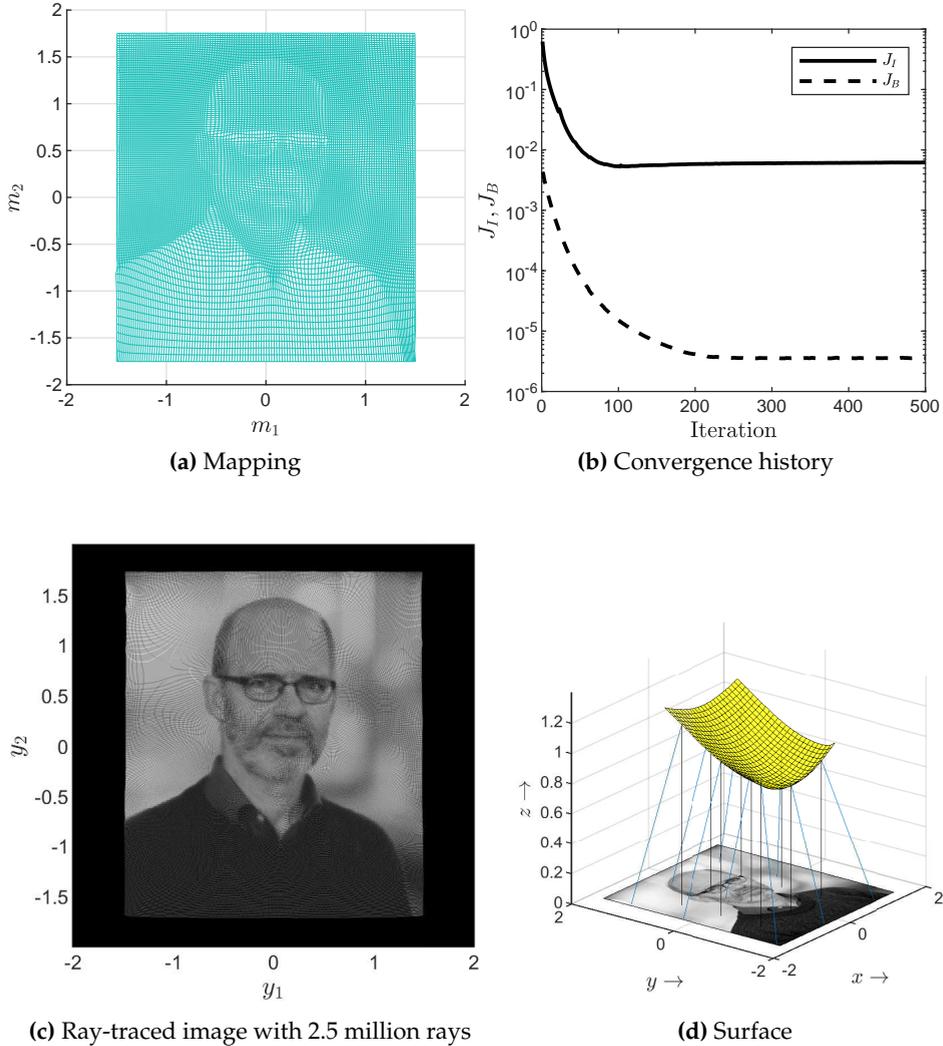


Figure 7.17: “Jan” problem: a picture of my supervisor in the near field, using $N = 500$. We calculate a G-convex solution u with parameter values $\alpha = 0.2$, $N_b = 4$. The mapping of the G-convex solution is shown in (a) and the convergence history in (b). Figure (c) shows the resulting ray-traced image and the G-convex surface is shown in (d) together with a subset of rays traced.

7.5 Summary

We showed a variety of numerical examples to illustrate the performance of our numerical algorithms. We included details of the convergence of the algorithms and tested the algorithms on challenging test problems with sources and targets corresponding to pictures. For an ellipsoidal lens problem, we compared the results from the GLS algorithm and the GJLS algorithm. Both procedures gave similar convergence results, with the GJLS algorithm as the most accurate. However, the GJLS approach took up more computation time since it requires an extra step in the iterative procedure to compute u . We subsequently reduced the computation time by not updating the surface u at every iteration, while still maintaining high solution accuracy.

Chapter 8

A Double Freeform Lens

In this chapter, we consider an optical system outside of the base cases: a lens with two freeform surfaces with a point source and far-field target, as illustrated schematically in Figure 8.1. For this combination of source and target only one optical surface suffices, as we have seen in Chapter 3, but we will show that using two surfaces gives more freedom in the design of the system.

For the double freeform systems in Chapter 3 we considered parallel and point sources *and* parallel and point targets. For these systems the optical path length for all rays is equal to the same constant, i.e., one of Hamilton's characteristic functions becomes a constant. The second surface can be directly computed when the mapping and the first surface are known, or vice versa.

When we consider a far-field target for a double freeform system, we cannot directly compute the second freeform surface from the first one. In this case, the target wavefront is neither collimated nor spherical, but has a general shape and none of Hamilton's characteristics become constant. In this chapter, we will show that for such a double freeform lens we can find two generating functions, one for each optical surface. We have an extra degree of freedom in the system by using an *intermediate target intensity* to compute the first freeform surface.

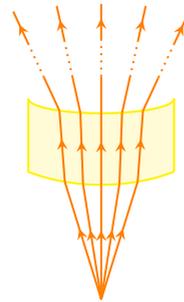


Figure 8.1: Double freeform lens with a point source and far-field target changing the direction of the rays at the first and the second optical surface.

If we concatenate more than two optical surfaces behind each other, we could in principle create a large number of such free parameters. We will show that we can add Hamilton's characteristic functions corresponding to each surface to get the total characteristic functions of the whole optical system. This chapter is one step towards computing multiple freeform surfaces for more complicated optical systems.

8.1 Mathematical formulation

In this section, we describe the double freeform lens mathematically in a two-step method. First, we consider the first freeform surface and use Hamilton's characteristic functions to derive a generating function. Second, we perform similar steps for the second freeform surface. Before we move to Hamilton's characteristic functions, we first explain the notation and introduce a new set of stereographic coordinates.

Figure 8.2 schematically illustrates a point source, a lens with refractive index n and a far-field target. A beam of light emanates from the point source located at the origin O of the Cartesian coordinate system. The point source emits rays of light travelling radially outward in the direction $\hat{s} = \hat{e}_r$, where \hat{e}_r is the radial basis vector in the spherical coordinate system.

The first surface of the lens is described by $\mathcal{L}_1 : r_1(\phi, \theta) = u(\phi, \theta) \hat{e}_r$, where $u(\phi, \theta) > 0$ is the radial parameter that describes the location of the surface, $0 \leq \phi \leq \pi$ is the zenith and $0 \leq \theta < 2\pi$ is the azimuth in the spherical coordinate system. The surface \mathcal{L}_1 refracts the ray \hat{s} in direction \hat{i} .

The second surface is given by $\mathcal{L}_2 : r_2(\phi, \theta) = u(\phi, \theta) \hat{e}_r + v(\phi, \theta) \hat{i}$, where $v(\phi, \theta) > 0$ is the parameter that describes the location of the surface relative to the first surface and is equal to the distance $d(P_1, P_2)$ between the point $P_1(u(\phi, \theta) \hat{e}_r)$, where the ray intersects the first surface, and the point P_2 , where the ray intersects the second surface. The surface \mathcal{L}_2 refracts the ray \hat{i} in direction \hat{t} .

The intensity of the source is given by $f(\phi, \theta)$ [lm/sr], and the required target intensity in the far field is denoted by $g(\psi_2, \chi_2)$ [lm/sr], where (ψ_2, χ_2) represents a different set of spherical coordinates, with zenith $0 \leq \psi_2 \leq \pi$ and azimuth $0 \leq \chi_2 < 2\pi$. The origin of the coordinate system describing the target is the lens approximated as a point in space (i.e., the far-field approximation). Note that we use the subscript 2 to indicate the *final* target intensity after the rays have propagated through both freeform surfaces. Later in this section, we also define an *intermediate far-field target intensity* $h(\psi_1, \chi_1)$, with zenith $0 \leq \psi_1 \leq \pi$ and azimuth $0 \leq \chi_1 < 2\pi$, which we will prescribe as the target after the rays have only propagated through the first freeform surface.

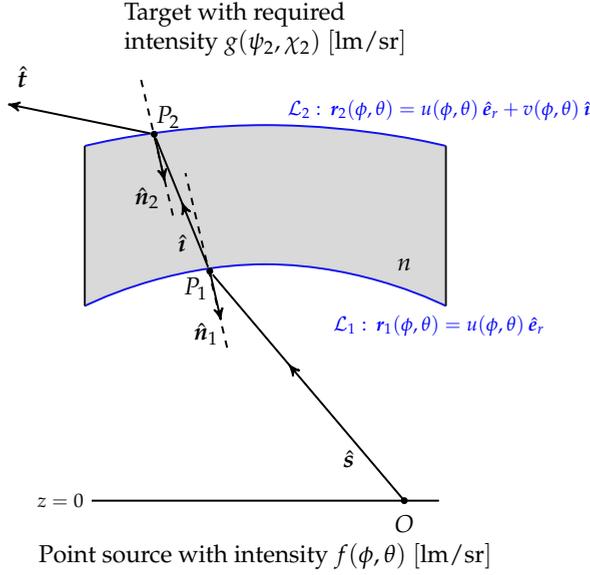


Figure 8.2: Double freeform lens converting the intensity $f(\phi, \theta)$ of a point source into a far-field target intensity $g(\psi_2, \chi_2)$.

We transform the coordinates of the light rays from spherical to stereographic. This is convenient since the vectors $\hat{s} = (s_1, s_2, s_3)$, $\hat{i} = (i_1, i_2, i_3)$ and $\hat{t} = (t_1, t_2, t_3)$ are defined on the unit sphere S^2 . Hence, $|\hat{s}| = |\hat{i}| = |\hat{t}| = 1$. We define

$$\mathbf{x}(\hat{s}) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{1 + s_3} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \frac{1}{1 + \cos(\phi)} \begin{pmatrix} \sin(\phi) \cos(\theta) \\ \sin(\phi) \sin(\theta) \end{pmatrix}, \quad (8.1a)$$

$$\mathbf{y}_1(\hat{i}) = \begin{pmatrix} y_{11} \\ y_{12} \end{pmatrix} = \frac{1}{1 + i_3} \begin{pmatrix} i_1 \\ i_2 \end{pmatrix} = \frac{1}{1 + \cos(\psi_1)} \begin{pmatrix} \sin(\psi_1) \cos(\chi_1) \\ \sin(\psi_1) \sin(\chi_1) \end{pmatrix}, \quad (8.1b)$$

$$\mathbf{y}_2(\hat{t}) = \begin{pmatrix} y_{21} \\ y_{22} \end{pmatrix} = \frac{1}{1 + t_3} \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = \frac{1}{1 + \cos(\psi_2)} \begin{pmatrix} \sin(\psi_2) \cos(\chi_2) \\ \sin(\psi_2) \sin(\chi_2) \end{pmatrix}, \quad (8.1c)$$

with corresponding inverse projections

$$\hat{\mathbf{s}}(\mathbf{x}) = \hat{\mathbf{e}}_r = \frac{1}{1 + |\mathbf{x}|^2} \begin{pmatrix} 2x_1 \\ 2x_2 \\ 1 - |\mathbf{x}|^2 \end{pmatrix}, \quad (8.2a)$$

$$\hat{\mathbf{i}}(\mathbf{y}_1) = \frac{1}{1 + |\mathbf{y}_1|^2} \begin{pmatrix} 2y_{11} \\ 2y_{12} \\ 1 - |\mathbf{y}_1|^2 \end{pmatrix}, \quad \hat{\mathbf{t}}(\mathbf{y}_2) = \frac{1}{1 + |\mathbf{y}_2|^2} \begin{pmatrix} 2y_{21} \\ 2y_{22} \\ 1 - |\mathbf{y}_2|^2 \end{pmatrix}. \quad (8.2b)$$

We represent the incoming rays $\hat{\mathbf{s}}$, the intermediate rays $\hat{\mathbf{i}}$ and the outgoing rays $\hat{\mathbf{t}}$ using stereographic projections from the south pole $(0, 0, -1)$ of S^2 onto the plane $z = 0$, as drawn schematically in Figure 3.4a. Hence, we assume that the lens does not refract the light rays in the negative z -direction. The stereographic projections in (8.1) are undefined at the south pole, and we consider $s_3, t_3, t_3 \neq -1$ and $0 \leq \phi, \psi_{1,2} < \pi$. Hence, we assume that the light rays are directed mainly in the upward z -direction.

We define our source domain \mathcal{X} as the support of $\tilde{f}(\mathbf{x}) = f(\phi(\mathbf{x}), \theta(\mathbf{x}))$, and our target domain \mathcal{Y}_2 as the image under the mapping \mathbf{m} , i.e., $\mathcal{Y}_2 = \mathbf{m}(\mathcal{X})$, and we introduce $\tilde{g}(\mathbf{y}_2) = g(\psi_2(\mathbf{y}_2), \chi_2(\mathbf{y}_2))$. We refer to $\mathbf{m} : \mathcal{X} \rightarrow \mathcal{Y}_2$ as the composite map $\mathbf{y}_2 = \mathbf{m}(\mathbf{x})$ from the source set of stereographic coordinates \mathcal{X} to the target set of stereographic coordinates \mathcal{Y}_2 . It is a composition of two mappings and can be written as $\mathbf{m} = \mathbf{m}_2 \circ \mathbf{m}_1$, where we refer to $\mathbf{m}_1 : \mathcal{X} \rightarrow \mathcal{Y}_1$ as the intermediate map $\mathbf{y}_1 = \mathbf{m}_1(\mathbf{x})$ from \mathcal{X} to the image under the mapping \mathbf{m}_1 , i.e., from the source set of stereographic coordinates \mathcal{X} to the (intermediate) target set of stereographic coordinates \mathcal{Y}_1 such that $\mathcal{Y}_1 = \mathbf{m}_1(\mathcal{X})$, and we refer to $\mathbf{m}_2 : \mathcal{Y}_1 \rightarrow \mathcal{Y}_2$ as the successive map $\mathbf{y}_2 = \mathbf{m}_2(\mathbf{y}_1)$ from \mathcal{Y}_1 to the image under the mapping \mathbf{m}_2 . Hence, we have $\mathcal{Y}_2 = \mathbf{m}_2(\mathcal{Y}_1) = (\mathbf{m}_2 \circ \mathbf{m}_1)(\mathcal{X})$.

We choose an intermediate target intensity $\tilde{h}(\mathbf{y}_1) = h(\psi_1(\mathbf{y}_1), \chi_1(\mathbf{y}_1))$, as an intermediate far-field intensity. Choosing a different intermediate target intensity changes the shape of both freeform optical surfaces. The added degree of freedom for the double lens design, in comparison to the systems outlined in Section 3.7, lies in the choice of an intermediate target distribution. In the numerical experiments in Chapter 9, we will show that we can choose the intermediate target intensity as, for instance, an interpolation between the source and final target intensity based on an initial approximate mapping from source to target.

In the next two sections, we use Hamilton's characteristic functions to find the generating functions. In Section 8.1.1 we focus on the first surface and in Section 8.1.2 we attend to the second surface.

We will not show the mappings explicitly for both optical surfaces. The

expression for the first surface is very long and the derivation extremely tedious. Hence, for the second freeform surface we will also not derive the mapping explicitly.

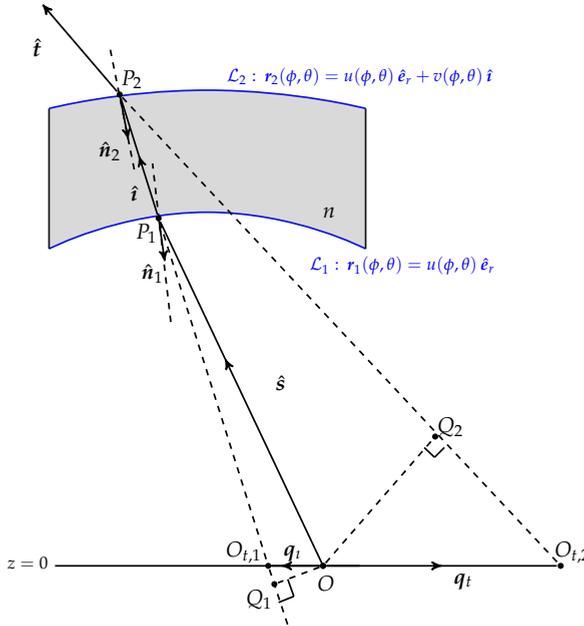


Figure 8.3: Illustration of the derivation of Hamilton's angular characteristic T for a double freeform lens.

8.1.1 The first freeform surface

Figure 8.3 illustrates an incident ray propagating in the direction \hat{s} intercepting the freeform lens surface \mathcal{L}_1 at point P_1 and refracting in the direction \hat{i} .

The position and direction coordinates of the ray \hat{s} at $z = 0$ are given by the two-vectors $\mathbf{q}_s = \mathbf{0}$ and $\mathbf{p}_s = (s_1, s_2)$, respectively. The position and direction coordinates of the ray \hat{i} at $z = 0$ are given by \mathbf{q}_i and $\mathbf{p}_i = (n_{i1}, n_{i2})$, respectively. Note that the point P_1 is given by $u(\hat{s}) \hat{s} = (u(\hat{s}) \mathbf{p}_s, u(\hat{s}) s_3)$.

In the following, we find Hamilton's angular characteristic function T from O to $O_{t,1}$ in the same way as for the point-to-far-field lens in Section 3.5. For the outgoing ray we have $\mathbf{p}_i = (n_{i1}, n_{i2})$ as opposed to $\mathbf{p}_t = (t_1, t_2)$ for the single freeform surface.

The angular characteristic $T_1(\mathbf{p}_i)$

The point characteristic V_1 between point $O(\mathbf{q}_s, 0)$ on the source plane and $O_{t,1}(\mathbf{q}_i, 0)$ (virtual image of the first surface) is given by

$$\begin{aligned} V_1(\mathbf{q}_s, \mathbf{q}_i) &= u(\hat{\mathbf{s}}) - n d(P_1, O_{t,1}), \\ d(P_1, O_{t,1}) &= \sqrt{|\mathbf{q}_i - u(\hat{\mathbf{s}}) \mathbf{p}_s|^2 + (u(\hat{\mathbf{s}}) s_3)^2}, \end{aligned} \quad (8.3)$$

where $n d(P_1, O_{t,1})$ denotes the optical path length between P_1 and $O_{t,1}$, which is equal to the Euclidean distance multiplied by n .

In order to find a generating function, we introduce Hamilton's angular characteristic $T_1(\mathbf{p}_s, \mathbf{p}_i)$ as the optical path length from O to Q_1 in Figure 8.3, and show it is independent of \mathbf{p}_s . Hamilton's angular characteristic for the first surface, which depends on the direction coordinate of the source ray and the intermediate ray, is given by

$$T_1(\mathbf{p}_s, \mathbf{p}_i) = V_1(\mathbf{q}_s, \mathbf{q}_i) + \mathbf{q}_s \cdot \mathbf{p}_s - \mathbf{q}_i \cdot \mathbf{p}_i. \quad (8.4)$$

Similar to the point-to-far-field lens system in Section 3.5, the angular characteristic T_1 is independent of the direction coordinate \mathbf{p}_s , since the position coordinate \mathbf{q}_s at the source plane is given using (2.126) as

$$\mathbf{q}_s = \frac{\partial T_1}{\partial \mathbf{p}_s} = \mathbf{0}. \quad (8.5)$$

Hence, $T_1 = T_1(\mathbf{p}_i)$ and we have

$$\begin{aligned} T_1(\mathbf{p}_i) &= V_1(\mathbf{q}_s, \mathbf{q}_i) - \mathbf{q}_i \cdot \mathbf{p}_i \\ &= u(\hat{\mathbf{s}}) - n d(P_1, O_{t,1}) - \mathbf{q}_i \cdot \mathbf{p}_i. \end{aligned} \quad (8.6)$$

Using Figure 8.3 we find that

$$\mathbf{p}_i = n \frac{u(\hat{\mathbf{s}}) \mathbf{p}_s - \mathbf{q}_i}{d(P_1, O_{t,1})}, \quad i_3 = \frac{u(\hat{\mathbf{s}}) s_3}{d(P_1, O_{t,1})}, \quad (8.7)$$

and, using (8.3) and writing $u = u(\hat{\mathbf{s}})$, we obtain

$$\begin{aligned} T_1(\mathbf{p}_i) &= u - \frac{n}{d(P_1, O_{t,1})} \left[|u \mathbf{p}_s - \mathbf{q}_i|^2 + \mathbf{q}_i \cdot (u \mathbf{p}_s - \mathbf{q}_i) + (u s_3)^2 \right] \\ &= u - \frac{n}{d(P_1, O_{t,1})} [(u \mathbf{p}_s - \mathbf{q}_i) \cdot u \mathbf{p}_s] - n i_3 (u s_3). \end{aligned}$$

Rearranging terms and substituting \mathbf{p}_i from (8.7) gives

$$\begin{aligned} T_1(\mathbf{p}_i) &= u - n u \left[\frac{1}{d(P_1, O_{t,1})} (u \mathbf{p}_s - \mathbf{q}_i) \cdot \mathbf{p}_s + s_3 t_3 \right] \\ &= u - n u (\mathbf{p}_i / n \cdot \mathbf{p}_s + s_3 t_3) \\ &= u(1 - n \hat{\mathbf{s}} \cdot \hat{\mathbf{i}}). \end{aligned} \quad (8.8)$$

The generating function

Solving (8.8) for $u(\hat{\mathbf{s}})$ we obtain

$$u(\hat{\mathbf{s}}) = \frac{T_1(\mathbf{p}_i)}{1 - n \hat{\mathbf{s}} \cdot \hat{\mathbf{i}}}. \quad (8.9)$$

Note that it is not always true that $1 - n \hat{\mathbf{s}} \cdot \hat{\mathbf{i}} \neq 0$; this depends on the choice of source and intermediate target domain and we will check this during our numerical procedure later in the next chapter. Changing to stereographic coordinates using (8.2a) and (8.2b) gives

$$u(\mathbf{x}) = T_1(\mathbf{p}_i) \left(1 - n + \frac{2n |\mathbf{x} - \mathbf{y}_1|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}_1|^2)} \right)^{-1}, \quad (8.10)$$

where, for ease of notation, we continue to use the variable u to represent the optical surface, but now as a function of \mathbf{x} . We construct the generating function G_1 from the relation $u(\mathbf{x}) = G_1(\mathbf{x}, \mathbf{y}_1, w)$ with $w = T_1(\mathbf{p}_i)$, as

$$G_1(\mathbf{x}, \mathbf{y}_1, w) = w \left(1 - n + \frac{2n |\mathbf{x} - \mathbf{y}_1|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}_1|^2)} \right)^{-1}. \quad (8.11)$$

Note that $w = T_1(\mathbf{p}_i)$ is dependent on the outgoing ray $\hat{\mathbf{i}}$ and hence $w = w(\mathbf{y}_1)$ is a function of \mathbf{y}_1 . The function H_1 is the angular characteristic $T(\mathbf{p}_i)$ rewritten in stereographic coordinates, i.e.,

$$H_1(\mathbf{x}, \mathbf{y}_1, w) = w \left(1 - n + \frac{2n |\mathbf{x} - \mathbf{y}_1|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}_1|^2)} \right). \quad (8.12)$$

The cost function

Alternatively, taking the logarithm by introducing $u_1(\mathbf{x}) = -\log(u(\mathbf{x}))$ and $u_2(\mathbf{y}) = -\log(T(\mathbf{p}_i))$ we can rewrite (8.9) as

$$u_2(\mathbf{y}) - u_1(\mathbf{x}) = -\log \left(1 - n + \frac{2n |\mathbf{x} - \mathbf{y}_1|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}_1|^2)} \right) = c(\mathbf{x}, \mathbf{y}_1), \quad (8.13)$$

where $c(\mathbf{x}, \mathbf{y}_1)$ is a logarithmic cost function in optimal transport theory.

Energy conservation

By transferring the light from source to the intermediate target we require that all light from the source ends up at the intermediate target and energy is conserved, i.e.,

$$\int_{\mathcal{A}} f(\phi, \theta) \, d\mathcal{S}(\phi, \theta) = \int_{\hat{\mathcal{I}}(\mathcal{A})} h(\psi_1, \chi_1) \, d\mathcal{S}(\psi_1, \chi_1), \quad (8.14)$$

for an arbitrary set $\mathcal{A} \subset S^2$ and image set $\hat{\mathcal{I}}(\mathcal{A}) \subset S^2$.

Following the arguments of Section 3.5, if we substitute the inverse projections $\hat{\mathbf{s}} = \hat{\mathbf{s}}(\mathbf{x})$ and $\hat{\mathbf{i}} = \hat{\mathbf{i}}(\mathbf{y}_1)$ from (8.2a) and (8.2b) into (8.14), and substitute the mapping $\mathbf{y}_1 = \mathbf{m}_1(\mathbf{x})$ we obtain the Jacobian equation

$$\det(\mathbf{D}\mathbf{m}_1(\mathbf{x})) = \frac{(1 + |\mathbf{m}_1(\mathbf{x})|^2)^2}{(1 + |\mathbf{x}|^2)^2} \frac{\tilde{f}(\mathbf{x})}{\tilde{h}(\mathbf{m}_1(\mathbf{x}))}, \quad (8.15)$$

where $\tilde{h}(\mathbf{y}_1) = h(\psi(\mathbf{y}_1), \chi(\mathbf{y}_1))$. Note that the right-hand side also depends on the surface u since $\mathbf{m}_1 = \mathbf{m}_1(\mathbf{x}, u, \nabla u)$.

We define the corresponding transport boundary condition to (8.15) as

$$\mathbf{m}_1(\partial\mathcal{X}) = \partial\mathcal{Y}_1, \quad (8.16)$$

stating that all light from the boundary of the source \mathcal{X} is mapped to the boundary of the target \mathcal{Y}_1 .

8.1.2 The second freeform surface

At $z = 0$, the position and direction coordinates of the ray $\hat{\mathbf{t}}$ are given by \mathbf{q}_t and $\mathbf{p}_t = (t_1, t_2)$, respectively. Note that the position vector of the point P_2 is given by

$$u(\hat{\mathbf{s}}) \hat{\mathbf{s}} + v(\hat{\mathbf{s}}) \hat{\mathbf{i}} = (u(\hat{\mathbf{s}}) \mathbf{p}_s + v(\hat{\mathbf{s}}) \mathbf{p}_i / n, u(\hat{\mathbf{s}}) s_3 + v(\hat{\mathbf{s}}) t_3), \quad (8.17)$$

as shown in Figure 8.3.

In this section, we find Hamilton's angular characteristic function T from O to $O_{t,2}$, i.e., from the source to a final target point. We show that this is equal to the sum of the characteristic function T_1 from the previous section and the angular characteristic T from $O_{t,1}$ to $O_{t,2}$.

The angular characteristic $T(\mathbf{p}_t)$

We start by writing down the angular characteristic $T_2(\mathbf{p}_i, \mathbf{p}_t)$, which is the angular characteristic from $O_{t,1}$ to $O_{t,2}$. The expression for $T_2(\mathbf{p}_i, \mathbf{p}_t)$ is

$$\begin{aligned} T_2(\mathbf{p}_i, \mathbf{p}_t) &= V_2(\mathbf{q}_i, \mathbf{q}_t) + \mathbf{q}_i \cdot \mathbf{p}_i - \mathbf{q}_t \cdot \mathbf{p}_t \\ &= n d(P_1, O_{t,1}) + n v(\hat{\mathbf{s}}) - d(P_2, O_{t,2}) + \mathbf{q}_i \cdot \mathbf{p}_i - \mathbf{q}_t \cdot \mathbf{p}_t, \end{aligned} \quad (8.18)$$

where $n d(P_1, O_{t,1})$ is the optical path length from $O_{t,1}$ to P_1 , $n v(\hat{\mathbf{s}})$ is the optical path length from P_1 to P_2 , and $d(P_2, O_{t,2})$ is the distance from P_2 to the virtual image $O_{t,2}$. Since in general neither $\mathbf{q}_i = \mathbf{0}$ nor $\mathbf{p}_i = \mathbf{0}$, we cannot find T_2 as a function of the target coordinate \mathbf{p}_t only. By adding $T_1(\mathbf{p}_i)$ in (8.6) to $T_2(\mathbf{p}_i, \mathbf{p}_t)$ in (8.18) we see that $\mathbf{q}_i \cdot \mathbf{p}_i$ and $n d(P_1, O_{t,1})$ cancel and that

$$T_1(\mathbf{p}_i) + T_2(\mathbf{p}_i, \mathbf{p}_t) = u(\hat{\mathbf{s}}) + n v(\hat{\mathbf{s}}) - d(P_2, O_{t,2}) - \mathbf{q}_t \cdot \mathbf{p}_t. \quad (8.19)$$

Below we will show that this is equal to the total characteristic $T(\mathbf{p}_t)$ from O to $O_{t,1}$, which is independent of the source coordinate \mathbf{p}_s .

The total point characteristic between point $O(\mathbf{q}_s, 0)$ on the source plane and $O_{t,2}(\mathbf{q}_t, 0)$ (virtual image of the second surface) is given by

$$\begin{aligned} V(\mathbf{q}_s, \mathbf{q}_t) &= u(\hat{\mathbf{s}}) + n d(P_1, P_2) - d(P_2, O_{t,2}), \\ d(P_2, O_{t,2}) &= \sqrt{|(u(\hat{\mathbf{s}}) \mathbf{p}_s + v(\hat{\mathbf{s}}) \mathbf{p}_i/n) - \mathbf{q}_t|^2 + (u(\hat{\mathbf{s}}) s_3 + v(\hat{\mathbf{s}}) t_3)^2}, \end{aligned} \quad (8.20)$$

where $n d(P_1, P_2) = n v(\hat{\mathbf{s}})$ denotes the optical path length between P_1 and P_2 , and $d(P_2, O_{t,2})$ is the Euclidean distance between P_2 and $O_{t,2}$.

As in the previous section, we introduce Hamilton's angular characteristic $T(\mathbf{p}_s, \mathbf{p}_t)$ as the optical path length from O to Q_2 in Figure 8.3, and show it is independent of \mathbf{p}_s . Hamilton's angular characteristic for the second surface, which depends on the direction of the source ray and the final ray, is given by

$$T(\mathbf{p}_s, \mathbf{p}_t) = V(\mathbf{q}_s, \mathbf{q}_t) + \mathbf{q}_s \cdot \mathbf{p}_s - \mathbf{q}_t \cdot \mathbf{p}_t. \quad (8.21)$$

Just like T_1 , the angular characteristic T is independent of the direction coordinate \mathbf{p}_s , since the position coordinate \mathbf{q}_s at the source plane is given using (2.126) as

$$\mathbf{q}_s = \frac{\partial T}{\partial \mathbf{p}_s} = \mathbf{0}. \quad (8.22)$$

Thus, $T = T(\mathbf{p}_t)$ and

$$\begin{aligned} T(\mathbf{p}_t) &= V(\mathbf{q}_s, \mathbf{q}_t) - \mathbf{q}_t \cdot \mathbf{p}_t \\ &= u(\hat{\mathbf{s}}) + n v(\hat{\mathbf{s}}) - d(P_2, O_{t,2}) - \mathbf{q}_t \cdot \mathbf{p}_t, \end{aligned} \quad (8.23)$$

which is indeed equal to (8.19). Hence, for systems with multiple freeform surfaces we can add the characteristic functions to get the total characteristic function of the system. Here, we have shown this holds for the angular characteristic T , but it is not difficult to show that it holds for V , W and W^* as well.

Using Figure 8.3 we can see that

$$\mathbf{p}_t = \frac{u(\hat{\mathbf{s}}) \mathbf{p}_s + v(\hat{\mathbf{s}}) \mathbf{p}_i/n - \mathbf{q}_t}{d(P_2, O_{t,2})}, \quad t_3 = \frac{u(\hat{\mathbf{s}}) s_3 + v(\hat{\mathbf{s}}) l_3}{d(P_2, O_{t,2})}, \quad (8.24)$$

and, using (8.20) and writing $u = u(\hat{\mathbf{s}})$ and $v = v(\hat{\mathbf{s}})$, we write (8.23) as

$$\begin{aligned} T(\mathbf{p}_t) &= u + n v - \frac{1}{d(P_2, O_{t,2})} \left[|u \mathbf{p}_s + v \mathbf{p}_i/n - \mathbf{q}_t|^2 \right. \\ &\quad \left. + \mathbf{q}_t \cdot (u \mathbf{p}_s + v \mathbf{p}_i/n - \mathbf{q}_t) + (u s_3 + v l_3)^2 \right] \\ &= u + n v - \frac{1}{d(P_2, O_{t,2})} \left[(u \mathbf{p}_s + v \mathbf{p}_i/n - \mathbf{q}_t) \cdot (u \mathbf{p}_s + v \mathbf{p}_i/n) \right] \\ &\quad - t_3 (u s_3 + v l_3). \end{aligned}$$

Using (8.24) and reordering terms gives

$$\begin{aligned} T(\mathbf{p}_t) &= u + n v - \left[\frac{1}{d(P_2, O_{t,2})} (u \mathbf{p}_s + v \mathbf{p}_i/n - \mathbf{q}_t) \cdot \right. \\ &\quad \left. (u \mathbf{p}_s + v \mathbf{p}_i/n) + u s_3 t_3 + v l_3 t_3 \right] \\ &= u + n v - [\mathbf{p}_t \cdot (u \mathbf{p}_s + v \mathbf{p}_i/n) + u s_3 t_3 + v l_3 t_3] \\ &= u(1 - \hat{\mathbf{s}} \cdot \hat{\mathbf{t}}) + v(n - \hat{\mathbf{i}} \cdot \hat{\mathbf{t}}). \end{aligned} \quad (8.26)$$

Hence, we arrive at

$$T(\mathbf{p}_t) = u(\hat{\mathbf{s}}) (1 - \hat{\mathbf{s}} \cdot \hat{\mathbf{t}}) + v(\hat{\mathbf{s}}) (n - \hat{\mathbf{i}} \cdot \hat{\mathbf{t}}). \quad (8.27)$$

We remark that the map $\mathbf{y}_2 = \mathbf{m}_2(\mathbf{y}_1)$ from the intermediate target \mathcal{Y}_1 to the final target \mathcal{Y}_2 can also be seen as a mapping from a *generalized source domain* to the final target, where \mathbf{q}_i and \mathbf{p}_i are the position and direction coordinates of the generalized source domain. A generalized source domain still has zero étendue, since phase space volume is conserved, but is neither a parallel beam nor a point source.

The generating function

Solving (8.27) for $v(\hat{\mathbf{s}})$ we obtain

$$v(\hat{\mathbf{s}}) = \frac{T(\mathbf{p}_t) - u(\hat{\mathbf{s}}) (1 - \hat{\mathbf{s}} \cdot \hat{\mathbf{t}})}{n - \hat{\mathbf{i}} \cdot \hat{\mathbf{t}}}. \quad (8.28)$$

Changing to stereographic coordinates using (8.2) gives

$$v(\mathbf{x}) = \left(T(\mathbf{p}_t) - u(\mathbf{x}) \frac{2|\mathbf{x} - \mathbf{y}_2|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}_2|^2)} \right) \times \left(n - 1 + \frac{2|\mathbf{y}_1 - \mathbf{y}_2|^2}{(1 + |\mathbf{y}_1|^2)(1 + |\mathbf{y}_2|^2)} \right)^{-1}, \quad (8.29)$$

where we use the variable v to represent the optical surface, but now as a function of \mathbf{x} . We construct the generating function G_2 from the relation $v(\mathbf{x}) = G_2(\mathbf{x}, \mathbf{y}_2, w)$ with $w = T(\mathbf{p}_t)$, as

$$G_2(\mathbf{x}, \mathbf{y}_2, w) = \left(w - u(\mathbf{x}) \frac{2|\mathbf{x} - \mathbf{y}_2|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}_2|^2)} \right) \times \left(n - 1 + \frac{2|\mathbf{y}_1 - \mathbf{y}_2|^2}{(1 + |\mathbf{y}_1|^2)(1 + |\mathbf{y}_2|^2)} \right)^{-1}. \quad (8.30)$$

Note that $w = T(\mathbf{p}_t)$ is dependent on the outgoing ray $\hat{\mathbf{t}}$ and hence $w = w(\mathbf{y}_2)$ is a function of \mathbf{y}_2 . The unique inverse function $H_2(\mathbf{x}, \mathbf{y}_2, w)$ with $w = v(\mathbf{x})$ is the angular characteristic $T(\mathbf{p}_t)$ rewritten in stereographic coordinates as

$$H_2(\mathbf{x}, \mathbf{y}_2, w) = u(\mathbf{x}) \frac{2|\mathbf{x} - \mathbf{y}_2|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}_2|^2)} + w \left(n - 1 + \frac{2|\mathbf{y}_1 - \mathbf{y}_2|^2}{(1 + |\mathbf{y}_1|^2)(1 + |\mathbf{y}_2|^2)} \right). \quad (8.31)$$

The cost function

We cannot derive an optimal-transport cost function using any transformation of variables of (8.29). If we introduce $u_1(\mathbf{x}) = v(\mathbf{x})$, $u_2(\mathbf{y}) = T(\mathbf{p}_t)$, and

$$c_1(\mathbf{x}, \mathbf{y}_2) = u(\mathbf{x}) \frac{2|\mathbf{x} - \mathbf{y}_2|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}_2|^2)},$$

$$c_2(\mathbf{x}, \mathbf{y}_2) = n - 1 + \frac{2|\mathbf{y}_1(\mathbf{x}) - \mathbf{y}_2|^2}{(1 + |\mathbf{y}_1(\mathbf{x})|^2)(1 + |\mathbf{y}_2|^2)},$$

we would get a relation of the form

$$u_2(\mathbf{y}) - c_2(\mathbf{x}, \mathbf{y}_2)u_1(\mathbf{x}) = c_1(\mathbf{x}, \mathbf{y}_2), \quad (8.32)$$

which we cannot reduce to the form (3.1), since $c_2(\mathbf{x}(\hat{\mathbf{s}}), \mathbf{y}_2(\hat{\mathbf{t}}))$ is not a constant. We have $c_2(\mathbf{x}(\hat{\mathbf{s}}), \mathbf{y}_2(\hat{\mathbf{t}})) = n - \hat{\mathbf{i}}(\hat{\mathbf{s}}) \cdot \hat{\mathbf{t}} \geq 0$ for $n \geq 1$, cf. (8.28).

Energy conservation

By transferring the light from the source to the final target we require that all light from the source ends up at the target and energy is conserved, i.e.,

$$\int_{\mathcal{A}} f(\phi, \theta) \, d\mathcal{S}(\phi, \theta) = \int_{\hat{\mathbf{t}}(\mathcal{A})} g(\psi_2, \chi_2) \, d\mathcal{S}(\psi_2, \chi_2), \quad (8.33)$$

for an arbitrary set $\mathcal{A} \subset \mathbb{S}^2$ and image set $\hat{\mathbf{t}}(\mathcal{A}) \subset \mathbb{S}^2$.

As in the previous section, we follow the arguments of Section 3.5. If we substitute $\hat{\mathbf{s}} = \hat{\mathbf{s}}(\mathbf{x})$ and $\hat{\mathbf{t}} = \hat{\mathbf{t}}(\mathbf{y}_2)$ from (8.2a) and (8.2b) into (8.33), and subsequently substitute the mapping $\mathbf{y}_2 = \mathbf{m}(\mathbf{x}) = \mathbf{m}_2(\mathbf{m}_1(\mathbf{x}))$ we obtain the Jacobian equation

$$\det(D\mathbf{m}(\mathbf{x})) = \frac{(1 + |\mathbf{m}(\mathbf{x})|^2)^2}{(1 + |\mathbf{x}|^2)^2} \frac{\tilde{f}(\mathbf{x})}{\tilde{g}(\mathbf{m}(\mathbf{x}))}. \quad (8.34)$$

Hence, we have a first-order partial differential equation for the composite mapping \mathbf{m} . We write the transport boundary condition as

$$\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}_2. \quad (8.35)$$

For the first surface of the double freeform lens we can construct the non-quadratic, logarithmic, optimal-transport cost function

$$c(\mathbf{x}, \mathbf{y}_1) = -\log \left(1 - n + \frac{2n |\mathbf{x} - \mathbf{y}_1|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}_1|^2)} \right), \quad (8.36)$$

the generating function

$$G_1(\mathbf{x}, \mathbf{y}_1, w) = w \left(1 - n + \frac{2n |\mathbf{x} - \mathbf{y}_1|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}_1|^2)} \right)^{-1}, \quad (8.37)$$

with corresponding inverse

$$H_1(\mathbf{x}, \mathbf{y}_1, w) = w \left(1 - n + \frac{2n |\mathbf{x} - \mathbf{y}_1|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}_1|^2)} \right). \quad (8.38)$$

Combining the mapping with energy conservation gives the Jacobian equation

$$\det(D\mathbf{m}_1(\mathbf{x})) = \frac{(1 + |\mathbf{m}_1(\mathbf{x})|^2)^2}{(1 + |\mathbf{x}|^2)^2} \frac{\tilde{f}(\mathbf{x})}{\tilde{h}(\mathbf{m}_1(\mathbf{x}))}. \quad (8.39)$$

For the second surface of the double freeform lens we cannot construct an optimal transport cost function. The generating function is

$$G_2(\mathbf{x}, \mathbf{y}_2, w) = \left(w - u(\mathbf{x}) \frac{2|\mathbf{x} - \mathbf{y}_2|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}_2|^2)} \right) \times \left(n - 1 + \frac{2|\mathbf{y}_1 - \mathbf{y}_2|^2}{(1 + |\mathbf{y}_1|^2)(1 + |\mathbf{y}_2|^2)} \right)^{-1}, \quad (8.40)$$

with corresponding inverse

$$H_2(\mathbf{x}, \mathbf{y}_2, w) = u(\mathbf{x}) \frac{2|\mathbf{x} - \mathbf{y}_2|^2}{(1 + |\mathbf{x}|^2)(1 + |\mathbf{y}_2|^2)} + w \left(n - 1 + \frac{2|\mathbf{y}_1 - \mathbf{y}_2|^2}{(1 + |\mathbf{y}_1|^2)(1 + |\mathbf{y}_2|^2)} \right), \quad (8.41)$$

where $u(\mathbf{x})$ is the first freeform surface with associated mapping $\mathbf{y}_1 = \mathbf{m}_1(\mathbf{x})$. Combining the mapping with energy conservation gives the Jacobian equation

$$\det(D\mathbf{m}(\mathbf{x})) = \frac{(1 + |\mathbf{m}(\mathbf{x})|^2)^2}{(1 + |\mathbf{x}|^2)^2} \frac{\tilde{f}(\mathbf{x})}{\tilde{g}(\mathbf{m}(\mathbf{x}))}. \quad (8.42)$$

8.2 The GJLS algorithm for a double freeform lens

In this section, we explain how to compute a double freeform lens using the GJLS algorithm twice.

To compute the first freeform surface $u(x)$ and first mapping $y_1 = m_1(x)$, we first apply the GJLS algorithm from Section 6.3 on (8.39), using $H = H_1$ in (8.38) and using

$$F = F_1(x, m_1(x), u(x)) = \frac{(1 + |m_1(x)|^2)^2}{(1 + |x|^2)^2} \frac{\tilde{f}(x)}{\tilde{h}(m_1(x))}, \quad (8.43)$$

in (6.105). We can also choose to compute $u(x)$ and the mapping $y_1 = m_1(x)$ using the GLS algorithm, since we formulated an optimal-transport cost function.

Subsequently, we substitute $u(x)$ and $y_1 = m_1(x)$ into H_2 in (8.41) and

$$F = F_2(x, m(x), v(x)) = \frac{(1 + |m(x)|^2)^2}{(1 + |x|^2)^2} \frac{\tilde{f}(x)}{\tilde{g}(m(x))}, \quad (8.44)$$

from (8.42) into (6.105), and run the GJLS algorithm again to compute $v(x)$ and the composite mapping $y_2 = m(x) = (m_2 \circ m_1)(x)$.

For the second surface we have the property $G_w > 0$, cf. (8.28) and (8.40) since $n - \hat{i} \cdot \hat{t} > 0$ with $n > 1$, which results in the max/min pair in (4.48) for a G-convex solution u or in the min/max pair in (4.49) for a G-concave solution u .

For the first surface, this is not necessarily true; see (8.9) and (8.11). For $G_w > 0$ we require $1 - n \hat{s} \cdot \hat{i} > 0$. Whether this inequality is satisfied depends on the choice of source and target domains. Otherwise $G_w < 0$, which results in a max/max pair for a G-convex solution and min/min pair for a G-concave solution, analogous to (4.48) and (4.49). This does not have an effect on our numerical method, but we need to make sure that $G_w \neq 0$.

8.3 Summary

In this chapter, we considered a double freeform lens with a point source and far-field target. We found two generating functions using Hamilton's characteristic functions. The expressions found for the first freeform surface are very similar to those for the point-to-far-field lens in Section 3.5. For the second freeform surface, we showed that we can concatenate characteristic functions.

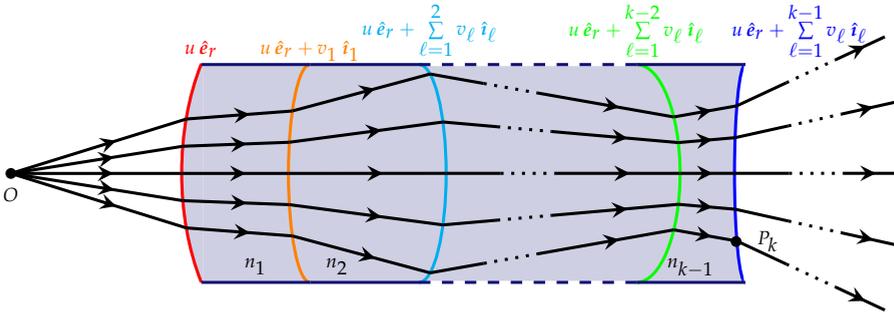


Figure 8.4: Freeform lens with k freeform surfaces for a point source and a far-field target.

For a general number of optical surfaces k , we could rewrite the sum in (8.19) as

$$\begin{aligned} T_1(\mathbf{p}_{i,1}) + T_2(\mathbf{p}_{i,1}, \mathbf{p}_{i,2}) + T_3(\mathbf{p}_{i,2}, \mathbf{p}_{i,3}) + \dots + T_k(\mathbf{p}_{i,k-1}, \mathbf{p}_t) &= T(\mathbf{p}_t) \\ &= u(\hat{\mathbf{s}}) + n_1 v_1(\hat{\mathbf{s}}) + n_2 v_2(\hat{\mathbf{s}}) + \dots + n_{k-1} v_{k-1}(\hat{\mathbf{s}}) - d(P_k, O_{t,k}) - \mathbf{q}_t \cdot \mathbf{p}_t, \end{aligned} \quad (8.45)$$

where $\mathbf{p}_{i,1} \dots \mathbf{p}_{i,k-1}$ are the intermediate momenta, the second optical surface is called v_1 , the third v_2 , \dots and the last v_{k-1} , and the refractive indices are n_1, n_2, \dots, n_{k-1} ; see Figure 8.4. P_k is a point on the last surface given by

$$u(\hat{\mathbf{s}}) \hat{\mathbf{s}} + v_1(\hat{\mathbf{s}}) \hat{\mathbf{i}}_1 + \dots + v_{k-1}(\hat{\mathbf{s}}) \hat{\mathbf{i}}_{k-1}, \quad (8.46)$$

where $\hat{\mathbf{i}}_1, \dots, \hat{\mathbf{i}}_{k-1}$ are the intermediate rays. $O_{t,k}$ is the final target point on the source plane, which is not drawn in Figure 8.4. This point can be found by extending the ray emanating from P_k back towards the source, as we did for $O_{t,1}$ and $O_{t,2}$ in Section 8.1.1 and 8.1.2. Writing out the full expression for $T(\mathbf{p}_t)$ in (8.45) we can derive a generating function for the last surface in the system. To derive a generating function for the second-to-last surface we would need the sum from T_1 to T_{k-1} , and for the third-to-last from T_1 to T_{k-2} , etc.

Of course, we would also need to define $k - 1$ intermediate target distributions, which we can choose freely. Hence, we can create a large number of such free parameters. In this thesis, we will stick to the double freeform lens for now. In the next chapter, we will show some numerical results and explain our choice of the intermediate target intensity as an interpolation between the source and final target intensity based on an initial mapping.

Chapter 9

Numerical Results – Part II

In this chapter, we present two numerical examples for the double freeform lens. We compute a double freeform lens by running the GJLS algorithm twice, once for each surface of the system. We present one numerical example to illustrate the accuracy and efficiency of the algorithm. In the second example, we compute a double freeform lens that transforms the light of a point source into a picture of Van Gogh on a screen in the far field. We introduce a tuning parameter β which determines the intermediate target intensity as an interpolation between the source intensity and final target intensity. By varying this parameter we compute multiple double freeform lenses that are solutions to the same problem but differ in design. The simulations in this chapter are adapted from [129].

9.1 Exact double freeform lens

To test the accuracy of the algorithm, we solve the generalized Monge-Ampère equations for a double freeform lens where we pre-compute the right-hand sides in (8.39) and (8.42) corresponding to two known surfaces, as we did in Section 7.1.1 and 7.3.2. We choose

$$u(\mathbf{x}) = 1 + |\mathbf{x}|^2, \quad v(\mathbf{x}) = 1, \quad (9.1)$$

such that $u(\mathbf{x})$ represents an elliptic paraboloid. We use the implicit relation for the mapping in (4.77) to solve for the mappings $\mathbf{m}_1(\mathbf{x})$ and $\mathbf{m}(\mathbf{x})$ using H_1 in (8.38) and H_2 in (8.41), respectively.

We consider a square source domain $\mathcal{X} = [-1, 1]^2$. Using the mapping $\mathbf{m}_1(\mathbf{x})$ we compute the target domain \mathcal{Y}_1 and the corresponding target boundary. Using the mapping $\mathbf{m}(\mathbf{x})$ we compute the target domain \mathcal{Y}_2 and the

corresponding target boundary. We use H_1 in (8.38) and compute the right-hand side $F_1(\mathbf{x}, \mathbf{m}_1(\mathbf{x}), u(\mathbf{x}))$ in (8.43) using the exact $\mathbf{m}_1(\mathbf{x})$ and $u(\mathbf{x})$. We use the initial guess \mathbf{m}^0 given in (6.11) and a spherical initial surface, which results in an SND matrix \mathbf{P}^0 . Verifying that $G_w(\mathbf{x}, \mathbf{m}_1^0, u^0) < 0$ at all points in the domain gives that we consider a min/min pair in Section 4.4.2 and thus compute a G-concave u . We run the algorithm to compute $u(\mathbf{x})$ and the intermediate mapping $\mathbf{m}_1(\mathbf{x})$. We compute the surface u at every iteration, i.e., we do not attempt to reduce the computation time as we did in Section 7.3.3. The results are shown in Figure 9.1.

After computing $\mathbf{m}_1(\mathbf{x})$ and $u(\mathbf{x})$ we run the algorithm again using H_2 in (8.41). We evaluate the right-hand side $F_2(\mathbf{x}, \mathbf{m}(\mathbf{x}), v(\mathbf{x}))$ in (8.44) using the exact $\mathbf{m}(\mathbf{x})$ and $v(\mathbf{x})$. We use the initial mapping in (6.11) and set v^0 to a constant value. We compute a G-convex v since $G_w(\mathbf{x}, \mathbf{m}^0, v^0) > 0$ and $\text{tr}(\mathbf{P}^0) \leq 0$. We run the algorithm to compute $v(\mathbf{x})$ and the composite mapping $\mathbf{m}(\mathbf{x})$, computing the surface v at every iteration. The results are shown in Figure 9.2.

Note that for the second surface we have the property $G_w > 0$, cf. (8.40) since $n - \hat{\mathbf{i}} \cdot \hat{\mathbf{t}} > 0$ with $n > 1$, which results in the max/min pair in (4.48) for a G-convex solution v or in the min/max pair in (4.49) for a G-concave solution v . For the first surface, this is not necessarily true; see (8.9) and (8.11). For $G_w > 0$ we require $1 - n \hat{\mathbf{s}} \cdot \hat{\mathbf{i}} > 0$. Whether this inequality is satisfied depends on the choice of source and target domains. Otherwise $G_w < 0$, which results in a max/max pair for a G-convex solution and min/min pair for a G-concave solution. As stated above, we compute a G-concave solution for u (min/min pair), after verifying that $G_w(\mathbf{x}, \mathbf{m}_1^0, u^0) < 0$ at all points of the domain, and a G-convex v (max/min pair). We note that we can also choose other combinations of G-convex and G-concave surfaces for the double freeform lens, which results in different designs. This point is included in our recommendations for future research in the next chapter.

We use refractive index $n = 1.5$ and $\alpha = 0.2$. Figures 9.1b and 9.2b show the difference of the mappings for the two components m_1 and m_2 with the exact solutions for several $N \times N$ grids with logarithmic least-squares fits. Figures 9.1c and 9.2c show the difference of the surfaces with the exact solutions. We observe a second-order convergence to the exact solution. The convergence of J_I and J_B are shown in Figure 9.1d and 9.2d. Note that the functionals J_I and J_B reach a plateau at a certain iteration number, due to discretization and rounding errors. The first and second surface for $N = 100$ are plotted in Figure 9.3a. The second surface deviates slightly from the exact solution, with the absolute difference displayed in Figure 9.3b. Details on the number of iterations and computation time of the algorithm are presented in Table 7.2,

where we used the stopping criterion (7.7). The number of iterations required increases sublinearly with N . The computation time increases quadratically. Lastly, we see that J_I and J_B have approximately fourth-order convergence.

Algorithm to compute $u(x)$ and $y_1 = m_1(x)$

Grids	20×20	40×40	60×60	80×80	100×100	Fits
Iterations	67	97	124	148	170	$\propto N^{0.58}$
Time [s]	4	11	21	36	56	$\propto N^{1.6}$
J_I	2.0×10^{-5}	1.2×10^{-6}	2.3×10^{-7}	7.4×10^{-8}	3.0×10^{-8}	$\propto N^{-4.0}$
J_B	3.2×10^{-7}	2.2×10^{-8}	4.3×10^{-9}	1.4×10^{-9}	5.7×10^{-10}	$\propto N^{-3.9}$

Algorithm to compute $v(x)$ and $y = m(x)$

Iterations	79	148	206	259	309	$\propto N^{0.8}$
Time [s]	3	9	22	44	83	$\propto N^{2.0}$
J_I	2.5×10^{-6}	1.2×10^{-7}	2.0×10^{-8}	6.0×10^{-9}	2.4×10^{-9}	$\propto N^{-4.3}$
J_B	9.6×10^{-8}	4.1×10^{-9}	7.5×10^{-10}	2.3×10^{-10}	9.3×10^{-11}	$\propto N^{-4.3}$

Table 9.1: “Exact-lens” problem: number of iterations, total computation time (in seconds) and residuals in the GJLS algorithm.

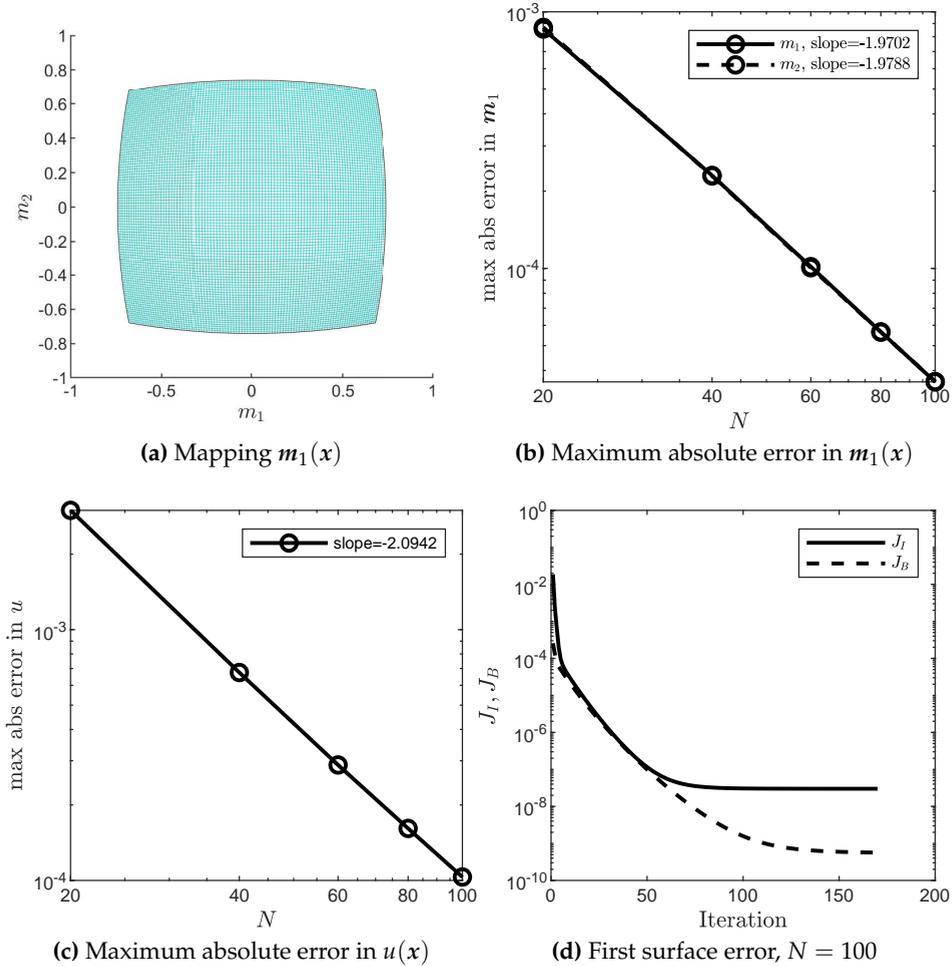


Figure 9.1: “Exact-lens” problem: we calculate solutions to u and $m_1(x)$ with parameter values $\alpha = 0.2$ and $n = 1.5$. The mapping of the G-concave solution is shown in (a) and the maximum absolute error in (b). Figure (c) shows the maximum absolute error in u and (d) presents the convergence history.

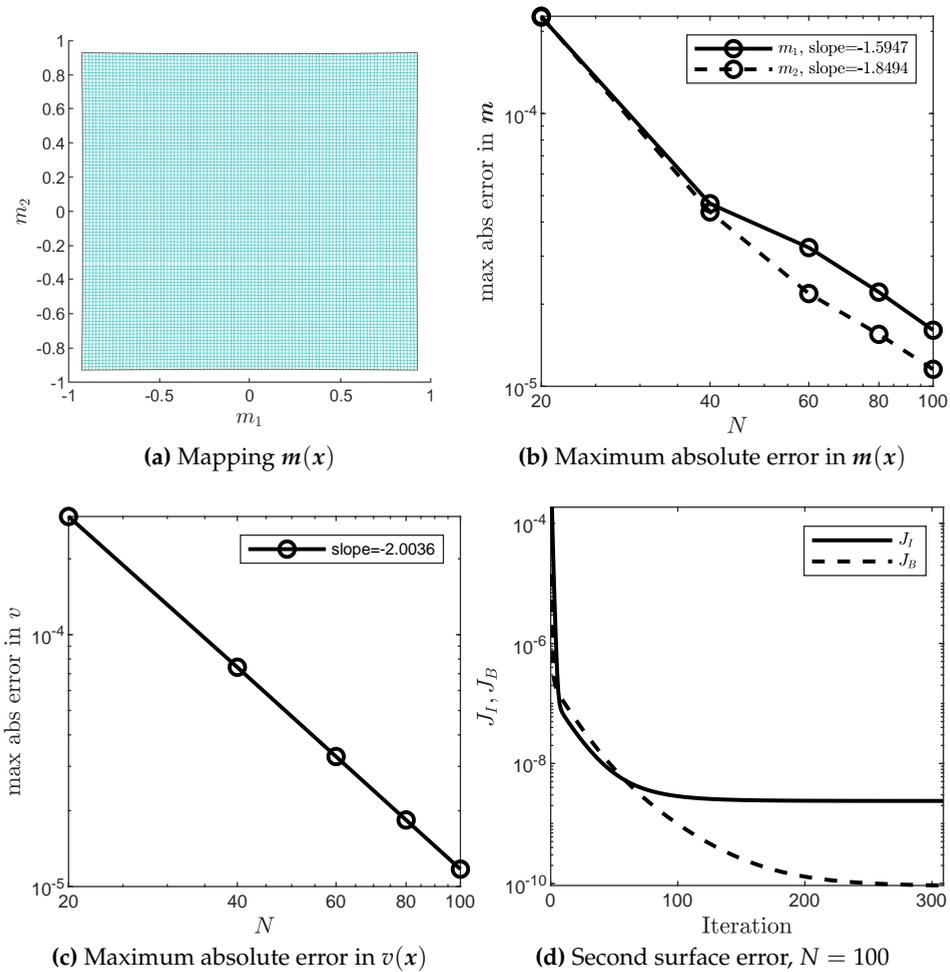
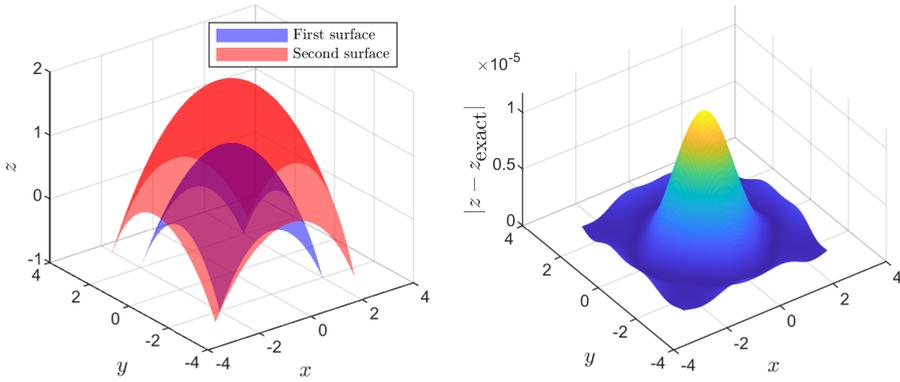


Figure 9.2: “Exact-lens” problem: we calculate solutions to v and $y = m(x)$ with parameter values $\alpha = 0.2$ and $n = 1.5$. The mapping of the G-convex solution is shown in (a) and the maximum absolute error in (b). Figure (c) shows the maximum absolute error in v and (d) presents the convergence history.



(a) First and second surfaces, $N = 100$ (b) Absolute error second surface $N = 100$

Figure 9.3: “Exact-lens” problem: the surfaces and absolute error of the second surface for $N = 100$.

9.2 Van Gogh double freeform lens

We consider a square source domain $\mathcal{X} = [-0.5, 0.5]^2$ with a Lambertian source intensity $f(\phi, \theta) = \cos(\phi)$ inside the square, i.e., the emittance is proportional to the cosine of the zenith ϕ described in Section 3.1.2. Using (3.7), we derive that $\tilde{f}(x) = (1 - |x|^2)/(1 + |x|^2)$ in stereographic coordinates.

The twice refracted rays are projected on a screen P in the far field, parallel to the plane $z = 0$. The required illuminance $L(\xi, \eta)$ [lm/m^2], with (ξ, η) the Cartesian coordinates on the projection screen, is derived from the grayscale values of a famous painting by Van Gogh: *Self portrait*, 1887, The Art Institute of Chicago [143]. As in Section 7.2 and 7.4, we increase values of $\tilde{g}(\mathbf{y}_2)$ that are below a threshold of 15% of its maximum value to this threshold. The target distribution $\tilde{g}(\mathbf{y}_2)$ is a deformation of the illuminance $L(\xi, \eta)$; the conversion from $L(\xi, \eta)$ to $\tilde{g}(\mathbf{y}_2)$ is explained in Section 3.1.4.

We discretize the source domain by a 200×200 grid and use the parameter values $\alpha = 0.1$ and $n = 1.5$.

To compute the first freeform surface, we need to choose an intermediate target intensity $\tilde{h}(\mathbf{y}_1) = h(\psi_1(\mathbf{y}_1), \chi_1(\mathbf{y}_1))$. We can see the computation of \tilde{h} as a dynamic optimal-transport (or generating-function) problem, i.e., we can think of a whole spectrum of intermediate target intensities starting from the source intensity but progressively translating and scaling towards the final target intensity. Taking a regular interpolation between the source and final

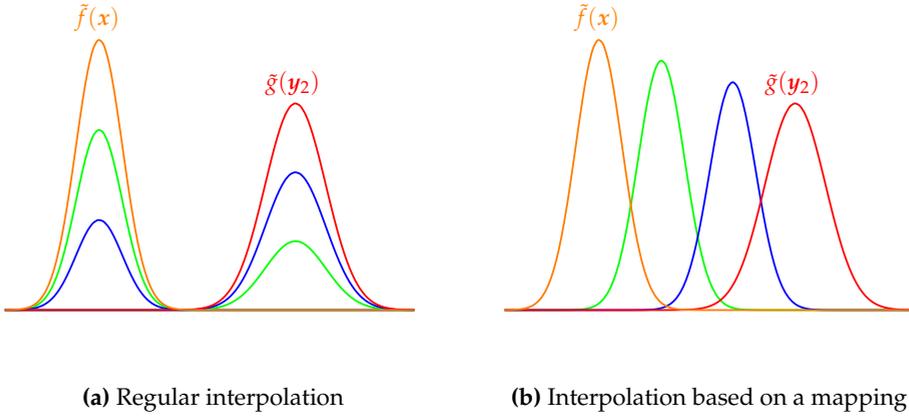


Figure 9.4: Example of a cross-section of the intensities $\tilde{f}(x)$ (orange) and $\tilde{g}(y_2)$ (red). The intermediate target intensities (green and blue) are interpolations of the source and target intensity. The interpolation in (a) is a regular interpolation of the intensities while the interpolation in (b) based on an initial mapping from source to target results in a progressive translation and scaling.

target intensity, e.g.,

$$\tilde{h}(y_1) = \beta \tilde{f}(x) + (1 - \beta) \tilde{g}(y_2), \quad 0 \leq \beta \leq 1, \quad (9.2)$$

results in mixtures of the two intensities but not in progressive translation and scaling. Figure 9.4a shows a schematic figure of the regular interpolation, while Figure 9.4b is the interpolation we are after. To obtain this translation and scaling, we proceed by finding the intermediate target intensity using a mapping from the source to the final target.

To compute an intermediate target intensity we first run the least-squares procedure to compute a mapping \tilde{m} from the source to the final target. We can use any generating function or cost function for this purpose. For simplicity we consider the generating function

$$G(x, y, z) = x \cdot y + z. \quad (9.3)$$

Using this generating function, the generalized Monge-Ampère equation in (6.105) reduces to the standard Monge-Ampère equation involving the determinant of the Hessian matrix of u and the matrix C in (4.81) is the identity matrix (multiplied by a minus sign). We use the GJLS algorithm to compute a mapping $y_2 = \tilde{m}(x)$ from \mathcal{X} to \mathcal{Y}_2 . Using the generating function (9.3) gives us an initial mapping \tilde{m} equivalent to the initial mapping used in most ray-mapping methods.

We consider the intermediate mapping

$$\tilde{m}_1(x, \beta) = \beta x + (1 - \beta) \tilde{m}(x), \quad 0 \leq \beta \leq 1, \quad (9.4)$$

as a weighted average of the identity mapping and the initial mapping $\tilde{m}(x)$. Subsequently, we set the intermediate intensity \tilde{h} as

$$\tilde{h}(y_1, \beta) = \frac{f(x)}{\det(D\tilde{m}_1(x, \beta))}, \quad (9.5)$$

where $\tilde{h}(y_1, \beta)$ is a function of y_1 and β , which we usually denote as $\tilde{h}(y_1)$. The intermediate target intensities for several β ($0 \leq \beta \leq 1$) is given in the first row in Figure 9.5. Choosing the intermediate target intensity in this way is akin to choosing Wasserstein barycentric averages with a quadratic cost function in [11].

Once we have computed $\tilde{h}(y_1)$ we use H_1 in (8.38) and use the algorithm to solve the Jacobian equation (8.39) to compute the intermediate mapping $m_1(x)$ and $u(x)$, as explained in Section 8.2. We use the initial guess m^0 given in (6.11) and a spherical initial surface, which results in an SND matrix P^0 . Verifying that $G_w(x, m_1^0, u^0) < 0$ at all points in the domain gives that we consider a min/min pair in Section 4.4.2 and thus compute a G-concave u . We compute the surface u at every iteration.

Subsequently, we use H_2 in (8.41) and run the algorithm to solve (8.42) to compute the composite mapping $m(x)$ and $v(x)$. We compute a G-convex v since $G_w(x, m^0, v^0) > 0$ and $\text{tr}(P^0) \leq 0$. Again, we compute the surfaces v at every iteration.

Figure 9.5 shows the results of the mappings and surfaces for various β . When $\beta = 1$ the mapping m_1 is simply the identity mapping and u is a spherical surface which does not alter the direction of the incoming rays. For smaller β more detail of the Van Gogh picture is incorporated in the first optical surface.

Table 9.2 shows details on the number of iterations and computation times. The number of iterations and computation time to compute the first surface increase for smaller β , with a maximum computation time of 263 seconds. To compute the second surface the number of iterations and computation time are similar for all values of β , with a mean computation time of 143 seconds.

In Figure 9.6 the ray-trace results of the lens are displayed, using our self-programmed ray-tracing algorithm explained in Section 7.7, performed with 10 million rays from quasi-random positions (quasi-Monte Carlo) on the source domain and 200×200 bins. The ray-trace results match the portrait of Van Gogh and look the same to the naked eye for all β . Hence, only the results

for $\beta = 1$ are shown in Figure 9.6. None of the traced rays missed the first or the second surface. Note that in the derivations of the generating functions, we assume that each ray \hat{s} refracts at the optical surfaces and always reaches the target plane. Hence, we ignore the occurrence of total internal reflection (TIR). The vectorial law of refraction [66, p. 140] gives

$$\hat{i} = \frac{1}{n} \hat{s} - \left(\frac{1}{n} (\hat{s} \cdot \hat{n}_1) + \sqrt{1 - \frac{1}{n^2} (1 - (\hat{s} \cdot \hat{n}_1)^2)} \right) \hat{n}_1, \quad (9.6a)$$

$$\hat{t} = n \hat{i} - \left(n (\hat{i} \cdot \hat{n}_2) + \sqrt{1 - n^2 (1 - (\hat{i} \cdot \hat{n}_2)^2)} \right) \hat{n}_2, \quad (9.6b)$$

requiring that the arguments of the square roots $1 - 1/n^2 (1 - (\hat{s} \cdot \hat{n}_1)^2)$ and $1 - n^2 (1 - (\hat{i} \cdot \hat{n}_2)^2)$ are nonnegative for the first and second surface, respectively. Here, \hat{n}_1 is the normal to the first surface and \hat{n}_2 is the normal to the second surface. Both conditions are satisfied for all rays traced, so TIR does not occur.

To verify the quality of the ray-trace results we take the difference between the target intensity on the projection screen (interpolated bilinearly onto the ray-tracing grid) and the ray-tracing irradiance. For all values of β the RMS error is approximately 0.07 and the correlation 0.91, and the ray-tracing irradiance closely matches the target intensity. Thus, even for complicated and detailed target distributions such as the picture of Van Gogh, we can distribute the refractive power over both surfaces of the lens in any way we like. This may lead to more flexibility in designing compact optical components in lamps and other applications, which requires implementations of our algorithm in modern optical software technologies. Points of future research will be discussed in the next chapter.

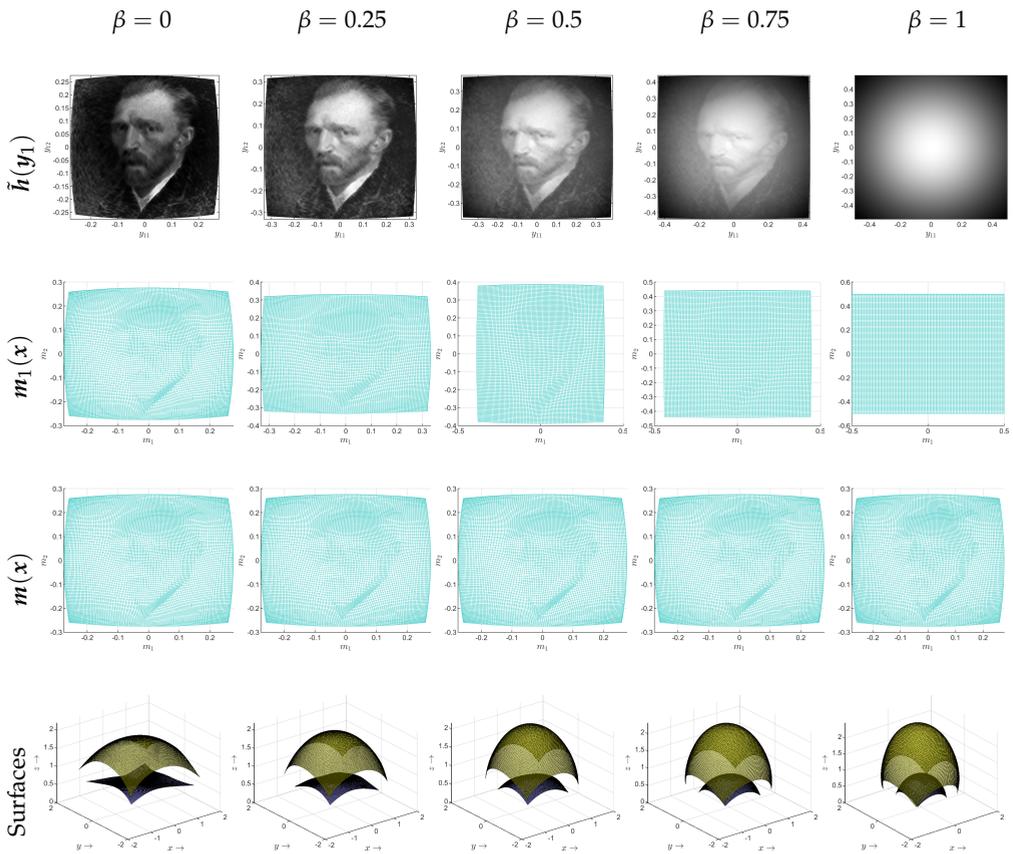


Figure 9.5: “Van-Gogh-lens” problem: the intermediate target intensities $\tilde{h}(y_1)$, mappings $m_1(x)$ and $m(x)$ and optical surfaces $u(x)$ (bottom surface) and $v(x)$ (top surface) plotted in Euclidean space. Van Gogh: ©The Art Institute of Chicago [143].

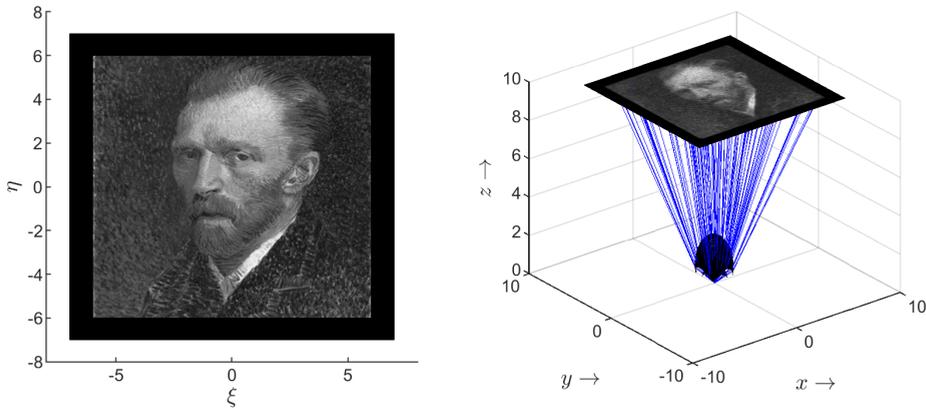


Figure 9.6: “Van-Gogh-lens” problem: the ray-trace results of the double freeform lens for $\beta = 1$. For all values of β the result looks the same. Van Gogh: ©The Art Institute of Chicago [143].

Algorithm to compute $u(x)$ and $y_1 = m_1(x)$

β	0	0.25	0.5	0.75	1
Iterations	500	150	150	150	10
Time [s]	263	79	81	80	5
J_I	6.3×10^{-4}	2.4×10^{-5}	1.1×10^{-5}	4.6×10^{-6}	9.7×10^{-27}
J_B	1.4×10^{-7}	3.8×10^{-8}	1.0×10^{-8}	1.7×10^{-9}	7.2×10^{-30}

Algorithm to compute $v(x)$ and $y = m(x)$

Iterations	150	148	150	150	150
Time [s]	142	140	145	143	143
J_I	3.0×10^{-2}	3.2×10^{-2}	3.4×10^{-2}	3.5×10^{-2}	3.5×10^{-2}
J_B	1.4×10^{-7}	1.6×10^{-7}	1.6×10^{-7}	1.5×10^{-7}	1.1×10^{-7}

Table 9.2: “Van-Gogh-lens” problem: number of iterations, total computation time (in seconds) and residuals in the GJLS algorithm.

9.3 Summary

In this chapter, we computed a double freeform lens using the GJLS algorithm. In the first example, we considered a target intensity corresponding to a known surface and showed that the algorithm indeed converges to the exact solution. In the second example, we challenged the algorithm to compute a double freeform lens that converts the light of a point source into a picture of Van Gogh on a screen in the far field. We introduced the parameter β to distribute the refractive power over both surfaces of the lens and computed multiple solutions that vary in shape. This added flexibility in the design could be valuable for the production process in a lab or factory. If multiple designs are available, they can be compared to optimize qualities such as suitability for optical diamond turning techniques, i.e., whether molds can be produced efficiently, and the compactness of the final design.

Chapter 10

Conclusions and Recommendations

10.1 Summary and conclusions

In this thesis, we developed a generic framework to describe optical systems using generated Jacobian equations. We started from Maxwell's equations and derived the laws of geometrical optics. We presented the theory on Hamiltonian optics and used Hamilton's characteristic functions to find generating functions for the 16 base-case optical systems, i.e., systems consisting of a minimum number of either reflector or lens surfaces for a parallel source or point source and a far-field, near-field, point or parallel target. For 5 out of the 16 base-case optical systems, we showed the full mathematical descriptions, while for the remaining cases we simply presented the generating functions (and cost functions if they exist).

We showed that the optical mapping can be derived by considering a G-convex or G-concave solution for the location of the optical surface u . Combining the optical mapping with energy conservation resulted in a generated Jacobian equation or generalized Monge-Ampère equation (if a cost function in optimal transport theory exists).

We extended a least-squares method previously used for optimal-transport problems to a generating-function framework. First, we presented the optimal-transport approach as the *generalized least-squares* (GLS) algorithm, which takes the cost function of the optical system as input. We extended this method to polar (stereographic) coordinates. The algorithm works by calculating the optical mapping in an iterative procedure, which involves two point-wise nonlinear minimization steps and solving a coupled boundary value problem. Subsequently, we compute the optical surface from the mapping by solving a

Neumann problem.

However, not for all optical systems such a cost function exists, but a generating function does. We modified the GLS algorithm to the *generated Jacobian least-squares* (GJLS) algorithm which takes the generating function of an optical system as input. The main difference with the optimal-transport approach is that the Neumann problem for the optical surface is now also incorporated into the iterative procedure. By using G-convexity theory to find a solution, we extended the applicability of the least-squares procedure to a much wider range of optical systems.

In the numerical results, we presented many experiments of the GLS and GJLS numerical approaches. We transformed a picture of a frog into a prince, discussed the accuracy and efficiency of the algorithm using a few test problems (e.g., a tilted flat reflector surface and square-to-circle problem), and computed a peanut lens for road lighting purposes. We also compared the performance of the GJLS algorithm to the GLS algorithm for an exact solution given as an ellipsoidal lens surface. We concluded that the GJLS algorithm performs better in accuracy and similarly in computation time. The computation time can be significantly reduced by updating the optical surface less frequently, i.e., not at every iteration.

As the most challenging test case, we used the least-squares approach to compute a double freeform lens. First, we ran the algorithm using the generating function for the first surface u with corresponding optical mapping from the source to an intermediate target. Subsequently, we substituted the mapping and surface u in the generating function for the second surface v and re-ran the least-squares algorithm to compute the final mapping from the source to the final target and second surface v . We tested the algorithm for a lens converting the light of a point source into a picture of Van Gogh on a screen in the far field. By choosing different intermediate target intensities as an interpolation between the source and final target intensity based on an approximate mapping, we have introduced a tuning parameter to distribute the refractive power over the two lens surfaces. This tuning parameter could be used to optimize the compactness of the system, which is a point of future research. Along with this point I will discuss more new questions for future research below.

10.2 Future research

During the writing process of a few of my papers, my supervisor Jan always pointed out my bold statements about future research. There are yet many things that can be added to the mix. Below I have listed the main items:

- In Chapter 3, we included a nonexhaustive list of optical systems with generating functions. We could think of more systems, such as systems which are composed of a combination of reflectors and lenses.
- In this thesis, we introduced polar stereographic coordinates for the GLS algorithm. This is a suitable coordinate system when considering point sources emitting cone-shaped bundles of light, because the source domain is circular in stereographic coordinates. Polar and/or elliptical coordinates have yet to be added to the GJLS algorithm.
- Each system has a different generating function and cost function (if it exists). We can approximate the cost function by taking a Taylor expansion of the cost function truncated after the second-order term. Consequently, this Taylor expansion is quadratic and we can compare the performance of the algorithm with the original cost function and the quadratic cost function [154]. From numerical experiments it was shown that the main indicator for differences between the cost function and the Taylor expansion in the resulting light distributions at the target is the relative difference of the cost function with the Taylor expansion. Such derivations and experiments can be performed for more optical systems. Moreover, can we do similar experiments with generating functions and can we characterize how far a generating function is away from the parallel-to-far-field generating function (corresponding to a quadratic cost function)?
- Another point of future research is whether we can apply our numerical algorithm to problems not related to optics. For instance, generating functions and generated Jacobian equations can also be derived for matching problems in economics [69], where $u(x)$ represents the maximal utility for a buyer x , and the map $m(x)$ assigns to every buyer x the commodity $y = m(x)$ that they should buy to maximize utility.
- In this thesis, we did not consider generalized or extended light sources. Generalized light sources are light sources of zero étendue which are not parallel or point sources. Since étendue is conserved in a lighting system, for the double freeform lens with a point source of zero étendue, we can think of the intermediate light rays \hat{i} to originate from a generalized light source. We saw that we can only find a generating function for the second surface by writing down the characteristic function for the whole system, starting from the point source, so that the characteristic function T is independent from the source coordinates, i.e., $T(\mathbf{p}_t)$ in (8.23) is dependent on \mathbf{p}_t while we cannot eliminate \mathbf{p}_t in $T_2(\mathbf{p}_t, \mathbf{p}_t)$ in (8.18).

Thus, for systems with generalized light sources with direction vectors \hat{i} we would need to insert a surface in front of the light source such that it seems we consider a parallel or point source after all. We can compute this surface easily using the computation step for u in our numerical algorithm, since we know the mapping from \hat{s} to \hat{i} .

For extended light sources we arrive in a different realm. Extended light sources are light sources of nonzero étendue. Since étendue is conserved in a lighting system, we cannot convert such a light source back to a parallel or point source and we cannot find a characteristic function that is independent of the position and direction coordinates at the source. Our method using G-convex and G-concave functions would fail for this formulation and it is not straightforward to see how we can incorporate extended light sources in our framework. One could think of an extended light source as a collection of point sources with each point having a target intensity distribution. The total target intensity could then be calculated as the convolution of all these contributions. Another potential strategy to incorporate extended light sources is to use our numerical algorithm with a zero-étendue source and use an optimization procedure to iteratively modify the target distribution based on ray-tracing results with the extended source.

- Currently, my colleagues Vi and Teun are investigating whether we can combine the least-squares numerical procedures with optical phenomena such as scattering, Fresnel reflections and aberrations. For instance, to include scattering we can consider multiple rays reflecting/refracting off a surface in directions which deviate from the specular laws of reflection/refraction. In certain cases this yields a relatively simple convolution integral to incorporate the diffusion. One can then deconvolve the prescribed target intensity with the probability density function related to the scattering to get an intermediate specular distribution such that diffuse problems can be treated using the least-squares algorithms [88].
- In this thesis, we assume that the optical surfaces are G-convex/G-concave solutions, but are there other possibilities? For instance, we can also think of saddle-shaped surfaces or periodic arrays of G-convex, G-concave and/or saddle surfaces. In this thesis, we restricted ourselves to a positive Jacobian of the mapping $\det(Dm)$, but saddle-shaped solutions can be found by considering a negative Jacobian of the mapping. In this case, we arrive in the realm of hyperbolic second-order nonlinear PDEs, which are treated by my colleague Maikel. Our least-squares procedures fail for these equations. For parallel-to-far-field problems

with cost function $c(\mathbf{x}, \mathbf{y}) = -\mathbf{x} \cdot \mathbf{y}$, these equations have saddle-shaped solutions, but our algorithm attempts to find convex or concave solutions. Solutions to hyperbolic problems can be found using the method of characteristics [10].

- For the double freeform lens with a point source and far-field target considered in this thesis, we computed a G-concave first surface and G-convex second surface in Chapter 7. We note that we can also choose other combinations of G-convex and G-concave surfaces for the double freeform lens, which results in different designs. More investigation is necessary to determine which combinations are possible for different experiments.

We cannot always choose between a G-convex or G-concave solution for any particular source and target domain. The initial mapping \mathbf{m}^0 and mixed Hessian matrix $\mathbf{C}(\mathbf{x}, \mathbf{m}^0, u^0)$ should be such that the matrix $\mathbf{P} = \mathbf{C} \mathbf{D}\mathbf{m}^0$ is SND or SPD at all points in the domain. Hence, the initial map \mathbf{m}^0 and initial surface u^0 should be chosen carefully. Computing a double freeform lens further complicates these choices, since we also compute an intermediate optical map.

- Can we compute optical systems where some of the light rays follow a different path, e.g., part of the light rays are refracted and part of the light rays undergo TIR?
- Can we incorporate effects from wave optics in the numerical algorithm, such as coherence, dispersion and diffraction?

Hopefully, my current colleagues, supervisors, future students and perhaps myself will consider the above items in the future and keep exploring the possibilities of generated Jacobian equations in freeform optical design.

Appendix A

G-convex and G-concave Functions

A.1 G-convex and G-concave functions

For the majority of generating functions in Table 3.1 it holds that $G_w > 0$. For this case we show that we get a max/min pair in case of a G-convex solution u , and a min/max pair in case of a G-concave solution u . First, we derive two properties using the intermediate value theorem in Lemma A.1.1 and Lemma A.1.2. The proof for the max/min pair is detailed in Section A.1.1. A min/max pair can be derived analogously and the result is given in Section A.1.2. Note that we can repeat the derivations in Section A.1.1 and A.1.2 for $G_w < 0$, which we will not show in this appendix. This results in a max/max pair (G-convex and H-convex) or min/min (G-concave and H-concave) pair, respectively.

Lemma A.1.1. *Consider the variables $w_1, w_2 \in [0, \infty)$ and let G be a generating function $G : \mathcal{G} \subset \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}$ with $G_w > 0$. By the intermediate value theorem, we have $G(\mathbf{x}, \mathbf{y}, w_1) - G(\mathbf{x}, \mathbf{y}, w_2) = G_w(\mathbf{x}, \mathbf{y}, \xi)(w_1 - w_2)$, with $\xi \in \text{int}(w_1, w_2)$. Let $w_1 < w_2$, then: $G(\mathbf{x}, \mathbf{y}, w_1) - G(\mathbf{x}, \mathbf{y}, w_2) < 0$ since $G_w > 0$. We have shown that*

$$w_1 < w_2 \quad \implies \quad G(\mathbf{x}, \mathbf{y}, w_1) < G(\mathbf{x}, \mathbf{y}, w_2). \quad (\text{A.1a})$$

It also follows that

$$w_1 > w_2 \quad \implies \quad G(\mathbf{x}, \mathbf{y}, w_1) > G(\mathbf{x}, \mathbf{y}, w_2). \quad (\text{A.1b})$$

We define the function $H : \mathcal{H} \subset \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}$ such that for fixed $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ it holds that

$$u(\mathbf{x}) = G(\mathbf{x}, \mathbf{y}, w(\mathbf{y})) \iff w(\mathbf{y}) = H(\mathbf{x}, \mathbf{y}, u(\mathbf{x})), \quad (\text{A.2})$$

assuming a unique inverse G exists.

Lemma A.1.2. *Consider the variables $w_1, w_2 \in [0, \infty)$ and let G be a generating function $G : \mathcal{G} \subset \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}$ with $G_w > 0$. We let $w_1 = u(\mathbf{x})$ and $w_2 = w(\mathbf{y})$. Using (A.2) we have*

$$w_1 = G(\mathbf{x}, \mathbf{y}, w_2) \iff w_2 = H(\mathbf{x}, \mathbf{y}, w_1). \quad (\text{A.3a})$$

Hence,

$$w_1 = G(\mathbf{x}, \mathbf{y}, H(\mathbf{x}, \mathbf{y}, w_1)). \quad (\text{A.3b})$$

Differentiating with respect to w_1 gives

$$1 = G_w H_w \implies H_w = \frac{1}{G_w} > 0, \quad (\text{A.3c})$$

since $G_w > 0$. By the intermediate value theorem, the function H has the property $H(\mathbf{x}, \mathbf{y}, w_1) - H(\mathbf{x}, \mathbf{y}, w_2) = H_w(\mathbf{x}, \mathbf{y}, \xi)(w_1 - w_2)$, with $\xi \in \text{int}(w_1, w_2)$.

Let $w_1 < w_2$, then: $H(\mathbf{x}, \mathbf{y}, w_1) - H(\mathbf{x}, \mathbf{y}, w_2) < 0$. We have shown that

$$w_1 < w_2 \implies H(\mathbf{x}, \mathbf{y}, w_1) < H(\mathbf{x}, \mathbf{y}, w_2). \quad (\text{A.3d})$$

It also follows that

$$w_1 > w_2 \implies H(\mathbf{x}, \mathbf{y}, w_1) > H(\mathbf{x}, \mathbf{y}, w_2). \quad (\text{A.3e})$$

A.1.1 G-convex and H-concave pair

Claim A.1.1. *Let G be a generating function $G : \mathcal{G} \subset \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}$ with $G_w > 0$. If we can construct a function $H : \mathcal{H} \subset \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}$ such that for all $\mathbf{x} \in \mathcal{X}$ and for all $\mathbf{y} \in \mathcal{Y}$ it holds that*

$$u(\mathbf{x}) = G(\mathbf{x}, \mathbf{y}, w(\mathbf{y})) \iff w(\mathbf{y}) = H(\mathbf{x}, \mathbf{y}, u(\mathbf{x})), \quad (\text{A.4a})$$

then for all $\mathbf{x} \in \mathcal{X}$ and for all $\mathbf{y} \in \mathcal{Y}$ it holds that

$$u(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{Y}} G(\mathbf{x}, \mathbf{y}, w(\mathbf{y})) \iff w(\mathbf{y}) = \min_{\mathbf{x} \in \mathcal{X}} H(\mathbf{x}, \mathbf{y}, u(\mathbf{x})), \quad (\text{A.4b})$$

i.e., the G -convex solution $u(\mathbf{x})$ and H -concave solution $w(\mathbf{y})$ form a pair.

Proof. We define a G -convex solution $u(\mathbf{x})$ as follows:

Definition A.1.1. $\forall \mathbf{x} \in \mathcal{X} : u(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{Y}} G(\mathbf{x}, \mathbf{y}, w(\mathbf{y}))$.

Then it follows that

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y} : \quad u(x) \geq G(x, y, w(y)). \quad (\text{A.5a})$$

By Lemma A.1.2, stated in (A.3e), we have

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y} : \quad H(x, y, u(x)) \geq H(x, y, G(x, y, w(y))) \iff (\text{A.5b})$$

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y} : \quad H(x, y, u(x)) \geq w(y) \iff (\text{A.5c})$$

$$\forall y \in \mathcal{Y} : \quad w(y) \leq \min_{x \in \mathcal{X}} H(x, y, u(x)). \quad (\text{A.5d})$$

Using (A.2) we also have, by definition,

$$\forall y \in \mathcal{Y} : \quad w(y) \geq \min_{x \in \mathcal{X}} H(x, y, u(x)). \quad (\text{A.5e})$$

Hence,

$$\forall y \in \mathcal{Y} : \quad w(y) = \min_{x \in \mathcal{X}} H(x, y, u(x)). \quad (\text{A.5f})$$

In conclusion,

$$\forall x \in \mathcal{X} : \quad u(x) = \max_{y \in \mathcal{Y}} G(x, y, w(y)) \implies$$

$$\forall y \in \mathcal{Y} : \quad w(y) = \min_{x \in \mathcal{X}} H(x, y, u(x)). \quad (\text{A.5g})$$

To prove the converse, we define the H-concave solution $w(y)$ as

Definition A.1.2. $\forall y \in \mathcal{Y} : \quad w(y) = \min_{x \in \mathcal{X}} H(x, y, u(x)).$

Then it follows that

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y} : \quad w(y) \leq H(x, y, u(x)). \quad (\text{A.6a})$$

By Equation (A.1.1), stated in (A.1b), and a similar argument as above, we have

$$\forall x \in \mathcal{X} : \quad u(x) \geq \max_{y \in \mathcal{Y}} G(x, y, w(y)). \quad (\text{A.6b})$$

Using (A.2) we also have, by definition,

$$\forall x \in \mathcal{X} : \quad u(x) \leq \max_{y \in \mathcal{Y}} G(x, y, w(y)). \quad (\text{A.6c})$$

Hence,

$$\forall x \in \mathcal{X} : \quad u(x) = \max_{y \in \mathcal{Y}} G(x, y, w(y)). \quad (\text{A.6d})$$

In conclusion,

$$\begin{aligned} \forall \mathbf{y} \in \mathcal{Y} : \quad w(\mathbf{y}) &= \min_{x \in \mathcal{X}} H(x, \mathbf{y}, u(x)) \quad \implies \\ \forall x \in \mathcal{X} : \quad u(x) &= \max_{y \in \mathcal{Y}} G(x, \mathbf{y}, w(\mathbf{y})). \end{aligned} \quad (\text{A.6e})$$

Combining (A.5g) and (A.6e) gives

$$\begin{aligned} \forall x \in \mathcal{X} : \quad u(x) &= \max_{y \in \mathcal{Y}} G(x, \mathbf{y}, w(\mathbf{y})) \quad \iff \\ \forall \mathbf{y} \in \mathcal{Y} : \quad w(\mathbf{y}) &= \min_{x \in \mathcal{X}} H(x, \mathbf{y}, u(x)), \end{aligned} \quad (\text{A.7a})$$

i.e., the G-convex solution $u(x)$ and H-concave solution $w(\mathbf{y})$ form a pair. \square

A.1.2 G-concave and H-convex pair

The proof for a G-concave and H-convex pair is similar to the proof in the previous section. The conclusion is that if we can construct an inverse function $H : \mathcal{H} \subset \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}$ such that for all $x \in \mathcal{X}$ and for all $\mathbf{y} \in \mathcal{Y}$ it holds that

$$u(x) = G(x, \mathbf{y}, w(\mathbf{y})) \iff w(\mathbf{y}) = H(x, \mathbf{y}, u(x)), \quad (\text{A.8a})$$

then for all $x \in \mathcal{X}$ and for all $\mathbf{y} \in \mathcal{Y}$ it holds that

$$u(x) = \min_{y \in \mathcal{Y}} G(x, \mathbf{y}, w(\mathbf{y})) \iff w(\mathbf{y}) = \max_{x \in \mathcal{X}} H(x, \mathbf{y}, u(x)), \quad (\text{A.8b})$$

i.e., the G-concave solution $u(x)$ and H-convex solution $w(\mathbf{y})$ form a pair.

Note that we can repeat the derivations in Section A.1.1 and A.1.2 for $G_w < 0$. This results in a max/max pair (G-convex and H-convex) or min/min (G-concave and H-concave) pair, respectively.

Appendix B

The Finite Volume Method

B.1 The finite volume method in Cartesian coordinates

In this section, we show how to use the finite volume method to solve the boundary value problem for m in (6.79).

We write $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2]$ and $\mathbf{C} = [c_1, c_2]$ using column-vector notation, with $\mathbf{p}_1, \mathbf{p}_2, c_1, c_2 \in \mathbb{R}^2$. Writing out the boundary value problem in (6.79) gives

$$\begin{aligned} \frac{\partial}{\partial x_1} \left[|c_1|^2 \frac{\partial m_1}{\partial x_1} + c_1 \cdot c_2 \frac{\partial m_2}{\partial x_1} \right] + \frac{\partial}{\partial x_2} \left[|c_1|^2 \frac{\partial m_1}{\partial x_2} + c_1 \cdot c_2 \frac{\partial m_2}{\partial x_2} \right] \\ = \frac{\partial}{\partial x_1} (c_1 \cdot \mathbf{p}_1) + \frac{\partial}{\partial x_2} (c_1 \cdot \mathbf{p}_2), \end{aligned} \quad (\text{B.1a})$$

$$\begin{aligned} \frac{\partial}{\partial x_1} \left[c_1 \cdot c_2 \frac{\partial m_1}{\partial x_1} + |c_2|^2 \frac{\partial m_2}{\partial x_1} \right] + \frac{\partial}{\partial x_2} \left[c_1 \cdot c_2 \frac{\partial m_1}{\partial x_2} + |c_2|^2 \frac{\partial m_2}{\partial x_2} \right] \\ = \frac{\partial}{\partial x_1} (c_2 \cdot \mathbf{p}_1) + \frac{\partial}{\partial x_2} (c_2 \cdot \mathbf{p}_2), \end{aligned} \quad (\text{B.1b})$$

and the boundary equations

$$(1 - \alpha) m_1 + \alpha (|c_1|^2 \nabla m_1 \cdot \hat{\mathbf{n}} + c_1 \cdot c_2 \nabla m_2 \cdot \hat{\mathbf{n}}) = (1 - \alpha) b_1 + \alpha c_1 \cdot \mathbf{P} \hat{\mathbf{n}}, \quad (\text{B.1c})$$

$$(1 - \alpha) m_2 + \alpha (c_1 \cdot c_2 \nabla m_1 \cdot \hat{\mathbf{n}} + |c_2|^2 \nabla m_2 \cdot \hat{\mathbf{n}}) = (1 - \alpha) b_2 + \alpha c_2 \cdot \mathbf{P} \hat{\mathbf{n}}. \quad (\text{B.1d})$$

We let $\mathbf{F} = \mathbf{C}^\top \mathbf{C} \mathbf{D} \mathbf{m}$ and $\mathbf{R} = \mathbf{C}^\top \mathbf{P}$, so that (6.79) can be written as

$$\nabla \cdot \mathbf{F} = \nabla \cdot \mathbf{R}, \quad (\text{B.2a})$$

$$(1 - \alpha) \mathbf{m} + \alpha \mathbf{F} \cdot \hat{\mathbf{n}} = (1 - \alpha) \mathbf{b} + \alpha \mathbf{R} \cdot \hat{\mathbf{n}}. \quad (\text{B.2b})$$

We introduce the row vectors f_1^T, f_2^T of F and r_1^T, r_2^T of R as

$$\begin{aligned}
 F &= \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{pmatrix} = \begin{pmatrix} f_1^T \\ f_2^T \end{pmatrix}, \\
 f_1 &= \begin{pmatrix} f_{11} \\ f_{12} \end{pmatrix} = \begin{pmatrix} |c_1|^2 \frac{\partial m_1}{\partial x_1} + c_1 \cdot c_2 \frac{\partial m_2}{\partial x_1} \\ |c_1|^2 \frac{\partial m_1}{\partial x_2} + c_1 \cdot c_2 \frac{\partial m_2}{\partial x_2} \end{pmatrix}, \\
 f_2 &= \begin{pmatrix} f_{21} \\ f_{22} \end{pmatrix} = \begin{pmatrix} c_1 \cdot c_2 \frac{\partial m_1}{\partial x_1} + |c_2|^2 \frac{\partial m_2}{\partial x_1} \\ c_1 \cdot c_2 \frac{\partial m_1}{\partial x_2} + |c_2|^2 \frac{\partial m_2}{\partial x_2} \end{pmatrix}, \\
 R &= \begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix} = \begin{pmatrix} r_1^T \\ r_2^T \end{pmatrix}, \\
 r_1 &= \begin{pmatrix} r_{11} \\ r_{12} \end{pmatrix} = \begin{pmatrix} c_1 \cdot p_1 \\ c_1 \cdot p_2 \end{pmatrix}, \quad r_2 = \begin{pmatrix} r_{21} \\ r_{22} \end{pmatrix} = \begin{pmatrix} c_2 \cdot p_1 \\ c_2 \cdot p_2 \end{pmatrix}. \quad (\text{B.3})
 \end{aligned}$$

We first focus on Equation (B.1a). Using (B.3) we can write Equation (B.1a) as

$$\frac{\partial f_{11}}{\partial x_1} + \frac{\partial f_{12}}{\partial x_2} = \frac{\partial r_{11}}{\partial x_1} + \frac{\partial r_{12}}{\partial x_2}, \quad (\text{B.4})$$

and more compactly in the divergence form

$$\nabla \cdot f_1 = \nabla \cdot r_1. \quad (\text{B.5})$$

Integrating over a control volume Ω_C and using Gauss's theorem gives

$$\oint_{\partial\Omega_C} f_1 \cdot \hat{n} \, ds = \int_{\Omega_C} \nabla \cdot f_1 \, dA = \int_{\Omega_C} \nabla \cdot r_1 \, dA = \oint_{\partial\Omega_C} r_1 \cdot \hat{n} \, ds, \quad (\text{B.6})$$

where \hat{n} is the unit outward normal and $\partial\Omega_C$ is oriented counterclockwise.

We discretize the source domain \mathcal{X} as in (6.12) and introduce $x_{ij} = (x_C)$. We consider the control volume

$$\Omega_C = [x_{1,w}, x_{1,e}] \times [x_{2,s}, x_{2,n}], \quad (\text{B.7})$$

where $x_{1,w}$ is the x_1 -coordinate of the western cell face Γ_w , i.e.

$$x_{1,w} = \frac{x_{1,C} + x_{1,W}}{2}, \quad (\text{B.8})$$

etc. The boundary of the control volume is divided into four parts such that $\partial\Omega_C = \Gamma_e \cup \Gamma_n \cup \Gamma_w \cup \Gamma_s$. Integrating Equation (B.5) over the control volume

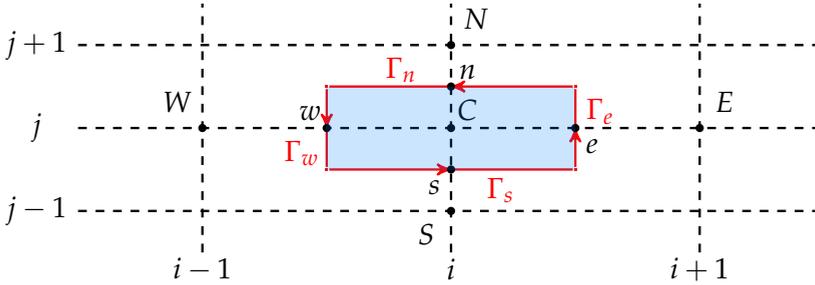


Figure B.1: Control volume for a cell-centered finite volume method. Note the difference between big E and small e , etc.

Ω_C using Gauss's theorem gives

$$\begin{aligned} & \int_{\Gamma_e} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds + \int_{\Gamma_n} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds + \int_{\Gamma_w} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds + \int_{\Gamma_s} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds \\ &= \int_{\Gamma_e} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds + \int_{\Gamma_n} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds + \int_{\Gamma_w} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds + \int_{\Gamma_s} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds. \end{aligned} \quad (\text{B.9})$$

The normals $\hat{\mathbf{n}}$ at each of the boundary segments are

$$\begin{aligned} \Gamma_e : \quad \hat{\mathbf{n}} &= \hat{\mathbf{e}}_{x_1}, & \Gamma_n : \quad \hat{\mathbf{n}} &= \hat{\mathbf{e}}_{x_2}, \\ \Gamma_w : \quad \hat{\mathbf{n}} &= -\hat{\mathbf{e}}_{x_1}, & \Gamma_s : \quad \hat{\mathbf{n}} &= -\hat{\mathbf{e}}_{x_2}. \end{aligned} \quad (\text{B.10})$$

Taking a midpoint approximation and using $\hat{\mathbf{n}} = \hat{\mathbf{e}}_{x_1}$ results in the approximation of the first integral in (B.9)

$$\int_{\Gamma_e} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds = \int_{x_{2,s}}^{x_{2,n}} f_{11}(x_{1,e}, x_2) \, dx_2 \approx F_{11,e} h_2, \quad (\text{B.11})$$

where $F_{11,e}$ is the numerical approximation of the flux term f_{11} at the interface point x_e and h_2 is the grid size as defined in Equation (6.12b). Analogously, we derive

$$\begin{aligned} \int_{\Gamma_n} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds &= - \int_{x_{1,e}}^{x_{1,w}} f_{12}(x_1, x_{2,n}) \, dx_1 \approx F_{12,n} h_1, \\ \int_{\Gamma_w} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds &= - \int_{x_{2,n}}^{x_{2,s}} -f_{11}(x_{1,w}, x_2) \, dx_2 \approx -F_{11,w} h_2, \\ \int_{\Gamma_s} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds &= \int_{x_{1,w}}^{x_{1,e}} -f_{12}(x_1, x_{2,s}) \, dx_1 \approx -F_{12,s} h_1, \end{aligned} \quad (\text{B.12})$$

with h_1 as defined in (6.12a). Similarly,

$$\begin{aligned}
 \int_{\Gamma_e} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds &= \int_{x_{2,s}}^{x_{2,n}} r_{11}(x_{1,e}, x_2) \, dx_2 \approx R_{11,e} h_2, \\
 \int_{\Gamma_n} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds &= - \int_{x_{1,e}}^{x_{1,w}} r_{12}(x_1, x_{2,n}) \, dx_1 \approx R_{12,n} h_1, \\
 \int_{\Gamma_w} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds &= - \int_{x_{2,n}}^{x_{2,s}} -r_{11}(x_{1,w}, x_2) \, dx_2 \approx -R_{11,w} h_2, \\
 \int_{\Gamma_s} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds &= \int_{x_{1,w}}^{x_{1,e}} -r_{12}(x_1, x_{2,s}) \, dx_1 \approx -R_{12,s} h_1,
 \end{aligned} \tag{B.13}$$

where $R_{11,e}$ is the numerical approximation of the flux term r_{11} at the interface point \mathbf{x}_e , etc.

Substituting (B.12) and (B.13) into (B.9) gives

$$h_2 (F_{11,e} - F_{11,w}) + h_1 (F_{12,n} - F_{12,s}) \approx h_2 (R_{11,e} - R_{11,w}) + h_1 (R_{12,n} - R_{12,s}). \tag{B.14}$$

We use central differences to approximate the first-order derivatives in the numerical approximation of the flux terms. For instance,

$$\begin{aligned}
 F_{11,e} &= |\mathbf{c}_1|_e^2 \frac{m_1(\mathbf{x}_E) - m_1(\mathbf{x}_C)}{h_1} + (\mathbf{c}_1 \cdot \mathbf{c}_2)_e \frac{m_2(\mathbf{x}_E) - m_2(\mathbf{x}_C)}{h_1}, \\
 F_{11,w} &= |\mathbf{c}_1|_w^2 \frac{m_1(\mathbf{x}_C) - m_1(\mathbf{x}_W)}{h_1} + (\mathbf{c}_1 \cdot \mathbf{c}_2)_w \frac{m_2(\mathbf{x}_C) - m_2(\mathbf{x}_W)}{h_1}, \\
 F_{12,n} &= |\mathbf{c}_1|_n^2 \frac{m_1(\mathbf{x}_N) - m_1(\mathbf{x}_C)}{h_2} + (\mathbf{c}_1 \cdot \mathbf{c}_2)_n \frac{m_2(\mathbf{x}_N) - m_2(\mathbf{x}_C)}{h_2}, \\
 F_{12,s} &= |\mathbf{c}_1|_s^2 \frac{m_1(\mathbf{x}_C) - m_1(\mathbf{x}_S)}{h_2} + (\mathbf{c}_1 \cdot \mathbf{c}_2)_s \frac{m_2(\mathbf{x}_C) - m_2(\mathbf{x}_S)}{h_2}, \\
 R_{11,e} &= (\mathbf{c}_1 \cdot \mathbf{p}_1)_e, \quad R_{11,w} = (\mathbf{c}_1 \cdot \mathbf{p}_1)_w, \\
 R_{12,n} &= (\mathbf{c}_1 \cdot \mathbf{p}_2)_n, \quad R_{12,s} = (\mathbf{c}_1 \cdot \mathbf{p}_2)_s.
 \end{aligned} \tag{B.15}$$

Substituting (B.15) into (B.14) and dividing by $h_1 h_2$ gives

$$\begin{aligned}
 & \frac{|c_1|_e^2}{h_1^2} m_1(x_E) + \frac{|c_1|_w^2}{h_1^2} m_1(x_W) + \frac{|c_1|_n^2}{h_2^2} m_1(x_N) + \frac{|c_1|_s^2}{h_2^2} m_1(x_S) \\
 & - \left(\frac{|c_1|_e^2}{h_1^2} + \frac{|c_1|_w^2}{h_1^2} + \frac{|c_1|_n^2}{h_2^2} + \frac{|c_1|_s^2}{h_2^2} \right) m_1(x_C) \\
 & + \frac{(c_1 \cdot c_2)_e}{h_1^2} m_2(x_E) + \frac{(c_1 \cdot c_2)_w}{h_1^2} m_2(x_W) \\
 & + \frac{(c_1 \cdot c_2)_n}{h_2^2} m_2(x_N) + \frac{(c_1 \cdot c_2)_s}{h_2^2} m_2(x_S) \\
 & - \left(\frac{(c_1 \cdot c_2)_e}{h_1^2} + \frac{(c_1 \cdot c_2)_w}{h_1^2} + \frac{(c_1 \cdot c_2)_n}{h_2^2} + \frac{(c_1 \cdot c_2)_s}{h_2^2} \right) m_2(x_C) \\
 & = \frac{1}{h_1} [(c_1 \cdot p_1)_e - (c_1 \cdot p_1)_w] + \frac{1}{h_2} [(c_1 \cdot p_2)_n - (c_1 \cdot p_2)_s]. \quad (B.16)
 \end{aligned}$$

For the second equation of the system (B.1b), we do the same to obtain

$$\begin{aligned}
 & \frac{(c_1 \cdot c_2)_e}{h_1^2} m_1(x_E) + \frac{(c_1 \cdot c_2)_w}{h_1^2} m_1(x_W) + \frac{(c_1 \cdot c_2)_n}{h_2^2} m_1(x_N) + \frac{(c_1 \cdot c_2)_s}{h_2^2} m_1(x_S) \\
 & - \left(\frac{(c_1 \cdot c_2)_e}{h_1^2} + \frac{(c_1 \cdot c_2)_w}{h_1^2} + \frac{(c_1 \cdot c_2)_n}{h_2^2} + \frac{(c_1 \cdot c_2)_s}{h_2^2} \right) m_1(x_C) \\
 & + \frac{|c_2|_e^2}{h_1^2} m_2(x_E) + \frac{|c_2|_w^2}{h_1^2} m_2(x_W) + \frac{|c_2|_n^2}{h_2^2} m_2(x_N) + \frac{|c_2|_s^2}{h_2^2} m_2(x_S) \\
 & - \left(\frac{|c_2|_e^2}{h_1^2} + \frac{|c_2|_w^2}{h_1^2} + \frac{|c_2|_n^2}{h_2^2} + \frac{|c_2|_s^2}{h_2^2} \right) m_2(x_C) \\
 & = \frac{1}{h_1} [(c_2 \cdot p_1)_e - (c_2 \cdot p_1)_w] + \frac{1}{h_2} [(c_2 \cdot p_2)_n - (c_2 \cdot p_2)_s]. \quad (B.17)
 \end{aligned}$$

At the boundaries, we use the boundary equations in Equation (B.1c) and (B.1d).

B.1.1 Incorporating boundary conditions

Consider the left boundary without corner points, such that we have $i = 1$ with $j = 2, \dots, N_2 - 1$. We consider the control volume

$$\Omega_C = [x_{1,C}, x_{1,e}] \times [x_{2,s}, x_{2,n}], \quad (B.18)$$

as shown in Figure B.2.

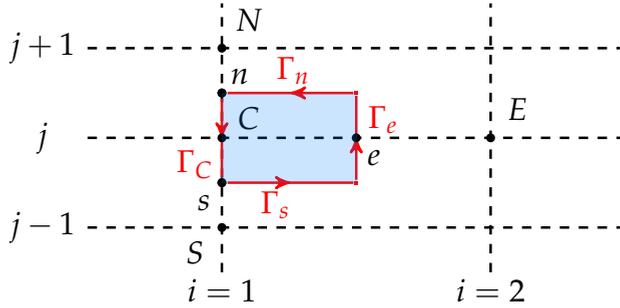


Figure B.2: Control volume for the left boundary of a cell-centered finite volume method.

The boundary of the control volume is divided into four parts such that $\partial\Omega_C = \Gamma_e \cup \Gamma_n \cup \Gamma_C \cup \Gamma_s$. Integrating Equation (B.5) over the control volume Ω_C using Gauss's theorem gives

$$\begin{aligned} & \int_{\Gamma_e} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds + \int_{\Gamma_n} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds + \int_{\Gamma_C} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds + \int_{\Gamma_s} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds \\ &= \int_{\Gamma_e} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds + \int_{\Gamma_n} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds + \int_{\Gamma_C} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds + \int_{\Gamma_s} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds. \end{aligned} \quad (\text{B.19})$$

This equation is similar to (B.9) but with Γ_w replaced by Γ_C . Taking approximations and substituting $\hat{\mathbf{n}}$ at each of the boundaries, where Γ_e , Γ_n and Γ_s are as in (B.10) and at Γ_C we have $\hat{\mathbf{n}} = (-1, 0)$, results in the approximation of the integrals

$$\begin{aligned} \int_{\Gamma_e} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds &= \int_{x_{2,s}}^{x_{2,n}} f_{11}(x_{1,e}, x_2) \, dx_2 \approx F_{11,e} h_2, \\ \int_{\Gamma_n} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds &= - \int_{x_{1,e}}^{x_{1,C}} f_{12}(x_1, x_{2,n}) \, dx_1 \approx \frac{F_{12,n} h_1}{2}, \quad (\text{no midpoint}) \\ \int_{\Gamma_C} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds &= - \int_{x_{2,n}}^{x_{2,s}} -f_{11}(x_{1,C}, x_2) \, dx_2 \approx -F_{11,C} h_2, \\ \int_{\Gamma_s} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds &= \int_{x_{1,C}}^{x_{1,e}} -f_{12}(x_1, x_{2,s}) \, dx_1 \approx -\frac{F_{12,s} h_1}{2}, \quad (\text{no midpoint}). \end{aligned} \quad (\text{B.20})$$

Similarly,

$$\begin{aligned}
 \int_{\Gamma_e} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds &= \int_{x_{2,s}}^{x_{2,n}} r_{11}(x_{1,e}, x_2) \, dx_2 \approx R_{11,e} h_2, \\
 \int_{\Gamma_n} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds &= - \int_{x_{1,e}}^{x_{1,C}} r_{12}(x_1, x_{2,n}) \, dx_1 \approx \frac{R_{12,n} h_1}{2}, \\
 \int_{\Gamma_C} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds &= - \int_{x_{2,n}}^{x_{2,s}} -r_{11}(x_{1,C}, x_2) \, dx_2 \approx -R_{11,C} h_2, \\
 \int_{\Gamma_s} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds &= \int_{x_{1,C}}^{x_{1,e}} -r_{12}(x_1, x_{2,s}) \, dx_1 \approx -\frac{R_{12,s} h_1}{2}.
 \end{aligned} \tag{B.21}$$

Substituting (B.20) and (B.21) into (B.19) gives

$$h_2 (F_{11,e} - F_{11,C}) + \frac{h_1}{2} (F_{12,n} - F_{12,s}) \approx h_2 (R_{11,e} - R_{11,C}) + \frac{h_1}{2} (R_{12,n} - R_{12,s}). \tag{B.22}$$

We use central differences to approximate the first-order derivatives in the numerical approximation of the flux terms. Hence, we obtain

$$\begin{aligned}
 F_{11,e} &= |\mathbf{c}_1|_e^2 \frac{m_1(\mathbf{x}_E) - m_1(\mathbf{x}_C)}{h_1} + (\mathbf{c}_1 \cdot \mathbf{c}_2)_e \frac{m_2(\mathbf{x}_E) - m_2(\mathbf{x}_C)}{h_1}, \\
 F_{12,n} &= |\mathbf{c}_1|_n^2 \frac{m_1(\mathbf{x}_N) - m_1(\mathbf{x}_C)}{h_2} + (\mathbf{c}_1 \cdot \mathbf{c}_2)_n \frac{m_2(\mathbf{x}_N) - m_2(\mathbf{x}_C)}{h_2}, \\
 F_{12,s} &= |\mathbf{c}_1|_s^2 \frac{m_1(\mathbf{x}_C) - m_1(\mathbf{x}_S)}{h_2} + (\mathbf{c}_1 \cdot \mathbf{c}_2)_s \frac{m_2(\mathbf{x}_C) - m_2(\mathbf{x}_S)}{h_2}, \\
 R_{11,e} &= (\mathbf{c}_1 \cdot \mathbf{p}_1)_e, \quad R_{11,C} = (\mathbf{c}_1 \cdot \mathbf{p}_1)_C, \\
 R_{12,n} &= (\mathbf{c}_1 \cdot \mathbf{p}_2)_n, \quad R_{12,s} = (\mathbf{c}_1 \cdot \mathbf{p}_2)_s,
 \end{aligned} \tag{B.23}$$

which is similar to (B.15) but leaving out $F_{11,w}$ and $R_{11,w}$ and adding $R_{11,C}$. We replace $F_{11,C}$ in (B.22) using the boundary equation (B.1c). Hence, with $\hat{\mathbf{n}} = (-1, 0)$ we get that the first boundary equation becomes

$$(1 - \alpha) m_1(\mathbf{x}_C) - \alpha F_{11,C} = (1 - \alpha) b_{1,C} - \alpha (\mathbf{c}_1 \cdot \mathbf{p}_1)_C, \tag{B.24}$$

and solving for $F_{11,C}$ gives

$$F_{11,C} = \left(\frac{1}{\alpha} - 1 \right) (m_1(\mathbf{x}_C) - b_{1,C}) + (\mathbf{c}_1 \cdot \mathbf{p}_1)_C. \tag{B.25}$$

Substituting (B.23) and (B.25) into (B.22) and dividing by $h_1 h_2$ gives

$$\begin{aligned}
& \frac{|c_1|_e^2}{h_1^2} m_1(x_E) + \frac{|c_1|_n^2}{2h_2^2} m_1(x_N) + \frac{|c_1|_s^2}{2h_2^2} m_1(x_S) - \left(\frac{|c_1|_e^2}{h_1^2} + \frac{|c_1|_n^2}{2h_2^2} + \frac{|c_1|_s^2}{2h_2^2} \right) m_1(x_C) \\
& + \frac{(c_1 \cdot c_2)_e}{h_1^2} m_2(x_E) + \frac{(c_1 \cdot c_2)_n}{2h_2^2} m_2(x_N) + \frac{(c_1 \cdot c_2)_s}{2h_2^2} m_2(x_S) \\
& - \left(\frac{(c_1 \cdot c_2)_e}{h_1^2} + \frac{(c_1 \cdot c_2)_n}{2h_2^2} + \frac{(c_1 \cdot c_2)_s}{2h_2^2} \right) m_2(x_C) - \frac{1}{h_1} \left(\frac{1}{\alpha} - 1 \right) m_1(x_C) \\
& = \frac{1}{h_1} [(c_1 \cdot p_1)_e - (c_1 \cdot p_1)_C] + \frac{1}{2h_2} [(c_1 \cdot p_2)_n - (c_1 \cdot p_2)_s] \\
& - \frac{1}{h_1} \left(\frac{1}{\alpha} - 1 \right) b_{1,C} + \frac{1}{h_1} (c_1 \cdot p_1)_C. \tag{B.26}
\end{aligned}$$

For the second equation of the system (B.1b) we follow the same procedure and we use boundary equation (B.1d) to obtain

$$\begin{aligned}
& \frac{|c_2|_e^2}{h_1^2} m_2(x_E) + \frac{|c_2|_n^2}{2h_2^2} m_2(x_N) + \frac{|c_2|_s^2}{2h_2^2} m_2(x_S) - \left(\frac{|c_2|_e^2}{h_1^2} + \frac{|c_2|_n^2}{2h_2^2} + \frac{|c_2|_s^2}{2h_2^2} \right) m_2(x_C) \\
& + \frac{(c_1 \cdot c_2)_e}{h_1^2} m_1(x_E) + \frac{(c_1 \cdot c_2)_n}{2h_2^2} m_1(x_N) + \frac{(c_1 \cdot c_2)_s}{2h_2^2} m_1(x_S) \\
& - \left(\frac{(c_1 \cdot c_2)_e}{h_1^2} + \frac{(c_1 \cdot c_2)_n}{2h_2^2} + \frac{(c_1 \cdot c_2)_s}{2h_2^2} \right) m_1(x_C) - \frac{1}{h_1} \left(\frac{1}{\alpha} - 1 \right) m_2(x_C) \\
& = \frac{1}{h_1} [(c_2 \cdot p_1)_e - (c_2 \cdot p_1)_C] + \frac{1}{2h_2} [(c_2 \cdot p_2)_n - (c_2 \cdot p_2)_s] \\
& - \frac{1}{h_1} \left(\frac{1}{\alpha} - 1 \right) b_{2,C} + \frac{1}{h_1} (c_2 \cdot p_1)_C. \tag{B.27}
\end{aligned}$$

We repeat this procedure for all boundaries.

For corner points, e.g., the bottom-left corner $i = 1, j = 1$, we perform a similar procedure, but now with a quarter control volume

$$\Omega_C = [x_{1,C}, x_{1,e}] \times [x_{2,C}, x_{2,n}]. \tag{B.28}$$

The boundary of the control volume is divided into four parts such that $\partial\Omega_C = \Gamma_e \cup \Gamma_n \cup \Gamma_{C,w} \cup \Gamma_{C,s}$, where $\Gamma_{C,w}$ is the segment connecting x_n and x_C , and $\Gamma_{C,s}$ is the segment connecting x_C and x_e . The derivation of the equations is analogous to the ones presented above.

We solve the linear system using Matlab's *mldivide* for the first component m_1 of m . The linear system results from the domain equations (B.16) for all interior grid points $x_{ij} \in \mathcal{X}$, the boundary equations for all grid points on the boundary $x_{ij} \in \partial\mathcal{X}$, e.g. (B.26) for the left boundary, and the equations for all

four corner points. In these equations we substitute the value of m_2 from the previous iteration. Subsequently, using the new m_1 we solve the linear system for the second component m_2 , resulting from the domain equations (B.17), boundary equations such as (B.27) and corner equations. We compute \mathbf{m}^{n+1} in one step, i.e., we do not perform multiple iterations where we substitute the newly found m_2 into the equation for m_1 . We experimented with multiple iterations until this resubstitution barely changes m_1 and m_2 (with a tolerance of 10^{-8}). Performing one iteration is sufficient for most example problems.

B.2 The finite volume method in polar coordinates

In this section, we show how to use the finite volume method to solve the boundary value problem for \mathbf{m} in (6.99).

Using the cost function matrix $\mathbf{C}(\omega, \mathbf{m}(\omega))$ in (6.94a) and $\mathbf{P}(\omega)$ in the matrix equation (6.95), we denote

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}, \quad \mathbf{P} = \begin{pmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{pmatrix}, \quad (\text{B.29a})$$

and

$$\mathbf{c}_1 = c_{11} \hat{\mathbf{e}}_\rho + c_{12} \hat{\mathbf{e}}_\zeta, \quad \mathbf{c}_2 = c_{21} \hat{\mathbf{e}}_\rho + c_{22} \hat{\mathbf{e}}_\zeta, \quad (\text{B.29b})$$

$$\mathbf{p}_1 = p_{11} \hat{\mathbf{e}}_\rho + p_{12} \hat{\mathbf{e}}_\zeta, \quad \mathbf{p}_2 = p_{12} \hat{\mathbf{e}}_\rho + p_{22} \hat{\mathbf{e}}_\zeta, \quad (\text{B.29c})$$

where $\hat{\mathbf{e}}_\rho = \cos(\zeta) \hat{\mathbf{e}}_{x_1} + \sin(\zeta) \hat{\mathbf{e}}_{x_2}$ and $\hat{\mathbf{e}}_\zeta = -\sin(\zeta) \hat{\mathbf{e}}_{x_1} + \cos(\zeta) \hat{\mathbf{e}}_{x_2}$. Note that \mathbf{P} is symmetric while \mathbf{C} may not be symmetric. In a polar basis, we can rewrite (6.99) as

$$\nabla \cdot \mathbf{F} = \nabla \cdot \mathbf{r}, \quad (\text{B.30})$$

with the divergence operator as described in (6.100) and

$$\mathbf{F} = \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{pmatrix}, \quad \mathbf{r} = \begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix}. \quad (\text{B.31a})$$

This is a system of two equations

$$\nabla \cdot \mathbf{f}_1 = \nabla \cdot \mathbf{r}_1, \quad (\text{B.32a})$$

$$\nabla \cdot \mathbf{f}_2 = \nabla \cdot \mathbf{r}_2, \quad (\text{B.32b})$$

where

$$\mathbf{f}_1 = f_{11} \hat{\mathbf{e}}_\rho + f_{12} \hat{\mathbf{e}}_\zeta, \quad \mathbf{f}_2 = f_{21} \hat{\mathbf{e}}_\rho + f_{22} \hat{\mathbf{e}}_\zeta, \quad (\text{B.33a})$$

$$\mathbf{r}_1 = r_{11} \hat{\mathbf{e}}_\rho + r_{12} \hat{\mathbf{e}}_\zeta, \quad \mathbf{r}_2 = r_{21} \hat{\mathbf{e}}_\rho + r_{22} \hat{\mathbf{e}}_\zeta. \quad (\text{B.33b})$$

The components of F and r read

$$f_{11} = |c_1|^2 \frac{\partial m_1}{\partial \rho} + c_1 \cdot c_2 \frac{\partial m_2}{\partial \rho}, \quad f_{12} = \frac{|c_1|^2}{\rho} \frac{\partial m_1}{\partial \zeta} + \frac{c_1 \cdot c_2}{\rho} \frac{\partial m_2}{\partial \zeta}, \quad (\text{B.34a})$$

$$f_{21} = |c_2|^2 \frac{\partial m_2}{\partial \rho} + c_1 \cdot c_2 \frac{\partial m_1}{\partial \rho}, \quad f_{22} = \frac{|c_2|^2}{\rho} \frac{\partial m_2}{\partial \zeta} + \frac{c_1 \cdot c_2}{\rho} \frac{\partial m_1}{\partial \zeta}, \quad (\text{B.34b})$$

$$r_{11} = c_1 \cdot p_1, \quad r_{12} = c_1 \cdot p_2, \quad (\text{B.34c})$$

$$r_{21} = c_2 \cdot p_1, \quad r_{22} = c_2 \cdot p_2. \quad (\text{B.34d})$$

We first consider (B.32a). Integrating over the control volume $\Omega_{i,j}$ and using Gauss's theorem gives

$$\oint_{\partial\Omega_{i,j}} f_1 \cdot \hat{n} \, ds = \int_{\Omega_{i,j}} \nabla \cdot f_1 \, dA = \int_{\Omega_{i,j}} \nabla \cdot r_1 \, dA = \oint_{\partial\Omega_{i,j}} r_1 \cdot \hat{n} \, ds, \quad (\text{B.35})$$

where \hat{n} is the unit outward normal and $\partial\Omega_{i,j}$ is oriented counterclockwise.

We cover the source domain $\mathcal{X} = \mathcal{D}_R$, a circle with radius R , with a polar coordinate grid. Let N_ρ and N_ζ be the number of grid points along the ρ - and ζ -coordinate lines, respectively. We number the grid points

$$\begin{aligned} \rho_i &= i h_\rho, & i &= 0, \dots, N_\rho, & \text{where} & & h_\rho &= \frac{R}{N_\rho}, \\ \zeta_j &= (j-1) h_\zeta, & j &= 1, \dots, N_\zeta, & \text{where} & & h_\zeta &= \frac{2\pi}{N_\zeta}. \end{aligned}$$

Hence, the boundary $\partial\mathcal{X} = \partial\mathcal{D}_R$ is discretized by (ρ_{N_ρ}, ζ_j) , $1 \leq j \leq N_\zeta$. For $i = 2, \dots, N_\rho - 1$, $j = 1, \dots, N_\zeta$, we consider the control volume with boundary

$$\partial\Omega_{i,j} = \Gamma_{i+\frac{1}{2},j} \cup \Gamma_{i,j+\frac{1}{2}} \cup \Gamma_{i-\frac{1}{2},j} \cup \Gamma_{i,j-\frac{1}{2}}, \quad (\text{B.36})$$

as shown in Figure B.3.

Integrating (B.32a) over the control volume $\Omega_{i,j}$ using Gauss's theorem gives

$$\begin{aligned} & \int_{\Gamma_{i+\frac{1}{2},j}} f_1 \cdot \hat{n} \, ds + \int_{\Gamma_{i,j+\frac{1}{2}}} f_1 \cdot \hat{n} \, ds + \int_{\Gamma_{i-\frac{1}{2},j}} f_1 \cdot \hat{n} \, ds + \int_{\Gamma_{i,j-\frac{1}{2}}} f_1 \cdot \hat{n} \, ds \\ &= \int_{\Gamma_{i+\frac{1}{2},j}} r_1 \cdot \hat{n} \, ds + \int_{\Gamma_{i,j+\frac{1}{2}}} r_1 \cdot \hat{n} \, ds + \int_{\Gamma_{i-\frac{1}{2},j}} r_1 \cdot \hat{n} \, ds + \int_{\Gamma_{i,j-\frac{1}{2}}} r_1 \cdot \hat{n} \, ds. \end{aligned} \quad (\text{B.37})$$

The normals \hat{n} at each of the boundary segments are

$$\begin{aligned} \Gamma_{i+\frac{1}{2},j} : \hat{n} &= \hat{e}_{\rho_{i+\frac{1}{2},j}}, & \Gamma_{i,j+\frac{1}{2}} : \hat{n} &= \hat{e}_{\zeta_{i,j+\frac{1}{2}}}, \\ \Gamma_{i-\frac{1}{2},j} : \hat{n} &= -\hat{e}_{\rho_{i-\frac{1}{2},j}}, & \Gamma_{i,j-\frac{1}{2}} : \hat{n} &= -\hat{e}_{\zeta_{i,j-\frac{1}{2}}}. \end{aligned} \quad (\text{B.38})$$

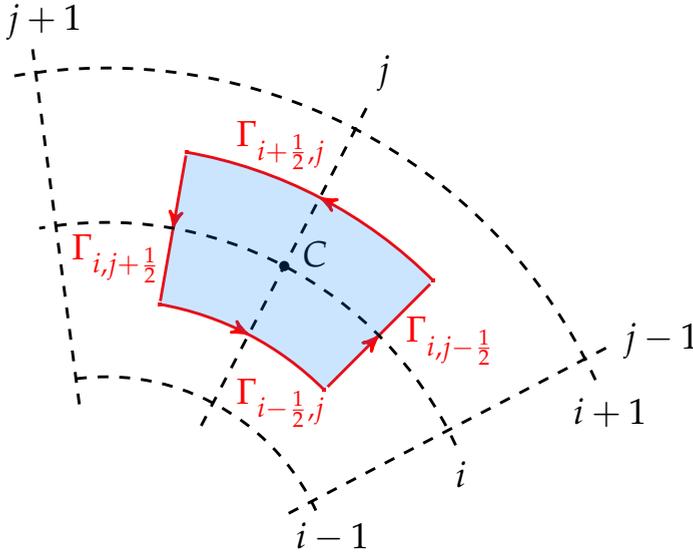


Figure B.3: Control volume for a cell-centered finite volume method on a polar grid.

Taking a midpoint approximation and using $\hat{\mathbf{n}} = \hat{\mathbf{e}}_{\rho_{i+\frac{1}{2},j}}$ results in the approximation of the first integral

$$\int_{\Gamma_{i+\frac{1}{2},j}} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds \approx F_{i+\frac{1}{2},j} \rho_{i+\frac{1}{2}} h_{\zeta}, \quad (\text{B.39a})$$

where $F_{i+\frac{1}{2},j}$ is the numerical approximation of the flux term $f_{11}(\rho_{i+\frac{1}{2}}, \zeta)$ at the interface point $(\rho_{i+\frac{1}{2}}, \zeta_j)$. Analogously, we derive

$$\int_{\Gamma_{i,j+\frac{1}{2}}} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds \approx F_{i,j+\frac{1}{2}} h_{\rho}, \quad (\text{B.39b})$$

$$\int_{\Gamma_{i-\frac{1}{2},j}} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds \approx -F_{i-\frac{1}{2},j} \rho_{i-\frac{1}{2}} h_{\zeta}, \quad (\text{B.39c})$$

$$\int_{\Gamma_{i,j-\frac{1}{2}}} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds \approx -F_{i,j-\frac{1}{2}} h_{\rho}. \quad (\text{B.39d})$$

Similarly,

$$\int_{\Gamma_{i+\frac{1}{2},j}} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds \approx R_{i+\frac{1}{2},j} \rho_{i+\frac{1}{2}} h_\zeta, \quad (\text{B.40a})$$

$$\int_{\Gamma_{i,j+\frac{1}{2}}} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds \approx R_{i,j+\frac{1}{2}} h_\rho, \quad (\text{B.40b})$$

$$\int_{\Gamma_{i-\frac{1}{2},j}} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds \approx -R_{i-\frac{1}{2},j} \rho_{i-\frac{1}{2}} h_\zeta, \quad (\text{B.40c})$$

$$\int_{\Gamma_{i,j-\frac{1}{2}}} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds \approx -R_{i,j-\frac{1}{2}} h_\rho, \quad (\text{B.40d})$$

where $R_{i+\frac{1}{2},j}$ is the numerical approximation of the flux term r_{11} at the grid point $(\rho_{i+\frac{1}{2}}, \zeta_j)$, etc.

Substituting (B.39) and (B.40) into (B.37) gives

$$\begin{aligned} & h_\zeta \left(F_{i+\frac{1}{2},j} \rho_{i+\frac{1}{2}} - F_{i-\frac{1}{2},j} \rho_{i-\frac{1}{2}} \right) + h_\rho \left(F_{i,j+\frac{1}{2}} - F_{i,j-\frac{1}{2}} \right) \\ & \approx h_\zeta \left(R_{i+\frac{1}{2},j} \rho_{i+\frac{1}{2}} - R_{i-\frac{1}{2},j} \rho_{i-\frac{1}{2}} \right) + h_\rho \left(R_{i,j+\frac{1}{2}} - R_{i,j-\frac{1}{2}} \right). \end{aligned} \quad (\text{B.41})$$

We use central differences to approximate the first-order derivatives in the numerical approximation of the flux terms:

$$\begin{aligned} F_{i+\frac{1}{2},j} &= |c_1|_{i+\frac{1}{2},j}^2 \frac{m_1(\rho_{i+1}, \zeta_j) - m_1(\rho_i, \zeta_j)}{h_\rho} + (c_1 \cdot c_2)_{i+\frac{1}{2},j} \frac{m_2(\rho_{i+1}, \zeta_j) - m_2(\rho_i, \zeta_j)}{h_\rho}, \\ F_{i-\frac{1}{2},j} &= |c_1|_{i-\frac{1}{2},j}^2 \frac{m_1(\rho_i, \zeta_j) - m_1(\rho_{i-1}, \zeta_j)}{h_\rho} + (c_1 \cdot c_2)_{i-\frac{1}{2},j} \frac{m_2(\rho_i, \zeta_j) - m_2(\rho_{i-1}, \zeta_j)}{h_\rho}, \\ F_{i,j+\frac{1}{2}} &= \frac{|c_1|_{i,j+\frac{1}{2}}^2}{\rho_i} \frac{m_1(\rho_i, \zeta_{j+1}) - m_1(\rho_i, \zeta_j)}{h_\zeta} + \frac{(c_1 \cdot c_2)_{i,j+\frac{1}{2}}}{\rho_i} \frac{m_2(\rho_i, \zeta_{j+1}) - m_2(\rho_i, \zeta_j)}{h_\zeta}, \\ F_{i,j-\frac{1}{2}} &= \frac{|c_1|_{i,j-\frac{1}{2}}^2}{\rho_i} \frac{m_1(\rho_i, \zeta_j) - m_1(\rho_i, \zeta_{j-1})}{h_\zeta} + \frac{(c_1 \cdot c_2)_{i,j-\frac{1}{2}}}{\rho_i} \frac{m_2(\rho_i, \zeta_j) - m_2(\rho_i, \zeta_{j-1})}{h_\zeta}, \\ R_{i+\frac{1}{2},j} &= (c_1 \cdot p_1)_{i+\frac{1}{2},j}, \quad R_{i-\frac{1}{2},j} = (c_1 \cdot p_1)_{i-\frac{1}{2},j}, \\ R_{i,j+\frac{1}{2}} &= (c_1 \cdot p_2)_{i,j+\frac{1}{2}}, \quad R_{i,j-\frac{1}{2}} = (c_1 \cdot p_2)_{i,j-\frac{1}{2}}, \end{aligned} \quad (\text{B.42})$$

where we make the identifications $(\rho_i, \zeta_{N_\zeta+1}) = (\rho_i, \zeta_1)$ and $(\rho_i, \zeta_0) = (\rho_i, \zeta_{N_\zeta})$ for $i = 1, \dots, N_\rho$.

Substituting (B.42) into (B.41) gives us the main discretization of the integral equation (B.37). For the second equation of the system in (B.32b), we follow

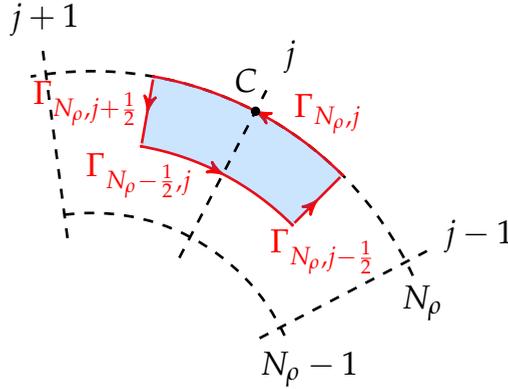


Figure B.4: Control volume for the outer boundary of a cell-centered finite volume method on a polar grid.

the same procedure to arrive at the main discretization. For the calculation of terms at the interface of the control volumes, we use linear interpolation.

In Section B.2.1, we explain briefly how the outer boundary, i.e. $i = N_\rho$, $j = 1, \dots, N_\zeta$, is incorporated into the main discretization. In Section B.2.2, we consider the inner boundary, i.e. the grid points (ρ_1, ζ_j) , with $1 \leq j \leq N_\zeta$.

After incorporating the boundary conditions, we solve the linear systems for m_1 and m_2 , by first solving for m_1 and using the new m_1 to solve for m_2 , using Matlab's builtin `mldivide`. In the next two sections, we explain briefly how the outer boundary equations are derived and describe in detail how the inner boundary is solved, since we have a coordinate singularity at the origin of the coordinate system.

B.2.1 Incorporating the outer boundary

Consider the outer boundary, i.e. $i = N_\rho$, $j = 1, \dots, N_\zeta$.

We consider the control volume with boundary

$$\partial\Omega_{N_\rho, j} = \Gamma_{N_\rho, j} \cup \Gamma_{N_\rho, j + \frac{1}{2}} \cup \Gamma_{N_\rho - \frac{1}{2}, j} \cup \Gamma_{N_\rho, j - \frac{1}{2}}, \quad (\text{B.43})$$

as shown in Figure B.4.

We integrate (B.32a) and (B.32b) over the control volume $\Omega_{N_\rho, j}$ and derive linear equations for m_1 and m_2 on the outer boundary using the boundary equations in (6.99b). The derivation is analogous to the derivation for the left-boundary equation described in Section B.1.

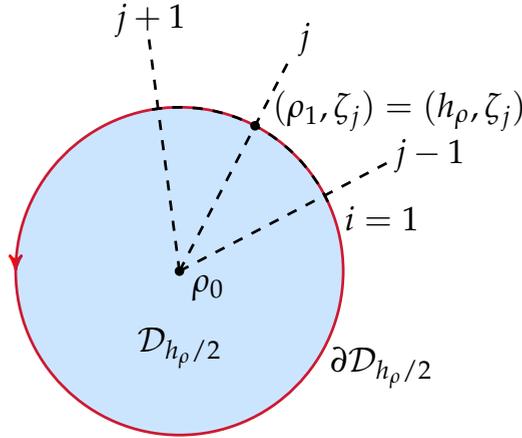


Figure B.5: Control volume for the inner boundary of a cell-centered finite volume method on a polar grid.

B.2.2 Incorporating the inner boundary

We consider the grid adjacent to the origin, i.e., the grid points (ρ_1, ζ_j) , with $1 \leq j \leq N_\zeta$. We denote the values of the components $m_1(\rho_0, \zeta_j) = m_1(\rho_0)$ and $m_2(\rho_0, \zeta_j) = m_2(\rho_0)$ at the origin by $m_{1,O}$ and $m_{2,O}$, as the vector \mathbf{m}_O . By considering a control volume as a disc around the origin, we will derive expressions for $m_{1,O}$ and $m_{2,O}$ and remove them from the main equations.

We consider the disc $\mathcal{D}_{h_\rho/2}$, centered at the origin with radius $h_\rho/2$, as control volume, as shown in Figure B.5.

Integrating (B.32a) over the control volume $\mathcal{D}_{h_\rho/2}$ using Gauss's theorem gives

$$\oint_{\partial \mathcal{D}_{h_\rho/2}} \mathbf{f}_1 \cdot \hat{\mathbf{n}} \, ds = \oint_{\partial \mathcal{D}_{h_\rho/2}} \mathbf{r}_1 \cdot \hat{\mathbf{n}} \, ds. \quad (\text{B.44})$$

Substituting the \mathbf{f}_1 and \mathbf{r}_1 terms from (B.34) and the unit outward normal $\hat{\mathbf{n}} = \hat{\mathbf{e}}_{\rho_{i,j}}$ gives

$$\oint_{\partial \mathcal{D}_{h_\rho/2}} |c_1|^2 \frac{\partial m_1}{\partial \rho} + c_1 \cdot c_2 \frac{\partial m_2}{\partial \rho} \, ds = \oint_{\partial \mathcal{D}_{h_\rho/2}} c_1 \cdot \mathbf{p}_1 \, ds. \quad (\text{B.45})$$

Similarly, for the second equation of the system (B.32b) we get

$$\oint_{\partial \mathcal{D}_{h_\rho/2}} |c_2|^2 \frac{\partial m_2}{\partial \rho} + c_1 \cdot c_2 \frac{\partial m_1}{\partial \rho} \, ds = \oint_{\partial \mathcal{D}_{h_\rho/2}} c_2 \cdot \mathbf{p}_1 \, ds. \quad (\text{B.46})$$

We can rewrite (B.45) by approximating the integrals to second-order accuracy to

$$\begin{aligned} m_{1,O} \sum_{j=1}^{N_\zeta} |c_1|_{\frac{1}{2},j}^2 + m_{2,O} \sum_{j=1}^{N_\zeta} (c_1 \cdot c_2)_{\frac{1}{2},j} \\ = \sum_{j=1}^{N_\zeta} |c_1|_{\frac{1}{2},j}^2 m_1(h_\rho, \zeta_j) + \sum_{j=1}^{N_\zeta} (c_1 \cdot c_2)_{\frac{1}{2},j} m_2(h_\rho, \zeta_j) - h_\rho \sum_{j=1}^{N_\zeta} (c_1 \cdot p_1)_{\frac{1}{2},j}, \end{aligned} \quad (\text{B.47})$$

and (B.46) to

$$\begin{aligned} m_{1,O} \sum_{j=1}^{N_\zeta} (c_1 \cdot c_2)_{\frac{1}{2},j} + m_{2,O} \sum_{j=1}^{N_\zeta} |c_2|_{\frac{1}{2},j}^2 \\ = \sum_{j=1}^{N_\zeta} |c_2|_{\frac{1}{2},j}^2 m_2(h_\rho, j) + \sum_{j=1}^{N_\zeta} (c_1 \cdot c_2)_{\frac{1}{2},j} m_1(h_\rho, \zeta_j) - h_\rho \sum_{j=1}^{N_\zeta} (c_2 \cdot p_1)_{\frac{1}{2},j}. \end{aligned} \quad (\text{B.48})$$

Using (B.47) and (B.48) we can solve for $m_{1,O}$ and $m_{2,O}$. For the grid points (ρ_1, ζ_j) , with $1 \leq j \leq N_\zeta$, we can also elaborate on the main discretization (B.41) for both (B.32a) and (B.32b) and substitute the expressions found for $m_{1,O}$ and $m_{2,O}$ into the discretization.

Note that after every iteration, we update the matrix C and also calculate $c_{1,O}$ and $c_{2,O}$ at the origin. We use this to be able to calculate values at the interface points via linear interpolation, such as $|c_1|_{\frac{1}{2},j}^2$.

B.3 Solving the Neumann problem for u

In this section, we explain how we use the finite volume method to solve the Neumann problem for u in (6.120). We use a cell-centered finite volume method on a Cartesian grid, as in Section B.1.

We rewrite (6.120a) to

$$\nabla \cdot (H_w^2 \nabla u) = \frac{1}{2} \frac{d}{dw} \left| \nabla_x H + H_w \nabla u \right|^2 - \nabla \cdot (H_w \nabla_x H), \quad (\text{B.49})$$

and further introduce $f = H_w^2 \nabla u$ and $r = H_w \nabla_x H$ such that

$$\nabla \cdot f = \frac{1}{2} \frac{d}{dw} \left| \nabla_x H + H_w \nabla u \right|^2 - \nabla \cdot r, \quad (\text{B.50})$$

and the boundary condition (6.120b) becomes

$$(f + r) \cdot \hat{n} = 0. \quad (\text{B.51})$$

We let

$$\mathbf{f} = f_1 \hat{\mathbf{e}}_{x_1} + f_2 \hat{\mathbf{e}}_{x_2}, \quad (\text{B.52})$$

$$\mathbf{r} = r_1 \hat{\mathbf{e}}_{x_1} + r_2 \hat{\mathbf{e}}_{x_2}, \quad (\text{B.53})$$

and we substitute $H(\mathbf{x}, \mathbf{y}, w)$ with $\mathbf{y} = \mathbf{m}^{n+1}$ and $w = u^n$ in all terms involving H in the right-hand side, i.e., we evaluate all H -terms, such that

$$f_1 = H_w^2(\mathbf{x}, \mathbf{m}^{n+1}, u^n) \frac{\partial u}{\partial x_1}, \quad f_2 = H_w^2(\mathbf{x}, \mathbf{m}^{n+1}, u^n) \frac{\partial u}{\partial x_2}, \quad (\text{B.54})$$

$$r_1 = H_w(\mathbf{x}, \mathbf{m}^{n+1}, u^n) H_{x_1}(\mathbf{x}, \mathbf{m}^{n+1}, u^n), \quad (\text{B.55})$$

$$r_2 = H_w(\mathbf{x}, \mathbf{m}^{n+1}, u^n) H_{x_2}(\mathbf{x}, \mathbf{m}^{n+1}, u^n). \quad (\text{B.56})$$

We perform these evaluations on the right-hand side of (B.49) with the previous u^n since it is very hard to solve (B.49) for u while taking into account *all* terms.

Integrating over a control volume Ω_C and using Gauss's theorem twice gives

$$\begin{aligned} \oint_{\partial\Omega_C} \mathbf{f} \cdot \hat{\mathbf{n}} \, ds &= \int_{\Omega_C} \frac{1}{2} \frac{d}{dw} \left| \nabla_x H(\mathbf{x}, \mathbf{m}^{n+1}, u^n) + H_w(\mathbf{x}, \mathbf{m}^{n+1}, u^n) \nabla u^n \right|^2 dx \\ &\quad - \oint_{\partial\Omega_C} \mathbf{r} \cdot \hat{\mathbf{n}} \, ds, \end{aligned} \quad (\text{B.57})$$

where $\hat{\mathbf{n}}$ is the unit outward normal and $\partial\Omega_C$ is oriented counterclockwise.

We discretize the source domain \mathcal{X} using a standard rectangular $N_1 \times N_2$ grid as given in (6.12) and let $\mathbf{x}_{ij} = \mathbf{x}_C$. We consider the control volume Ω_C given in (B.7), drawn as the cell in Figure B.1. The boundary of the control volume is divided into four parts and we have $\partial\Omega_C = \Gamma_e \cup \Gamma_n \cup \Gamma_w \cup \Gamma_s$. Rewriting (B.57) gives

$$\begin{aligned} &\int_{\Gamma_e} \mathbf{f} \cdot \hat{\mathbf{n}} \, ds + \int_{\Gamma_n} \mathbf{f} \cdot \hat{\mathbf{n}} \, ds + \int_{\Gamma_w} \mathbf{f} \cdot \hat{\mathbf{n}} \, ds + \int_{\Gamma_s} \mathbf{f} \cdot \hat{\mathbf{n}} \, ds \\ &= \int_{\Omega_C} \frac{1}{2} \frac{d}{dw} \left| \nabla_x H(\mathbf{x}, \mathbf{m}^{n+1}, u^n) + H_w(\mathbf{x}, \mathbf{m}^{n+1}, u^n) \nabla u^n \right|^2 dx \\ &\quad - \int_{\Gamma_e} \mathbf{r} \cdot \hat{\mathbf{n}} \, ds - \int_{\Gamma_n} \mathbf{r} \cdot \hat{\mathbf{n}} \, ds - \int_{\Gamma_w} \mathbf{r} \cdot \hat{\mathbf{n}} \, ds - \int_{\Gamma_s} \mathbf{r} \cdot \hat{\mathbf{n}} \, ds. \end{aligned} \quad (\text{B.58})$$

The normals $\hat{\mathbf{n}}$ at each of the boundary segments are given in (B.10). Taking a midpoint approximation and using $\hat{\mathbf{n}} = \hat{\mathbf{e}}_{x_1}$ results in the approximation of the first integral

$$\int_{\Gamma_e} \mathbf{f} \cdot \hat{\mathbf{n}} \, ds = \int_{x_{2,s}}^{x_{2,n}} f_1(x_{1,e}, x_2) \, dx_2 \approx F_{1,e} h_2, \quad (\text{B.59})$$

where $F_{1,e}$ is the numerical approximation of the flux term f_1 at the interface point x_e . Analogously, we derive

$$\begin{aligned}
 \int_{\Gamma_n} \mathbf{f} \cdot \hat{\mathbf{n}} \, ds &= - \int_{x_{1,e}}^{x_{1,w}} f_2(x_1, x_{2,n}) \, dx_1 \approx F_{2,n} h_1, \\
 \int_{\Gamma_w} \mathbf{f} \cdot \hat{\mathbf{n}} \, ds &= - \int_{x_{2,n}}^{x_{2,s}} -f_1(x_{1,w}, x_2) \, dx_2 \approx -F_{1,w} h_2, \\
 \int_{\Gamma_s} \mathbf{f} \cdot \hat{\mathbf{n}} \, ds &= \int_{x_{1,w}}^{x_{1,e}} -f_2(x_1, x_{2,s}) \, dx_1 \approx -F_{2,s} h_1.
 \end{aligned} \tag{B.60}$$

Similarly,

$$\begin{aligned}
 \int_{\Gamma_e} \mathbf{r} \cdot \hat{\mathbf{n}} \, ds &= \int_{x_{2,s}}^{x_{2,n}} r_1(x_{1,e}, x_2) \, dx_2 \approx R_{1,e} h_2, \\
 \int_{\Gamma_n} \mathbf{r} \cdot \hat{\mathbf{n}} \, ds &= - \int_{x_{1,e}}^{x_{1,w}} r_2(x_1, x_{2,n}) \, dx_1 \approx R_{2,n} h_1, \\
 \int_{\Gamma_w} \mathbf{r} \cdot \hat{\mathbf{n}} \, ds &= - \int_{x_{2,n}}^{x_{2,s}} -r_1(x_{1,w}, x_2) \, dx_2 \approx -R_{1,w} h_2, \\
 \int_{\Gamma_s} \mathbf{r} \cdot \hat{\mathbf{n}} \, ds &= \int_{x_{1,w}}^{x_{1,e}} -r_2(x_1, x_{2,s}) \, dx_1 \approx -R_{2,s} h_1,
 \end{aligned} \tag{B.61}$$

where $R_{1,e}$ is the numerical approximation of the flux term r_1 at the interface point x_e , etc.

Substituting (B.60) and (B.61) into (B.58) gives

$$\begin{aligned}
 &h_2 (F_{1,e} - F_{1,w}) + h_1 (F_{2,n} - F_{2,s}) \\
 &\approx \int_{\Omega_C} \frac{1}{2} \frac{d}{dw} \left| \nabla_x H(\mathbf{x}, \mathbf{m}^{n+1}, u^n) + H_w(\mathbf{x}, \mathbf{m}^{n+1}, u^n) \nabla u^n \right|^2 \, dx \\
 &- h_2 (R_{1,e} - R_{1,w}) - h_1 (R_{2,n} - R_{2,s}).
 \end{aligned} \tag{B.62}$$

We use central differences to approximate the first-order derivatives in the numerical approximation of the flux terms, i.e.,

$$\begin{aligned}
 F_{1,e} &= (H_w)_e^2 \frac{u(\mathbf{x}_E) - u(\mathbf{x}_C)}{h_1}, & F_{1,w} &= (H_w)_w^2 \frac{u(\mathbf{x}_C) - u(\mathbf{x}_W)}{h_1}, \\
 F_{2,n} &= (H_w)_n^2 \frac{u(\mathbf{x}_N) - u(\mathbf{x}_C)}{h_2}, & F_{2,s} &= (H_w)_s^2 \frac{u(\mathbf{x}_C) - u(\mathbf{x}_S)}{h_2}, \\
 R_{1,e} &= (H_w H_{x_1})_e, & R_{1,w} &= (H_w H_{x_1})_w, & R_{2,n} &= (H_w H_{x_2})_n, & R_{2,s} &= (H_w H_{x_2})_s,
 \end{aligned} \tag{B.63}$$

where $(H_w)_e^2 = (H_w(\mathbf{x}_e, \mathbf{m}^{n+1}(\mathbf{x}_e), u^n(\mathbf{x}_e)))^2$, $(H_{x_1})_e = H_{x_1}(\mathbf{x}_e, \mathbf{m}^{n+1}(\mathbf{x}_e), u^n(\mathbf{x}_e))$, etc., evaluated at the interface point \mathbf{x}_e , using a bilinear interpolation method, etc.

We can rewrite the remaining integral as

$$\begin{aligned}
 I_R[u^n, \mathbf{m}^{n+1}] &:= \int_{\Omega_C} \frac{1}{2} \frac{d}{dw} \left| \nabla_x H(\mathbf{x}, \mathbf{m}^{n+1}, u^n) + H_w(\mathbf{x}, \mathbf{m}^{n+1}, u^n) \nabla u^n \right|^2 dx \\
 &= \int_{\Omega_C} (\nabla_x H + H_w \nabla u) \cdot (\nabla_x H_w + H_{ww} \nabla u) dx \\
 &= \int_{\Omega_C} \nabla_x H \cdot \nabla_x H_w + H_{ww} \nabla_x H \cdot \nabla u + H_w \nabla u \cdot \nabla_x H_w + H_w H_{ww} |\nabla u|^2 dx.
 \end{aligned} \tag{B.64}$$

Using the midpoint rule we obtain

$$\begin{aligned}
 I_R[u^n, \mathbf{m}^{n+1}] &\approx h_1 h_2 \left((\nabla_x H \cdot \nabla_x H_w)_C + (H_{ww} H_{x_1} + H_w H_{x_1 w})_C \frac{u^n(\mathbf{x}_E) - u^n(\mathbf{x}_W)}{2 h_1} \right. \\
 &\quad + (H_{ww} H_{x_2} + H_w H_{x_2 w})_C \frac{u^n(\mathbf{x}_N) - u^n(\mathbf{x}_S)}{2 h_2} + (H_w H_{ww})_C \left(\frac{u^n(\mathbf{x}_E) - u^n(\mathbf{x}_W)}{2 h_1} \right)^2 \\
 &\quad \left. + (H_w H_{ww})_C \left(\frac{u^n(\mathbf{x}_N) - u^n(\mathbf{x}_S)}{2 h_2} \right)^2 \right) := h_1 h_2 Z(\mathbf{x}_C, \mathbf{m}^{n+1}, u^n),
 \end{aligned} \tag{B.65}$$

where $(H_{ww} H_{x_1})_C = H_{ww}(\mathbf{x}_C, \mathbf{m}^{n+1}(\mathbf{x}_C), u^n(\mathbf{x}_C)) H_{x_1}(\mathbf{x}_C, \mathbf{m}^{n+1}(\mathbf{x}_C), u^n(\mathbf{x}_C))$ and the other H -terms are evaluated at the centre point \mathbf{x}_C .

Substituting (B.63) and (B.65) into (B.62) and dividing by $h_1 h_2$ gives

$$\begin{aligned}
 &(H_w)_e^2 \frac{u(\mathbf{x}_E)}{h_1^2} + (H_w)_w^2 \frac{u(\mathbf{x}_W)}{h_1^2} + (H_w)_n^2 \frac{u(\mathbf{x}_N)}{h_2^2} + (H_w)_s^2 \frac{u(\mathbf{x}_S)}{h_2^2} \\
 &\quad - \left(\frac{(H_w)_e^2}{h_1^2} + \frac{(H_w)_w^2}{h_1^2} + \frac{(H_w)_n^2}{h_2^2} + \frac{(H_w)_s^2}{h_2^2} \right) u(\mathbf{x}_C) \\
 &= Z(\mathbf{x}_C, \mathbf{m}^{n+1}, u^n) - \frac{1}{h_1} (H_w H_{x_1})_e + \frac{1}{h_1} (H_w H_{x_1})_w \\
 &\quad - \frac{1}{h_2} (H_w H_{x_2})_n + \frac{1}{h_2} (H_w H_{x_2})_s.
 \end{aligned} \tag{B.66}$$

B.3.1 Incorporating boundary conditions

Consider the left boundary without corner points, i.e. we have $i = 1$ with $j = 2, \dots, N_2 - 1$. We consider the control volume

$$\Omega_C = [x_{1,C}, x_{1,e}] \times [x_{2,s}, x_{2,n}], \tag{B.67}$$

as shown in Figure B.2.

The boundary of the control volume is divided into four parts such that $\partial\Omega_C = \Gamma_e \cup \Gamma_n \cup \Gamma_C \cup \Gamma_s$. Integrating (B.50) over the control volume Ω_C using Gauss's theorem gives

$$\begin{aligned}
 & \int_{\Gamma_e} \mathbf{f} \cdot \hat{\mathbf{n}} \, ds + \int_{\Gamma_n} \mathbf{f} \cdot \hat{\mathbf{n}} \, ds + \int_{\Gamma_C} \mathbf{f} \cdot \hat{\mathbf{n}} \, ds + \int_{\Gamma_s} \mathbf{f} \cdot \hat{\mathbf{n}} \, ds \\
 &= \int_{\Omega_C} \frac{1}{2} \frac{d}{dw} \left| \nabla_x H(\mathbf{x}, \mathbf{m}^{n+1}, u^n) + H_w(\mathbf{x}, \mathbf{m}^{n+1}, u^n) \nabla u^n \right|^2 \, dx \\
 & \quad - \int_{\Gamma_e} \mathbf{r} \cdot \hat{\mathbf{n}} \, ds - \int_{\Gamma_n} \mathbf{r} \cdot \hat{\mathbf{n}} \, ds - \int_{\Gamma_C} \mathbf{r} \cdot \hat{\mathbf{n}} \, ds - \int_{\Gamma_s} \mathbf{r} \cdot \hat{\mathbf{n}} \, ds. \tag{B.68}
 \end{aligned}$$

Taking approximations and substituting $\hat{\mathbf{n}}$ in (B.10), with Γ_C replacing Γ_w , at each of the boundaries results in the approximation of the integrals

$$\begin{aligned}
 \int_{\Gamma_e} \mathbf{f} \cdot \hat{\mathbf{n}} \, ds &= \int_{x_{2,s}}^{x_{2,n}} f_1(x_{1,e}, x_2) \, dx_2 \approx F_{1,e} h_2, \\
 \int_{\Gamma_n} \mathbf{f} \cdot \hat{\mathbf{n}} \, ds &= - \int_{x_{1,e}}^{x_{1,C}} f_2(x_1, x_{2,n}) \, dx_1 \approx \frac{F_{2,n} h_1}{2}, \quad (\text{no midpoint}) \\
 \int_{\Gamma_C} \mathbf{f} \cdot \hat{\mathbf{n}} \, ds &= - \int_{x_{2,n}}^{x_{2,s}} -f_1(x_{1,C}, x_2) \, dx_2 \approx -F_{1,C} h_2, \\
 \int_{\Gamma_s} \mathbf{f} \cdot \hat{\mathbf{n}} \, ds &= \int_{x_{1,C}}^{x_{1,e}} -f_2(x_1, x_{2,s}) \, dx_1 \approx -\frac{F_{2,s} h_1}{2}, \quad (\text{no midpoint}). \tag{B.69}
 \end{aligned}$$

Similarly,

$$\begin{aligned}
 \int_{\Gamma_e} \mathbf{r} \cdot \hat{\mathbf{n}} \, ds &= \int_{x_{2,s}}^{x_{2,n}} r_1(x_{1,e}, x_2) \, dx_2 \approx R_{1,e} h_2, \\
 \int_{\Gamma_n} \mathbf{r} \cdot \hat{\mathbf{n}} \, ds &= - \int_{x_{1,e}}^{x_{1,C}} r_2(x_1, x_{2,n}) \, dx_1 \approx \frac{R_{2,n} h_1}{2}, \\
 \int_{\Gamma_C} \mathbf{r} \cdot \hat{\mathbf{n}} \, ds &= - \int_{x_{2,n}}^{x_{2,s}} -r_1(x_{1,C}, x_2) \, dx_2 \approx -R_{1,C} h_2, \\
 \int_{\Gamma_s} \mathbf{r} \cdot \hat{\mathbf{n}} \, ds &= \int_{x_{1,C}}^{x_{1,e}} -r_2(x_1, x_{2,s}) \, dx_1 \approx -\frac{R_{2,s} h_1}{2}. \tag{B.70}
 \end{aligned}$$

Substituting (B.69) and (B.70) into (B.68) gives

$$\begin{aligned}
 & h_2 (F_{1,e} - F_{1,C}) + \frac{h_1}{2} (F_{2,n} - F_{2,s}) \\
 &= \int_{\Omega_C} \frac{1}{2} \frac{d}{dw} \left| \nabla_x H(\mathbf{x}, \mathbf{m}^{n+1}, u^n) + H_w(\mathbf{x}, \mathbf{m}^{n+1}, u^n) \nabla u^n \right|^2 dx \\
 & \quad - h_2 (R_{1,e} - R_{1,C}) - \frac{h_1}{2} (R_{2,n} - R_{2,s}). \tag{B.71}
 \end{aligned}$$

Again, we approximate the remaining integral by

$$\begin{aligned}
 & \frac{h_1}{2} h_2 \left((\nabla_x H \cdot \nabla_x H_w)_c + (H_{ww} H_{x_1} + H_w H_{x_1 w})_c \frac{u^n(\mathbf{x}_E) - u^n(\mathbf{x}_C)}{h_1} \right. \\
 & \quad + (H_{ww} H_{x_2} + H_w H_{x_2 w})_c \frac{u^n(\mathbf{x}_N) - u^n(\mathbf{x}_S)}{2 h_2} + (H_w H_{ww})_c \left(\frac{u^n(\mathbf{x}_E) - u^n(\mathbf{x}_C)}{h_1} \right)^2 \\
 & \quad \left. + (H_w H_{ww})_c \left(\frac{u^n(\mathbf{x}_N) - u^n(\mathbf{x}_S)}{2 h_2} \right)^2 \right) := \frac{h_1}{2} h_2 Z(\mathbf{x}_C, \mathbf{m}^{n+1}, u^n), \tag{B.72}
 \end{aligned}$$

using the midpoint rule, where we denote $\mathbf{x}_c = (\mathbf{x}_C + \mathbf{x}_e)/2$ as the centre point of the half cell.

We replace $F_{1,C} = (H_w)_C^2 \frac{u(\mathbf{x}_e) - u(\mathbf{x}_w)}{h_1}$ using the boundary equation (B.51). Hence, with $\hat{\mathbf{n}} = (-1, 0)$ we get that the boundary equation becomes

$$H_w(\mathbf{x}_C, \mathbf{m}^{n+1}, u^n) H_{x_1}(\mathbf{x}_C, \mathbf{m}^{n+1}, u^n) + F_{1,C} = 0,$$

and solving for $F_{1,C}$ gives

$$F_{1,C} = -H_w(\mathbf{x}_C, \mathbf{m}^{n+1}, u^n) H_{x_1}(\mathbf{x}_C, \mathbf{m}^{n+1}, u^n). \tag{B.73}$$

Substituting the remaining flux terms (B.63) and (B.73) into (B.71) and dividing by $h_1 h_2$ gives

$$\begin{aligned}
 & (H_w)_e^2 \frac{u(\mathbf{x}_E)}{h_1^2} + (H_w)_n^2 \frac{u(\mathbf{x}_N)}{2 h_2^2} + (H_w)_s^2 \frac{u(\mathbf{x}_S)}{2 h_2^2} \\
 & \quad - \left(\frac{(H_w)_e^2}{h_1^2} + \frac{(H_w)_s^2}{2 h_2^2} + \frac{(H_w)_n^2}{2 h_2^2} \right) u(\mathbf{x}_C) \\
 &= \frac{Z(\mathbf{x}_C, \mathbf{m}^{n+1}, u^n)}{2} - \frac{1}{h_1} (H_w H_{x_1})_e - \frac{1}{2 h_2} (H_w H_{x_2})_n + \frac{1}{2 h_2} (H_w H_{x_2})_s. \tag{B.74}
 \end{aligned}$$

We obtain a finite difference equation for all the grid points in our grid, i.e., for interior grid points we have (B.66) and for the left boundary we have (B.74). The equations for the right, upper and lower boundaries and the corner points are derived analogously.

The mapping $\mathbf{y} = \mathbf{m}(x) = \bar{\mathbf{m}}(x, u, \nabla u)$ is a function of $u(x)$ and $\nabla u(x)$ given implicitly in (4.78). Hence, the Neumann problem (6.119) for u has multiple solutions and a corresponding discretization matrix with incomplete rank. Given the mapping \mathbf{m}^{n+1} during iteration n , the average height of the surface u^{n+1} can be fixed to calculate a unique solution. We calculate a unique solution by prescribing the average value of u as a constraint which adds an extra row to the discretization matrix, as explained in Section 6.1.4 with $u = u_1$.

Bibliography

- [1] F. Abedin and C. E. Gutiérrez. An iterative method for generated Jacobi equations. *Calc. Var. Partial Differential Equations*, 56(4):101, 2017.
- [2] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2nd edition, 2008.
- [3] S. B. Angenent, S. Haker, and A. R. Tannenbaum. Minimizing flows for the Monge-Kantorovich problem. *SIAM J. Math. Anal.*, 35(1):61–97, 2003.
- [4] R. Beltman. Solving the Monge-Ampère equation for a free-form reflector in arbitrary coordinate systems. Master's thesis, Eindhoven University of Technology, Eindhoven, the Netherlands, 2015.
- [5] R. Beltman, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. A least-squares method for the inverse reflector problem in arbitrary orthogonal coordinates. *J. Comput. Phys.*, 367:347–373, 2018.
- [6] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM J. Sci. Comput.*, 37(2):A1111–A1138, 2015.
- [7] J.-D. Benamou and V. Duval. Minimal convex extensions and finite difference discretisation of the quadratic Monge-Kantorovich problem. *Eur. J. Appl. Math.*, 30(6):1041–1078, 2019.
- [8] J.-D. Benamou, B. D. Froese, and A. M. Oberman. Two numerical methods for the elliptic Monge-Ampère equation. *Math. Model. Numer. Anal.*, 44(4):737–758, 2010.
- [9] J.-D. Benamou, B. D. Froese, and A. M. Oberman. Numerical solution of the optimal transportation problem using the Monge-Ampère equation. *J. Comput. Phys.*, 260:107–126, 2014.

- [10] M. W. M. C. Bertens, E. M. T. Vugts, M. J. H. Anthonissen, J. H. M. Boonkamp, and W. L. IJzerman. Numerical methods for the hyperbolic Monge-Ampère equation based on the method of characteristics. *arXiv preprint arXiv:2104.11659*, 2021.
- [11] N. Bonneel, G. Peyré, and M. Cuturi. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Trans. Graph.*, 35(4):71–81, 2016.
- [12] M. Born and E. Wolf. *Principles of Optics*. Cambridge University Press, Cambridge, UK, 7th edition, 1999.
- [13] J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. CMS Books in Mathematics. Springer, New York, 2005.
- [14] C. Bösel and H. Gross. Ray mapping approach for the efficient design of continuous freeform surfaces. *Opt. Express*, 24(13):14271–14282, 2016.
- [15] C. Bösel and H. Gross. Single freeform surface design for prescribed input wavefront and target irradiance. *J. Opt. Soc. Am. A*, 34(9):1490–1499, 2017.
- [16] C. Bösel and H. Gross. Double freeform illumination design for prescribed wavefronts and irradiances. *J. Opt. Soc. Am. A*, 35(2):236–243, 2018.
- [17] Y. Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305:805–808, 1987.
- [18] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure and Appl. Math*, 44(4):375–417, 1991.
- [19] K. Brix, Y. Hafizogullari, and A. Platen. Designing illumination lenses and mirrors by the numerical solution of Monge-Ampère equations. *J. Opt. Soc. Am. A*, 32(11):2227–2236, 2015.
- [20] K. Brix, Y. Hafizogullari, and A. Platen. Solving the Monge-Ampère equations for the inverse reflector problem. *Math. Models Methods Appl. Sci.*, 25(5):803–837, 2015.
- [21] A. Bruneton, A. Bäuerle, R. Wester, J. Stollenwerk, and P. Loosen. High resolution irradiance tailoring using multiple freeform surfaces. *Opt. Express*, 21(9):10563–10571, 2013.

-
- [22] D. A. Bykov, L. L. Doskolovich, A. A. Mingazov, E. A. Bezus, and N. L. Kazanskiy. Linear assignment problem in the design of freeform refractive optical elements generating prescribed irradiance distributions. *Opt. Express*, 26(21):27812–27825, 2018.
- [23] A. Caboussat, R. Glowinski, and D. Gourzoulidis. A least-squares/relaxation method for the numerical solution of the three-dimensional elliptic Monge-Ampère equation. *J. Sci. Comput.*, 77(1):53–78, 2018.
- [24] A. Caboussat, R. Glowinski, and D. C. Sorensen. A least-squares method for the numerical solution of the Dirichlet problem for the elliptic Monge-Ampère equation in dimension two. *ESAIM: Control Optim. Calc. Var.*, 19(3):780–810, 2013.
- [25] L. A. Caffarelli. Boundary regularity of maps with convex potentials–II. *Ann. of Math.*, 144(3):453–496, 1996.
- [26] L. A. Caffarelli and M. Milman, editors. *On the numerical solution of the problem of reflector design with given far-field scattering data*, volume 226. AMS, 1999.
- [27] L. A. Caffarelli and V. I. Oliker. Weak solutions of one inverse problem in geometric optics. *J. Math. Sci.*, 154(1):39–49, 2008.
- [28] C. Canavesi, W. J. Cassarly, and J. P. Rolland. Target flux estimation by calculating intersections between neighboring conic reflector patches. *Opt. Lett.*, 38(23):5012–5015, 2013.
- [29] P. M. M. de Castro, Q. Mérigot, and B. Thibert. Far-field reflector problem and intersection of paraboloids. *Numer. Math.*, 134(2):389–411, 2016.
- [30] F. Cavalletti and M. Huesmann. Existence and uniqueness of optimal transport maps. In *Annales de l’Institut Henri Poincaré (C) Non Linear Analysis*, volume 32, pages 1367–1377. Elsevier, 2015.
- [31] S. Chang, R. Wu, L. An, and Z. Zheng. Design beam shapers with double freeform surfaces to form a desired wavefront with prescribed illumination pattern by solving a Monge-Ampère type equation. *J. Opt.*, 18(12):125602, 2016.
- [32] R. Courant and D. Hilbert. *Methods of Mathematical Physics*, volume 2. John Wiley & Sons, 1989 edition, 1953.
-

- [33] J. A. Cuesta-Albertos and A. Tuero-Diaz. A characterization for the solution of the Monge-Kantorovich mass transference problem. *Stat. Probab. Lett.*, 16(2):147–152, 1993.
- [34] M. J. P. Cullen and R. J. Purser. Properties of the Lagrangian semigeostrophic equations. *J. Atmos. Sci.*, 46(17):2684–2697, 1989.
- [35] E. J. Dean and R. Glowinski. Numerical solution of the two-dimensional elliptic Monge-Ampère equation with Dirichlet boundary conditions: An augmented Lagrangian approach. *Comptes rendus Mathématique*, 336(9):779–784, 2003.
- [36] E. J. Dean and R. Glowinski. Numerical solution of the two-dimensional elliptic Monge-Ampère equation with Dirichlet boundary conditions: A least-squares approach. *Comptes rendus Mathématique*, 339(12):887–892, 2004.
- [37] E. J. Dean and R. Glowinski. Numerical methods for fully nonlinear elliptic equations of the Monge-Ampère type. *Comput. Method. Appl. M.*, 195(13):1344–1386, 2006.
- [38] P. Delanoë. Classical solvability in dimension two of the second boundary-value problem associated with the Monge-Ampère operator. *Ann. Inst. Henri Poincaré, Analyse Non Linéaire*, 8(5):443–457, 1991.
- [39] K. Desnijder, P. Hanselaer, and Y. Meuret. Ray mapping method for off-axis and non-paraxial freeform illumination lens design. *Opt. Lett.*, 44(4):771–774, 2019.
- [40] L. L. Doskolovich, D. A. Bykov, E. S. Andreev, E. A. Bezus, and V. I. Olikier. Designing double freeform surfaces for collimated beam shaping with optimal mass transportation and linear assignment problems. *Opt. Express*, 26(19):24602–24613, 2018.
- [41] L. L. Doskolovich, D. A. Bykov, A. A. Mingazov, and E. A. Bezus. Optimal mass transportation and linear assignment problems in the design of freeform refractive optical elements generating far-field irradiance distributions. *Opt. Express*, 27(9):13083–13097, 2019.
- [42] R. S. Ellis. Convex functions and the Legendre-Fenchel transform. In *Entropy, Large Deviations, and Statistical Mechanics*, pages 211–228. Springer, 1985.
- [43] L. C. Evans. *Partial Differential Equations*. American Mathematical Society, Providence, R.I., 2nd edition, 2010.

- [44] F. Z. Fang, X. D. Zhang, A. Weckenmann, G. X. Zhang, and C. Evans. Manufacturing and measurement of freeform optics. *CIRP Annals*, 62(2):823–846, 2013.
- [45] X. Feng, R. Glowinski, and M. Neilan. Recent developments in numerical methods for fully nonlinear second order partial differential equations. *SIAM Rev. Soc. Ind. Appl. Math.*, 55(2):205–267, 2013.
- [46] X. Feng and M. Neilan. Mixed finite element methods for the fully nonlinear Monge-Ampère equation based on the vanishing moment method. *SIAM J. Numer. Anal.*, 47(2):1226–1250, 2009.
- [47] X. Feng and M. Neilan. Vanishing moment method and moment solutions for fully nonlinear second order partial differential equations. *J. Sci. Comput.*, 38(1):74–98, 2009.
- [48] X. Feng and M. Neilan. Analysis of Galerkin methods for the fully nonlinear Monge-Ampère equation. *J. Sci. Comput.*, 47(3):303–327, 2011.
- [49] Z. Feng, D. Cheng, and Y. Wang. Iterative wavefront tailoring to simplify freeform optical design for prescribed irradiance. *Opt. Lett.*, 44(9):2274–2277, 2019.
- [50] Z. Feng, B. D. Froese, C.-Y. Huang, D. Ma, and R. Liang. Creating unconventional geometric beams with large depth of field using double freeform-surface optics. *Appl. Opt.*, 54(20):6277–6281, 2015.
- [51] Z. Feng, B. D. Froese, and R. Liang. Freeform illumination optics construction following an optimal transport map. *Appl. Opt.*, 55(16):4301–4306, Jun 2016.
- [52] Z. Feng, B. D. Froese, R. Liang, D. Cheng, and Y. Wang. Simplified freeform optics design for complicated laser beam shaping. *Appl. Opt.*, 56(33):9308–9314, 2017.
- [53] Z. Feng, L. Huang, G. Jin, and M. Gong. Designing double freeform optical surfaces for controlling both irradiance and wavefront. *Opt. Express*, 21(23):28693–28701, 2013.
- [54] A. Figalli, Y.-H. Kim, and R. J. McCann. When is multidimensional screening a convex program? *J. Econ. Theory*, 146(2):454–478, 2011.
- [55] C. Filosa, J. H. M. ten Thije Boonkamp, and W. L. IJzerman. Ray tracing method in phase space for two-dimensional optical systems. *Appl. Opt.*, 55(13):3599–3606, 2016.

- [56] C. Filosa, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Phase space ray tracing for a two-dimensional parabolic reflector. *Mathematics and Statistics*, 5(4):135–142, 2017.
- [57] F. R. Fournier, W. J. Cassarly, and J. P. Rolland. Fast freeform reflector generation using source-target maps. *Opt. Express*, 18(5):5295–5304, 2010.
- [58] B. D. Froese. A numerical method for the elliptic Monge-Ampère equation with transport boundary conditions. *SIAM J. Sci. Comput.*, 34(3):A1432–A1459, 2012.
- [59] B. D. Froese and A. M. Oberman. Convergent finite difference solvers for viscosity solutions of the elliptic Monge-Ampère equation in dimensions two and higher. *SIAM J. Numer. Anal.*, 49(4):1692–1714, 2011.
- [60] B. D. Froese and A. M. Oberman. Fast finite difference solvers for singular solutions of the elliptic Monge-Ampère equation. *J. Comput. Phys.*, 230(3):818–834, 2011.
- [61] B. D. Froese and A. M. Oberman. Convergent filtered schemes for the Monge-Ampère partial differential equation. *SIAM J. Numer. Anal.*, 51(1):423–444, 2013.
- [62] W. Gangbo and V. I. Oliker. Existence of optimal maps in the reflector-type problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 13(1):93–106, 2007.
- [63] R. J. Garver. On the nature of the roots of a quartic equation. *Mathematics News Letter*, pages 6–8, 1933.
- [64] I. M. Gelfand and S. V. Fomin. *Calculus of Variations*. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1963. Translated by R. A. Silverman.
- [65] P. Gimenez-Benitez, J. C. Miñano, J. Blen, R. M. Arroyo, J. Chaves, O. Dross, M. Hernández, and W. Falicoff. Simultaneous multiple surface optical design method in three dimensions. *Opt. Eng.*, 43(7):1489–1503, 2004.
- [66] A. S. Glassner. *An Introduction to Ray Tracing*. Elsevier, 1989.
- [67] T. Glimm and V. I. Oliker. Optical design of single reflector systems and the Monge-Kantorovich mass transfer problem. *J. Math. Sci.*, 117(3):4096–4108, 2003.

-
- [68] T. Graf and V. I. Oliker. An optimal mass transport approach to the near-field reflector problem in optical design. *Inverse Probl.*, 28(2):025001, 2012.
- [69] N. Guillen. A primer on generated Jacobian equations: Geometry, optics, economics. *Notices Amer. Math. Soc.*, 66(9), 2019.
- [70] N. Guillen and J. Kitagawa. Pointwise estimates and regularity in geometric optics and other generated Jacobian equations. *Comm. Pure Appl. Math.*, 70(6):1146–1220, 2017.
- [71] C. E. Gutiérrez. Refraction problems in geometric optics. In *Fully Nonlinear PDEs in Real and Complex Geometry and Optics*, pages 95–150. Springer, Cham, 2014.
- [72] C. E. Gutiérrez and Q. Huang. The refractor problem in reshaping light beams. *Arch. Ration. Mech. Anal.*, 193(2):423–443, 2009.
- [73] D. ter Haar. *The Old Quantum Theory: The Commonwealth and International Library: Selected Readings in Physics*. Elsevier, 1st edition, 1967.
- [74] E. Haber, T. ur Rehman, and A. R. Tannenbaum. An efficient numerical method for the solution of the L_2 optimal mass transfer problem. *SIAM J. Sci. Comput.*, 32(1):197–211, 2010.
- [75] W. R. Hamilton. Theory of systems of rays. *The Transactions of the Royal Irish Academy*, pages 69–174, 1828.
- [76] W. R. Hamilton. Second supplement to an essay on the theory of systems of rays. *The Transactions of the Royal Irish Academy*, pages 92–126, 1830.
- [77] W. R. Hamilton. Supplement to an essay on the theory of systems of rays. *The Transactions of the Royal Irish Academy*, pages 3–62, 1830.
- [78] E. Hecht. *Optics, Global Edition*. Pearson Education Limited, 5th edition, 2016.
- [79] M. J. Herzberger. Optics from Euclid to Huygens. *Appl. Opt.*, 5(9):1383–1393, 1966.
- [80] C. Huygens. *Traité de la lumière* (drafted 1678; published in Leyden by Van der Aa, 1690), translated by Silvanus P. Thompson as *Treatise on light*, 2005.

- [81] C. Huygens, T. Young, A. J. Fresnel, and F. Arago. *The Wave Theory of Light: Memoirs of Huygens, Young and Fresnel*, volume 15. American Book Company, 1900.
- [82] F. Jiang and N. S. Trudinger. On the second boundary value problem for Monge-Ampère type equations and geometric optics. *N.S. Arch. Ration. Mech. Anal.*, 229:547–566, 2018.
- [83] A. L. Karakhanyan and A. Sabra. Refractor surfaces determined by near-field data. *J. Differ. Equ.*, 2020.
- [84] E. L. Kawecki, O. Lakkis, and T. Pryer. A finite element method for the Monge-Ampère equation with transport boundary conditions. *arXiv preprint arXiv:1807.03535*, 2018.
- [85] S. A. Kochengin and V. I. Olikier. Determination of reflector surfaces from near-field scattering data. *Inverse Problems*, 13(2):363–373, 1997.
- [86] S. A. Kochengin and V. I. Olikier. Determination of reflector surfaces from near-field scattering data II. Numerical solution. *Numer. Math.*, 79(4):553–568, 1998.
- [87] S. A. Kochengin and V. I. Olikier. Computational algorithms for constructing reflectors. *Comput. Visual Sci.*, 6(1):15–21, 2003.
- [88] V. C. E. Kronberg, M. J. H. Anthonissen, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Modelling surface light scattering in the context of freeform optical design. *arXiv preprint arXiv:2106.01691*, 2021.
- [89] O. Lakkis and T. Pryer. A finite element method for second order nonvariational elliptic problems. *SIAM J. Sci. Comput.*, 33(2):786–801, 2011.
- [90] O. Lakkis and T. Pryer. A finite element method for nonlinear elliptic problems. *SIAM J. Sci. Comput.*, 35(4):A2025–A2045, 2013.
- [91] R. De Leo, C. E. Gutiérrez, and H. Mawi. On the numerical solution of the far field refractor problem. *Nonlinear Anal.*, 157:123–145, 2017.
- [92] Q.-R. Li, F. Santambrogio, and X.-J. Wang. Regularity in Monge’s mass transfer problem. *J. Math. Pures Appl.*, 102(6):1015–1040, 2014.
- [93] B. S. van Lith. *Principles of Computational Illumination Optics*. PhD thesis, Eindhoven University of Technology, Eindhoven, the Netherlands, 2017.

-
- [94] J. Liu. Light reflection is nonlinear optimization. *Calc. Var. Partial Differential Equations*, 46(3):861–878, 2013.
- [95] J. Liu and N. S. Trudinger. On the classical solvability of near field reflector problems. *Discrete Contin. Dyn. Syst.*, 36(2):895–916, 2016.
- [96] G. Loeper. On the regularity of solutions of optimal transportation problems. *Acta Math.*, 202(2):241–283, 2009.
- [97] G. Loeper and F. Rapetti. Numerical solution of the Monge-Ampère equation by a Newton’s algorithm. *Comptes rendus Mathématique*, 340(4):319–324, 2005.
- [98] Lumileds. LED Documents: Data sheets, 2021. Online; accessed June 15, 2021. <https://www.lumileds.com/support/documentation/datasheets/>.
- [99] R. K. Luneburg. *Mathematical Theory of Optics*. University of California Press, Berkeley, CA, USA, 1964.
- [100] D. Ma, Z. Feng, and R. Liang. Freeform illumination lens design using composite ray mapping. *Appl. Opt.*, 54(3):498–503, 2015.
- [101] M. J. J. J.-B. Maes. *Mathematical methods for reflector design*. PhD thesis, University of Amsterdam, Amsterdam, the Netherlands, 1997.
- [102] D. Malacara-Hernández. *Color Vision and Colorimetry: Theory and Applications*, volume 204 of *SPIE Press monograph*. SPIE, 2nd edition, 2011.
- [103] J. C. Maxwell. *A Dynamical Theory of the Electromagnetic Field*. Royal Society, 1865.
- [104] D. Michaelis, P. Schreiber, and A. Bräuer. Cartesian oval representation of freeform optics in illumination systems. *Opt. Lett.*, 36(6):918–920, 2011.
- [105] J. C. Miñano and J. C. Gonzalez. New method of design of nonimaging concentrators. *Appl. Opt.*, 31(16):3051–3060, 1992.
- [106] T. Möller and B. Trumbore. Fast, minimum storage ray-triangle intersection. *J. Graph. Tools*, 2(1):21–28, 1997.
- [107] G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l’Académie Royale des Sciences de Paris*, 1781.
-

- [108] I. Newton. *Optics: Treatise of Reflections, Refractions, Inflections & Colours of Light*. London: Printed for Sam. Smith, and Benj. Walford., 1704. Online; accessed June 15, 2021. <https://doi.org/10.5479/sil.302475.39088000644674>.
- [109] D. Nijkerk, B. van Venrooy, P. van Doorn, R. Henselmans, F. Draaisma, and A. Hoogstrate. The TROPOMI Telescope. In B. Cugny, E. Armandillo, and N. Karafolas, editors, *International Conference on Space Optics — ICSSO 2012*, volume 10564, pages 272–278. International Society for Optics and Photonics, SPIE, 2017.
- [110] G. Nöldeke and L. Samuelson. The implementation duality. *Econometrica*, 86(4):1283–1324, 2018.
- [111] A. M. Oberman. Convergent difference schemes for degenerate elliptic and parabolic equations: Hamilton-Jacobi equations and free boundary problems. *SIAM J. Numer. Anal.*, 44(2):879–895, 2006.
- [112] A. M. Oberman. Wide stencil finite difference schemes for the elliptic Monge-Ampère equation and functions of the eigenvalues of the Hessian. *Discrete Contin. Dyn. Syst. B*, 10(1):221, 2008.
- [113] H. Ohno. Design of a coaxial light guide producing a wide-angle light distribution. *Appl. Opt.*, 56(14):3977–3983, 2017.
- [114] H. Ohno and M. Kato. Total internal reflection shell for light-emitting diode bulbs. *Appl. Opt.*, 58(1):87–93, 2019.
- [115] V. I. Oliker. On reconstructing a reflecting surface from the scattering data in the geometric optics approximation. *Inverse Probl.*, 5(1):51–65, 1989.
- [116] V. I. Oliker. Optical design of freeform two-mirror beam-shaping systems. *J. Opt. Soc. Am. A*, 24(12):3741–3752, 2007.
- [117] V. I. Oliker. Controlling light with freeform multifocal lens designed with supporting quadric method (SQM). *Opt. Express*, 25(4):A58–A72, 2017.
- [118] V. I. Oliker, L. L. Doskolovich, and D. A. Bykov. Beam shaping with a plano-freeform lens pair. *Opt. Express*, 26(15):19406–19419, 2018.
- [119] V. I. Oliker and E. J. Newman. The energy conservation equation in the reflector mapping problem. *Appl. Math. Lett.*, 6(1):91–95, 1993.

- [120] V. I. Olikier, J. Rubinstein, and G. M. Wolansky. Supporting quadric method in optical design of freeform lenses for illumination control of a collimated light. *Adv. in Appl. Math.*, 62:160–183, 2015.
- [121] P. M. Pattison, M. Hansen, and J. Y. Tsao. LED lighting efficacy: Status and directions. *Comptes Rendus Physique*, 19(3):134–145, 2018.
- [122] G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.
- [123] W. H. Press, S. A. Teukolsky, B. P. Flannery, and W. T. Vetterling. *Numerical Recipes in Fortran 77: The Art of Scientific Computing*. Cambridge University Press, 2nd edition, 1996.
- [124] C. R. Prins. *Inverse Methods for Illumination Optics*. PhD thesis, Eindhoven University of Technology, Eindhoven, the Netherlands, 2014.
- [125] C. R. Prins, R. Beltman, J. H. M. ten Thije Boonkkamp, W. L. IJzerman, and T. W. Tukker. A least-squares method for optimal transport using the Monge-Ampère equation. *SIAM J. Sci. Comput.*, 37(6):B937–B961, 2015.
- [126] R. J. Purser and M. J. P. Cullen. A duality principle in semigeostrophic theory. *J. Atmos. Sci.*, 44(23):3449–3468, 1987.
- [127] H. Ries and J. Muschaweck. Tailored freeform optical surfaces. *J. Opt. Soc. Am. A*, 19(3):590–595, 2002.
- [128] H. Ries and A. Rabl. Edge-ray principle of nonimaging optics. *J. Opt. Soc. Am. A*, 11(10):2627–2632, 1994.
- [129] L. B. Romijn, J. H. M. ten Thije Boonkkamp, M. J. H. Anthonissen, and W. L. IJzerman. Generating-function approach for double freeform lens design. *J. Opt. Soc. Am. A*, 38(3):356–368, 2021.
- [130] L. B. Romijn, J. H. M. ten Thije Boonkkamp, M. J. H. Anthonissen, and W. L. IJzerman. An iterative least-squares method for generated Jacobian equations in freeform optical design. *SIAM J. Sci. Comput.*, 43(2):B298–B322, 2021.
- [131] L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Freeform lens design for a point source and far-field target. *J. Opt. Soc. Am. A*, 36(11):1926–1939, 2019.

- [132] L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Freeform lens design for a point source and far-field target. In *Optical Design and Fabrication 2019 (Freeform, OFT)*, page FT1B.2. Optical Society of America, 2019.
- [133] L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Inverse reflector design for a point source and far-field target. *J. Comput. Phys.*, 408:109283, 2020.
- [134] L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. A Monge-Ampère least-squares solver for the design of a freeform lens. In F. J. Vermolen and C. Vuik, editors, *Numerical Mathematics and Advanced Applications ENUMATH 2019*, pages 833–840, Lecture Notes in Computational Science and Engineering, 2020. Springer International Publishing.
- [135] A. H. van Roosmalen, M. J. H. Anthonissen, W. L. IJzerman, and J. H. M. ten Thije Boonkkamp. Design of a freeform two-reflector system to collimate and shape a point source distribution. *Opt. Express*, 29(16):25605–25625, 2021.
- [136] L. Rüschendorf and S. T. Rachev. A characterization of random variables with minimum L^2 -distance. *J. Multivar. Anal.*, 32(1):48–54, 1990.
- [137] A. I. Sabra and Abd al-Hamid Sabra. *Theories of Light: From Descartes to Newton*. Cambridge University Press, 1981.
- [138] F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progr. Nonlinear Differential Equations 87. Birkhäuser, Springer, Cham, 1st edition, 2015.
- [139] E. F. Schubert, T. Gessmann, and J. K. Kim. *Light Emitting Diodes*. American Cancer Society, 2005.
- [140] C. S. Smith and M. Knott. Note on the optimal transportation of distributions. *J. Optim. Theory Appl.*, 52(2):323–329, 1987.
- [141] J. H. M. ten Thije Boonkkamp, L. B. Romijn, and W. L. IJzerman. Generalized Monge-Ampère equations for illumination freeform design. In *Optical Design and Testing IX*, volume 11185, page 1118504. International Society for Optics and Photonics, SPIE, 2019.
- [142] J. H. M. ten Thije Boonkkamp, L. B. Romijn, N. K. Yadav, and W. L. IJzerman. Monge-Ampère type equations for freeform illumination optics. In T. E. Kidger and S. David, editors, *Illumination Optics V*, volume

-
- 10693, pages 52–66. International Society for Optics and Photonics, SPIE, 2018.
- [143] The Art Institute of Chicago. Vincent van Gogh: Self-Portrait, 1887. Online; accessed April 7, 2021. <https://www.artic.edu/artists/40610/vincent-van-gogh>.
- [144] J.-P. Tignol. *Galois' Theory of Algebraic Equations*. Longman Scientific & Technical, 1988.
- [145] N. S. Trudinger. On the local theory of prescribed Jacobian equations. *Discrete Contin. Dyn. Syst.*, 34(4):1663–1681, 2012.
- [146] J. I. E. Urbas. On the second boundary value problem for equations of Monge-Ampère type. *J. Reine Angew. Math.*, 487:115–124, 1997.
- [147] C. P. T. Villani. *Topics in Optimal Transportation*, volume 58. American Mathematical Society, 2003.
- [148] C. P. T. Villani. *Optimal Transport: Old and New*. Grundlehren Math. Wiss. 338. Springer, Berlin, Heidelberg, 2009.
- [149] B. Vohnsen. A short history of optics. *Phys. Scr.*, 2004(T109):75, 2004.
- [150] K. Wang, F. Chen, Z. Liu, X. Luo, and S. Liu. Design of compact freeform lens for application specific light-emitting diode packaging. *Opt. Express*, 18(2):413–425, 2010.
- [151] X.-J. Wang. On the design of a reflector antenna. *Inverse Probl.*, 12(3):351, 1996.
- [152] X.-J. Wang. On the design of a reflector antenna II. *Calc. Var. Partial Differential Equations*, 20(3):329–341, 2004.
- [153] S. Wei, Z. Zhu, Z. Fan, Y. Yan, and D. Ma. Double freeform surfaces design for beam shaping with non-planar wavefront using an integrable ray mapping method. *Opt. Express*, 27(19):26757–26771, 2019.
- [154] R. R. Wessels. A least-squares algorithm for the Monge-Ampère equation: analysis and application to freeform optics. Master's thesis, Eindhoven University of Technology, Eindhoven, the Netherlands, 2019.
- [155] R. Winston, J. C. Miñano, P. G. Benitez, et al. *Nonimaging Optics*. Elsevier, 2005.
-

- [156] K. B. Wolf. *Geometric Optics on Phase Space*. Theoretical and Mathematical Physics. Springer Science & Business Media, 1st edition, 2004.
- [157] R. Wu and H. Hua. Direct design of aspherical lenses for extended non-Lambertian sources in three-dimensional rotational geometry. *Opt. Express*, 24(2):1017–1030, 2016.
- [158] R. Wu, C. Y. Huang, X. Zhu, H.-N. Cheng, and R. Liang. Direct three-dimensional design of compact and ultra-efficient freeform lenses for extended light sources. *Optica*, 3(8):840–843, 2016.
- [159] R. Wu, P. Liu, Y. Zhang, Z. Zheng, H. Li, and X. Liu. A mathematical model of the single freeform surface design for collimated beam shaping. *Opt. Express*, 21(18):20974–20989, 2013.
- [160] R. Wu, L. Xu, P. Liu, Y. Zhang, Z. Zheng, H. Li, and X. Liu. Freeform illumination design: A nonlinear boundary problem for the elliptic Monge-Ampère equation. *Opt. Lett.*, 38(2):229–231, 2013.
- [161] N. K. Yadav. *Monge-Ampère Problems with Non-Quadratic Cost Function: Application to Freeform Optics*. PhD thesis, Eindhoven University of Technology, Eindhoven, the Netherlands, 2018.
- [162] N. K. Yadav, L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. A least-squares method for the design of two-reflector optical systems. *J. Phys. Photonics*, 1(3):034001, 2019.
- [163] N. K. Yadav, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. A least-squares method for a Monge-Ampère equation with non-quadratic cost function. *Numerical Mathematics and Advanced Applications ENUMATH 2017*, pages 301–309, 2019.
- [164] N. K. Yadav, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. A Monge-Ampère problem with non-quadratic cost function to compute freeform lens surfaces. *J. Sci. Comput.*, 80(1):475–499, 2019.
- [165] L. Yang, Y. Liu, Z. Ding, J. Zhang, X. Tao, Z. Zheng, and R. Wu. Design of freeform lenses for illuminating hard-to-reach areas through a light-guiding system. *Opt. Express*, 28(25):38155–38168, 2020.
- [166] Y. Zhang, R. Wu, P. Liu, Z. Zheng, H. Li, and X. Liu. Double freeform surfaces design for laser beam shaping with Monge-Ampère equation method. *Opt. Commun.*, 331:297–305, 2014.

Index

- angular characteristic T , 42, 77, 84, 210, 213
- anisotropic, 21
- Brewster's angle, 22
- c-concave function, 107
- c-convex analysis, 100, 106
- c-convex function, 107
- c-transform, 106
- chromaticity, 50
- conjugate pair, 102
- convex analysis, 100, 102
- convex function, 102
- convex set, 102
- cost function, 58, 99, 109, 147
- coupled boundary value problem, 165, 171
- critical angle, 29
- Descartes' sphere, 31
- direction cosines, 30
- dispersion, 25
- displacement current, 13
- étendue, 33
- eikonal equation, 17
- emittance, 51
- extended light source, 235
- far field, 53
- far-field approximation, 54, 59
- Fermat's principle, 21
- flux density, 50
- freeform optics, 4
- Fresnel reflections, 23, 236
- G-concave function, 117, 124, 239
- G-convex analysis, 100, 114
- G-convex function, 117, 124, 239
- G-exponential map, 118
- G-transform, 114
- generalized light source, 235
- generalized Monge-Ampère equation, 99, 113, 123
- generated Jacobian equation, 100, 115, 120–123
- generating function, 58, 116, 173
- H-concave function, 117, 124, 239
- H-convex function, 117, 124, 239
- Hamilton's characteristics, 29–51
- Hamiltonian, 30
- Hamiltonian system, 22, 29, 31
- Huygens-Fresnel principle, 26
- illuminance, 51
- imaging illumination optics, 3
- intermediate target intensity, 205, 206, 226
- irradiance, 49
- isotropic, 20

- Lambertian light source, 4, 226
- law of reflection, 22, 129
- law of refraction, 25
- LED, 1
- Legendre transformation, 33, 40, 42, 43
- Legendre-Fenchel transform, 102, 107, 117, 118
- light ray, 18
- luminous efficacy, 51
- luminous flux, 51
- luminous intensity, 51

- mixed characteristic of the first kind W , 40, 68
- mixed characteristic of the second kind W^* , 42, 88
- momentum, 31
- monochromaticity, 51

- near field, 53, 54
- Neumann problem, 166, 177
- nonimaging illumination optics, 3

- optical direction cosines, 30
- optical mapping
 - explicit, 71, 75, 82, 126
 - implicit, 111, 124, 126
 - initial guess, 149, 170
- optimal transport theory, 58, 108

- phase space, 33
- photometric quantities, 51
- plane of incidence, 23, 26, 34
- plane wave, 18

- point characteristic V , 37, 74
- polarization, 16
- Poynting vector, 15
- push-forward, 101, 109, 119

- radiant exitance, 50
- radiant flux, 49
- radiant intensity, 50
- radiometric quantities, 50
- ray equation, 20

- scattering, 6, 22, 236
- second boundary value problem, 119
- shipper's problem, 110, 118, 123
- SMS method, 141
- Snell's law, 25
- solid angle, 50
- spectral luminous flux, 51
- spectral radiant flux, 50
- SQM method, 138
- standard Monge-Ampère equation, 72, 105, 115
- steradian, 50
- stereographic projection, 54, 60, 128
- stopping criterion, 183

- Theorem of Malus and Dupin, 21, 53, 87
- transport boundary condition, 73, 76, 82, 86, 91, 119, 127

- zero étendue, 4, 47

Summary

Generated Jacobian Equations: Mathematical Theory and Numerics

The popularity of LED lighting systems has increased significantly in the last decade. A major advantage of using LEDs over conventional light sources is that an LED light operates at lower temperatures and plastic materials can be used for the optical components of the lamp. Thus, complicated shapes are possible. Modern optical components are freeform (i.e., non-axially symmetric) reflectors and lenses that transform the light from the LED source into the required light output of the lighting system.

Given a particular source light distribution and a given target light distribution, what optical system should be placed in between? In this work an inverse approach is used. The inverse approach based on first principles uses geometrical optics and conservation of energy to derive a partial differential equation (PDE) for the optical surface. An optical mapping can be constructed that connects coordinates on the source and target domains. Substituting the mapping into the relation for energy conservation leads to a nonlinear second-order elliptic PDE for the unknown location of the optical surface.

In this thesis, I developed a generic framework to derive second-order nonlinear PDEs for freeform illumination optics. I applied this framework to optical systems that can be described using a cost function in optimal transport theory. These cost functions can be derived using Hamilton's characteristic functions of optical path length. The simplest quadratic cost function corresponds to parallel-to-far-field systems and the standard Monge-Ampère equation as the second-order nonlinear elliptic PDE. I derived the cost functions for a range of optical systems. The associated PDE for these systems is also called a generalized Monge-Ampère equation.

However, many optical systems cannot be cast as optimal-transport problems, e.g., systems involving near-field targets or multiple freeform surfaces. For these systems I generalized the concept of a cost function to a generating function. In these cases, the PDE for the location of the optical surface is a so-called generated Jacobian equation. I laid the foundation for a solution strategy that can handle any optical system described by a generating function.

I presented an overview of 16 base-case optical systems that all fit into this framework.

The generalized Monge-Ampère equations were solved numerically using a generalized least-squares (GLS) algorithm. Originally, this method was developed for the standard Monge-Ampère equation by Corien Prins, who graduated in 2014, considering the parallel-to-far-field problem. It was extended to an optimal-transport framework to solve generalized Monge-Ampère equations by Nitin Yadav, who graduated in 2018, and myself. The GLS algorithm is presented in detail in this work. Later on, the generalized numerical procedure was extended to a generating-function framework, namely, the generated Jacobian least-squares (GJLS) algorithm. The GJLS algorithm is now one of the first algorithms capable of solving generated Jacobian equations. Moreover, to the best of our knowledge, our algorithm can handle the widest variety of optical systems.

I tested the GLS and GJLS algorithms on numerous problems, both examples to test the accuracy and efficiency of the algorithms and practical applications. The development of these algorithms has opened up new possibilities for system design. In Chapter 7, I included a numerical experiment to distribute the refractive power of a lens with a point source and far-field target over two freeform surfaces. To compute the first optical surface an intermediate far-field target intensity needs to be defined. By choosing different intermediate target intensities as an interpolation between the source and final target intensity based on an initial approximate mapping, I have introduced a tuning parameter to distribute the refractive power over the two lens surfaces. Varying this parameter allows for the design of more compact optical systems.

List of Publications

Journal articles

6. M. J. H. Anthonissen, L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. A unified mathematical framework for a class of fundamental freeform optical systems. In review. *Opt. Express*, 2021.
5. L. B. Romijn, J. H. M. ten Thije Boonkkamp, M. J. H. Anthonissen, and W. L. IJzerman. Generating-function approach for double freeform lens design. *J. Opt. Soc. Am. A*, 38(3):356–368, 2021.
4. L. B. Romijn, J. H. M. ten Thije Boonkkamp, M. J. H. Anthonissen, and W. L. IJzerman. An iterative least-squares method for generated Jacobian equations in freeform optical design. *SIAM J. Sci. Comput.*, 43(2):B298–B322, 2021.
3. L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Inverse reflector design for a point source and far-field target. *J. Comput. Phys.*, 408:109283, 2020.
2. L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Freeform lens design for a point source and far-field target. *J. Opt. Soc. Am. A*, 36(11):1926–1939, 2019.
1. N. K. Yadav, L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. A least-squares method for the design of two-reflector optical systems. *J. Phys. Photonics*, 1(3):034001, 2019.

Conference contributions

5. L. B. Romijn, M. J. H. Anthonissen, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Mathematics for point source freeform tailoring. In *Optical Design and Fabrication 2021 (Freeform, IODC, OFT)*, page JTh1A.4. Optical Society of America, 2021.

4. L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. A Monge-Ampère least-squares solver for the design of a freeform lens. In F. J. Vermolen and C. Vuik, editors, *Numerical Mathematics and Advanced Applications ENUMATH 2019*, pages 833–840. Lecture Notes in Computational Science and Engineering, Springer International Publishing, 2020.
3. L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Freeform lens design for a point source and far-field target. In *Optical Design and Fabrication 2019 (Freeform, OFT)*, page FT1B.2. Optical Society of America, 2019.
2. J. H. M. ten Thije Boonkkamp, L. B. Romijn, and W. L. IJzerman. Generalized Monge-Ampère equations for illumination freeform design. In *Optical Design and Testing IX*, volume 11185, page 1118504. International Society for Optics and Photonics, 2019.
1. J. H. M. ten Thije Boonkkamp, L. B. Romijn, N. K. Yadav, and W. L. IJzerman. Monge-Ampère type equations for freeform illumination optics. In T. E. Kidger and S. David, editors, *Illumination Optics V*, volume 10693, pages 52–66. International Society for Optics and Photonics, SPIE, 2018.

Oral presentations at scientific conferences

10. L. B. Romijn, M. J. H. Anthonissen, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Mathematics for point source freeform tailoring. In *Optical Design and Fabrication 2021 (Freeform, IODC, OFT), JTh1A.4 - Joint Freeform and IODC I*, Optical Society of America, Online conference, 28 June – 1 July, 2021.
9. L. B. Romijn. Mathematics for the design of advanced freeform optical systems. Nominated for the *Dutch Royal Mathematical Society (KWG) Prize*, Dutch Mathematical Congress (NMC), Online competition, 20 May, 2021.
8. L. B. Romijn, M. J. H. Anthonissen, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Generalized Monge-Ampère equations in freeform optical design. In *Parallel session PW2A Atomic, molecular and optical physics V*, Physics @ Veldhoven, Online conference, 18 – 20 January, 2021.

7. L. B. Romijn, M. J. H. Anthonissen, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Generalized Monge-Ampère equations in freeform optical design. In *MS47 Generalized Monge-Ampère Equations in Illumination Optics - Part II of II*, SIAM-CAIMS Annual Meeting, Online conference, 6 – 17 July, 2020.
6. L. B. Romijn, M. J. H. Anthonissen, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Numerically solving generated Jacobian equations in freeform optical design. In *TOM 2 - Computational, Adaptive and Freeform Optics*, European Optical Society Biennial Meeting (EOSAM), Online conference, 7 – 11 September, 2020.
5. L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. A Monge-Ampère least-squares solver for the design of a freeform lens. In *MS19: Numerical methods for Monge-Ampère equations (Part 1)*, ENU-MATH, Egmond aan Zee, The Netherlands, 30 September – 4 October, 2019.
4. L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Freeform lens design for a point source and far-field target. In *Optical Design and Fabrication 2019 (Freeform, OFT), FT1B - Freeform Illumination I*, Optical Society of America, Washington DC, USA, 10 – 12 June, 2019.
3. L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Monge-Ampère elliptic equations for freeform optics. SCS (Dutch-Flemish Scientific Computing Society) Spring Meeting, Antwerp, 17 May, 2019.
2. L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Inverse reflector problem for a point source. ILIAD, Eindhoven, The Netherlands, 13 November, 2018.
1. L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Inverse reflector problem for a point source. In *TOM 2 - Freeform Optics for Illumination, Augmented Reality and Virtual Reality*, European Optical Society Biennial Meeting (EOSAM), Delft, The Netherlands, 8 – 12 October, 2018.

Other publications

2. No more blinding lights: a PhD in smart lighting. In *Supremum*, 53(2), Gewis, Eindhoven University of Technology, 2021.
1. The design of freeform optical surfaces for LED-based applications. In *ILI Glow Magazine 2020*, Edition 13, pages 14–15. Intelligent Lighting Institute, 2020.

Curriculum Vitae



Lotte Romijn was born on the 7th of April 1993 in Eindhoven, the Netherlands. After finishing her gymnasium degree in 2011 at the Augustinianum in Eindhoven, the Netherlands, she studied Liberal Arts and Sciences at Amsterdam University College, the Netherlands. In 2014, she graduated and moved to Melbourne, Australia, to pursue her Master's degree in Applied Mathematics and Mathematical Physics at the University of Melbourne. For her Master's thesis she studied the biomechanics of the small-intestinal crypt by developing computational models of tissue monolayers. She won a prize for best Master's thesis in Applied Mathematics (Professor Wilson Prize), graduated *summa cum laude* and was the department's valedictorian in 2016.

In 2017, Lotte started a PhD project in the Computational Illumination Optics group at Eindhoven University of Technology (TU/e), the Netherlands. Her work involves the design of freeform reflector and lens surfaces for LED-based applications, under the supervision of Assoc. Prof. Jan ten Thije Boonkkamp, Asst. Prof. Martijn Anthonissen and Prof. Wilbert IJzerman. The results of this work are presented in this dissertation. Highlights of her PhD trajectory are the publication of five journal articles and five conference contributions, a nomination for the KWG PhD Prize of the Dutch Royal Mathematical Society, an invitation to present as invited speaker at a conference (OSA Freeform Optics, 28 June – 1 July, 2021, Rhode Island, USA), and her participation as secretary and treasurer of the PhD Council of the Department of Mathematics and Computer Science at the TU/e. As part of her PhD studies she undertook a Course on Modern Optics for Optical Designers: "CMOP part 1", High Tech Institute, September 2018 - February 2019. She also attended two summer schools on optimal transport theory: "Optimal transport: Numerical methods and applications" at Lake Como School of Advanced Studies in Como, Italy, 7 – 11 May, 2018, and "A numerical introduction to optimal transport" at Inria in Paris, France, 13 – 17 May, 2019.

Acknowledgments

I would like to express my thanks to a number of very important people:

To my supervisors Jan, Wilbert and Martijn. Jan's meticulousness, Wilbert's clear-eyed sense of practicality and Martijn's calm mediation are a match made in heaven.

To Enna, Diane and Jolijn, for all their administrative support.

To my CASA colleagues. I always enjoyed our lunch chats, the 'meerkamp', virtual 'borrels' and attending conferences together.

To my office mates Rien and Vi. It is always better to feel the need to punch your computer monitor together than alone.

To my grandparents, for being proud no matter what.

To my parents, my sister and her boyfriend, and Huub's family. Thank you for your support.

Especially to my mother, my powerhouse.

To Huub, for lighting up my life.