Numerical Methods for the Hyperbolic Monge-Ampère Equation with Applications to Optical Design

Numerical Methods for the Hyperbolic Monge-Ampère Equation with Applications to Optical Design

Maikel Bertens

Numerical Methods for the Hyperbolic Monge-Ampère Equation with Applications to Optical Design

Maikel Wilhelmus Marinus Cornelus Bertens

Maikel Wilhelmus Marinus Cornelus Bertens
Numerical Methods for the Hyperbolic Monge-Ampère Equation with Applications to Optical Design
Eindhoven University of Technology, 2023

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.

This work is part of the research programme Nederlandse Organisatie voor-Wetenschappelijk Onderzoek Toegepaste en Technische Wetenschappen (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: 978-90-386-5811-7

Cover design by Maikel Bertens (see Figure 4.18 of this thesis).
Printing by: Gildeprint - Enschede.

Numerical Methods for the Hyperbolic Monge-Ampère Equation with Applications to Optical Design

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof. dr. S. K. Lenaerts, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op woensdag 27 september 2023 om 16:00 uur

door

Maikel Wilhelmus Marinus Cornelus Bertens

geboren te 's-Hertogenbosch

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

voorzitter:	prof. dr. M. G. J. van den Brand
1e promotor:	prof. dr. ir. W. L. IJzerman
2e promotor:	dr. ir. J. H. M. ten Thije Boonkkamp
copromotor:	dr. ir. M. J. H. Anthonissen
leden:	prof. dr. M. A. Peletier
	prof. prof. dr. ir. R. W. C. P. Verstappen (RUG)
	prof. dr. J. D. Benamou (INRIA)

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

Amor fati

Friedrich W. Nietzsche

Contents

Chapter 1

Introduction

1.1 Motivation

To quote the Nobel Laureate Werner Karl Heisenberg (1967): "I wish to emphasize again that the progress of physics certainly will depend to a large extent on the progress of nonlinear mathematics, of methods of solving nonlinear equations" [40]. Our particular interest lies in one such nonlinear mathematical problem, viz. the Monge-Ampère equation and, more specifically, the design of freeform optical surfaces.

Since the introduction of light emitting diodes (LEDs) the luminous efficacy has increased from 10.3 lumen per watt (lm/W) for the modern incandescent light bulbs [63] to over 210 lm/W for LED replacement bulbs [64]. In a traditional incandescent light bulb a wire filament is heated until it glows. This requires a different physical setup to produce light efficiently compared to typical LED systems, because LEDs are comparatively small and produce less heat. This requires us to rethink the design of optical systems.

It has been found that the design of freeform optical systems, in which the light of a given light source distribution is transferred to a desired target distribution, can be mathematically described by the Monge-Ampère equation [65]. This equation has in recent years been subject to increasing research [5]. One reason for this is its wide applicability to, for example, differential geometry, calculus of variations, economics, meteorology, optimal transport, fluid dynamics, mathematical finance, aerodynamics, hydrodynamics, filtration theory, mesh generation and geometrical optics [9, 13, 49, 58].

For the design of freeform optical systems sixteen base cases have been derived [3]. These base cases are fundamental components for developing complex freeform optical systems. These base cases satisfy either a Monge-Ampère equation or a generalization of the Monge-Ampère equation viz. a

generated Jacobian equation. In this thesis we focus on the Monge-Ampère equation. For the design of freeform optical systems the Monge-Ampère equation comes in two variants, an elliptic and a hyperbolic one. Furthermore, for each variant and each system multiple distinct optical surfaces can be calculated. For elliptic solutions the freeform surfaces are either convex or concave and for hyperbolic solutions the surfaces are saddle shaped. Figure 1.1 shows three of these examples.

The increasing interest in the Monge-Ampère equation did significantly boost the number of numerical methods for the elliptic Monge-Ampère equation, some of which have successfully been used to construct freeform optical systems (see [68] for example). Unfortunately, the hyperbolic counter variant has received little attention. For optical systems, these solutions are relevant for, among others, reducing glare and light pollution. To see this, consider a lens array. Such an array consist of multiple LEDs, arranged in a grid-like structure, with lenses on top. In Figure 1.2 the freeform optical surfaces of two such structures are shown. On the left only concave elements are used, while on the right convex, concave and saddle elements are used. Using solely convex or concave elements, i.e., elliptic solutions, leads to an optical surface with discontinuities in the gradient, i.e., cusps. These cusps are hard and thus expensive to manufacture and are prone to manufacturing errors. Furthermore, even error-free cusps lead to unwanted scattering of light which in term contributes to light pollution and may cause glare. The addition of saddle shaped optical surfaces, and hence hyperbolic solutions to the Monge-Ampère equation, allows for surfaces with continuous gradients as shown on the right in Figure 1.2. These surfaces allow for smoother light profiles and better control of the light output of the optical systems.

A second application of such continuous surfaces arises from satellite communication. In this field light has to be sent over fast distances requiring immense precision and accurate detection. Cusps in the optical system of both a sending and receiving satellite may scatter the light in such a way that the message is irretrievable from the signal.

In this thesis the design of better optical systems is the main reason to develop numerical methods for the hyperbolic Monge-Ampère equation. The wide applicability of the Monge-Ampère equation proves to be a beneficial secondary effect for the scientific community at large.

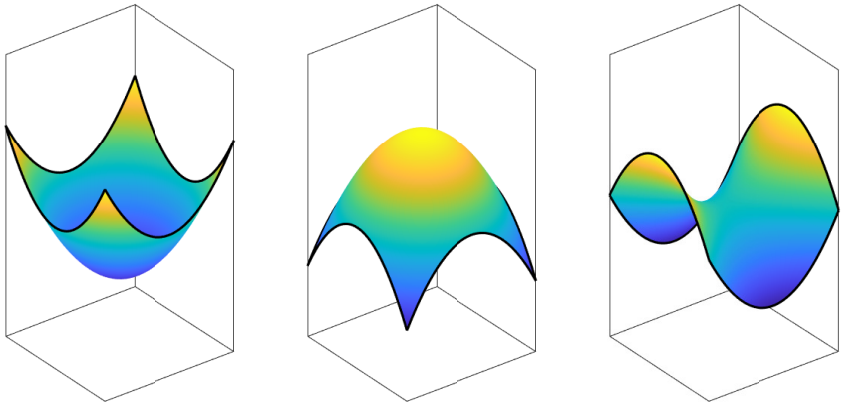


Figure 1.1: Optical surfaces which are either convex (left), concave (middle) or saddle shaped (right).

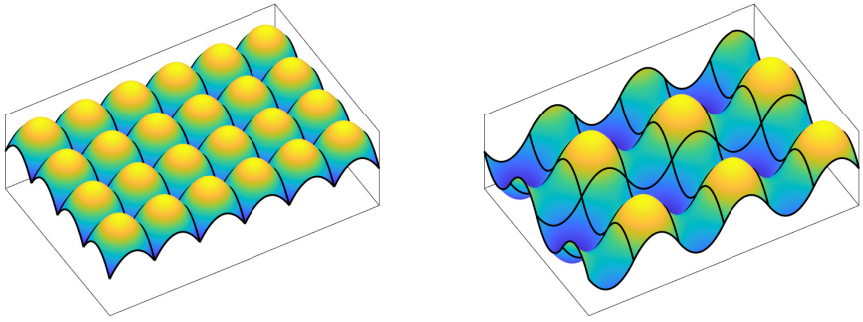


Figure 1.2: Lens array consisting of solely concave surfaces (left) and a lens array consisting of convex, concave and saddle surfaces (right).

1.2 Outline of this thesis

This thesis is organized as follows.

In Chapter 2, we introduce a few general properties of the Monge-Ampère equation. We start with a classification of quasi-linear PDEs in two variables as parabolic, elliptic or hyperbolic type. Subsequently we expand this categorization to nonlinear PDEs. This has proven to be non-trivial and thus we introduce multiple classifications which, for the Monge-Ampère equation, turn out to be equivalent. With these classifications in hand we consider

the null space of the Hessian, which is equivalent to a particular parabolic Monge-Ampère equation and derive solutions. Next, we consider solutions to the hyperbolic Monge-Ampère equation via complex functions. These analytical solutions are later used to test our numerical methods. One additional method of generating solutions is found via an integral formulation of the Monge-Ampère equation, which we derive next. The solutions to the parabolic Monge-Ampère equation and the solutions to the hyperbolic Monge-Ampère equation all show a lot of symmetry. Furthermore, if the relevant domains and the boundary conditions contain certain symmetries, we can find multiple solutions to the same Monge-Ampère equation. This is shown next. We conclude the chapter by a literature overview on numerical methods for the elliptic and hyperbolic Monge-Ampère equation.

In Chapter 3, we introduce the fundamentals to the method of characteristics. We first define characteristics and derive the characteristic condition. From this we derive the evolution of the solution along the characteristics, which, for a general second-order hyperbolic PDE, results in two mutually coupled systems of ordinary differential equations (ODEs). Afterwards we derive the ODE systems for the hyperbolic Monge-Ampère equation and discuss the necessary boundary conditions for the Monge-Ampère equation.

Three numerical methods for the coupled systems of ODEs are presented in Chapter 4. These methods are based on explicit one-step methods, viz. forward Euler, modified Euler and Runge-Kutta. We integrate the ODEs from one (vertical) grid line to the next and subsequently interpolate the solution along this second grid line. By repeating this process a numerical solution can be computed on the whole domain. Because analytical solutions are not always known, we utilize an integral formulation of the Monge-Ampère equation to evaluate the error of the numerical approximation of the solution. The numerical errors, that is both the error obtained by the integral formulation and the global discretization error, depend on both the one-step method and the spline interpolant used. We introduce a method to determine the distance between grid lines such that neither of those errors dominates the other. This allows us to estimate and test the rate of convergence for the numerical method for integration and interpolation combined. We present various examples, among which an example where all boundary segments require different boundary conditions, one example where the number of required boundary conditions varies along the boundary, and one example with discontinuous third derivatives, for which no analytical solution is known. Lastly, we also show three cases for the general Monge-Ampère equation where we not only consider integration in the positive x-direction, but integration in all four cardinal directions.

For designing optical systems the Monge-Ampère equation needs to satisfy the transport boundary condition, requiring that the gradient of the solution is mapped from the boundary of the domain to the boundary of an a priori fixed codomain. In Chapter 5, we introduce a least-squares method for the hyperbolic Monge-Ampère equation with transport boundary condition. The least-squares method is known in the literature and used to solve the elliptic Monge-Ampère equation. First, we discuss the theory of the least-squares method for the hyperbolic Monge-Ampère equation and subsequently adapt parts of the method, viz. the optimization in the interior domain, the boundary methods and we introduce two grid-shock correction methods. We present numerical results and compare three boundary methods, show their weaknesses and strengths and elaborate on the convergence of the algorithm for various test cases.

In Chapter 6, we discuss the design of optical systems. We first introduce the fundamentals of geometrical optics. Because light is an electromagnetic wave, it behaves according to the Maxwell equations. By applying a short-wavelength approximation, we find that light satisfies the so-called eikonal equation and thus behaves as rays when the optical elements are much larger than the wavelength of light. From the eikonal equation we subsequently derive the ray equation, describing the path of the ray. The ray equation is not ideal to work with in practice, and therefore we introduce the equivalent Fermat's principle, also known as the principle of shortest optical path. By the Euler-Lagrange equations for Fermat's principle, we move to Hamiltonian optics. With the use of Hamilton's characteristics, we are able to formulate the design of four optical systems as a Monge-Ampère equation. These systems satisfy a so-called cost balance, which stems from optimal transport theory.

In Chapter 7 we present an adaptation of the least-squares solver of Chapter 5 to include the cost balance. This adjusted least-squares solved is subsequently used to calculate optical designs of four distinct optical design problems. We show that for each optical design problem multiple solutions exists. We verify the optical designs by ray-tracing the obtained solution and compare the error of the ray-traced and exact target distribution.

In Chapter 8, we conclude this thesis with a summary and give recommendations for further research.

Chapter 2

The Monge-Ampère Equation

In this chapter we introduce the Monge-Ampère (MA) equation, a fully non-linear partial differential equation (PDE). Applications of the Monge-Ampère equation are found, a.o., in fluid dynamics to compute the velocity of an incompressible fluid from the pressure using the streamline formulation [48], in mathematical finance to determine optimal portfolio strategies [14] and in Riemannian geometry to compute the surface of a manifold given the Gauss curvature [18]. Furthermore, as we discuss in Chapter 6, the MA equation has applications in freeform optical design. For a good introduction to the Monge-Ampère equation, the different classes of solutions and some applications, we refer to Figalli's book [30].

The general Monge-Ampère equation for a variable u in two independent variables x, y is of the form

$$A_0(u_{xx}u_{yy} - u_{xy}^2) + A_1u_{xx} + A_2u_{xy} + A_3u_{yy} + A_4 = 0, \tag{2.1}$$

where A_0, A_1, A_2, A_3 and A_4 are functions, possibly dependent on x, y, u, u_x and u_y. The linearity in the Hessian $u_{xx}u_{yy} - u_{xy}^2$ is the defining feature of the Monge-Ampère equation and henceforth we assume $A_0 = 1$, unless specified otherwise. We adhere to the notation originally introduced by Monge, viz.

$$p = u_x, \qquad q = u_y, \qquad r = u_{xx}, \qquad s = u_{xy}, \qquad t = u_{yy}. \tag{2.2}$$

Instead of discussing (2.1) in its full generality, we often simplify it to the case $A_0 = 1$, $A_i = 0$ for $i = 1, 2, 3$ and $A_4 = \pm f^2$ with $f \in C(\mathbb{R}^2)$ and $f = f(x, y)$. We call (2.1) for this simplified case the standard Monge-Ampère equation.

For quasi-linear PDEs, i.e., partial differential equations which are linear with respect to all the highest order derivatives of the unknown function, standard definitions for classification as either elliptic, parabolic or hyperbolic

type exist. We present a generalization of those classifications to the nonlinear case here, that is, a generalization to PDEs which contain at least one term that is not linear in the highest order derivative. Next, we present parabolic solutions to the standard MA. These solutions are also known as the null-space of the Hessian as for this case $f \equiv 0$ and the Monge-Ampère operator coincides with the Hessian. We continue with a method to construct solutions for the hyperbolic Monge-Ampère equation via complex-valued functions. Subsequently, we present an equivalent integral formulation for the hyperbolic Monge-Ampère equation and use it to construct more solutions. Afterwards, we discuss symmetries of the Monge-Ampère equation where we also briefly elaborate on the elliptic variant. We conclude with a literature overview on numerical methods for the MA equation.

2.1 Classification

PDEs are often classified as either elliptic, hyperbolic, or parabolic. The classification proves useful as PDEs of the same type share characteristic traits, both mathematical and physical. For example, proving existence of solutions for elliptic PDEs often involves variants of the Lax-Millgram theorem, while the method of characteristics only works for hyperbolic PDEs. Furthermore, stability of finite difference methods varies per type when numerically solving the PDE. Physically, we have for hyperbolic PDEs wave-like behavior, where discontinuities in the boundary conditions propagate as discontinuities in the solution. Each point in the domain is only influenced by its so-called domain of dependence. Also, if the PDE is nonlinear, then shocks may develop even though the boundary conditions are smooth [55, p. 120]. In contrast, the solutions of elliptic PDEs are generally smooth even if the boundary conditions are not and the boundary conditions influence each point in the domain. Parabolic PDEs are usually time-dependent phenomena, representing diffusion-like processes including heat conduction and particle diffusion. Solutions are often smooth but singularities, e.g., due to point sources, may exist in space or time.

Before we classify the Monge-Ampère equation, let us consider the quasi-linear second order PDE

$$A_1 u_{xx} + A_2 u_{xy} + A_3 u_{yy} + A_4 = 0, \tag{2.3}$$

where $A_i = A_i(x, y)$ and $A_4 = A_4(x, y, u, p, q)$. Replacing u_{xx}, u_{xy} and u_{yy} by x^2, xy and y^2 in (2.3), respectively, we obtain

$$A_1 x^2 + A_2 xy + A_3 y^2 + A_4 = 0, \tag{2.4}$$

which describes either an ellipsoid, paraboloid or hyperboloid. We henceforth classify the PDE (2.3) as elliptic if (2.4) is an ellipsoid and similar for the other cases. We thus obtain the classification based on $\Delta = A_2^2 - 4A_1A_3$, viz.

$$\begin{cases} \text{elliptic} & \text{if } \Delta < 0 \\ \text{parabolic} & \text{if } \Delta = 0 \\ \text{hyperbolic} & \text{if } \Delta > 0. \end{cases} \tag{2.5}$$

Different but equivalent definitions exist, see for example [24, p. 312, 372, 399].

As the focus of this thesis is the hyperbolic Monge-Ampère equation, which is a fully nonlinear equation, we next define hyperbolicity in the context of general second order PDEs, which proves to be no simple matter. We quote, regarding linear versus nonlinear PDEs: "The mathematical methods devised to deal with these two classes of equations are often entirely different, and the behavior of solutions differs substantially." [55, p 8]. These difficulties have brought forth multiple different definitions for hyperbolicity, see for example [80, § 4] for a thorough consideration and the accompanying subtleties. Here we present a classification by linearizing the PDE and subsequently applying the classification for quasi-linear PDEs.

Let $F(x, y, u, p, q, r, s, t) = 0$ be a general second order nonlinear PDE. We classify $F = 0$ by classifying the linearization of the principle part, i.e., we classify $F = 0$ by classifying the second order nonlinear differential operator $L(u) := F(u_{xx}, u_{xy}, u_{yy})$, which formally represents $F(x, y, u, p, q, r, s, t)$ with x, y, u, p and q fixed. The Gateaux derivative [74, p. 121] of L reads

$$L_u(v) = F_r(u_{xx}, u_{xy}, u_{yy})v_{xx} + F_s(u_{xx}, u_{xy}, u_{yy})v_{xy} + F_t(u_{xx}, u_{xy}, u_{yy})v_{yy}, \tag{2.6}$$

at the solution u, where the derivatives F_r, F_s and F_t are to be understood as derivatives of F w.r.t. the first, second and third argument of $F(u_{xx}, u_{xy}, u_{yy})$, which we identify with the derivatives of $F(x, y, u, p, q, r, s, t)$ w.r.t. r, s and t, respectively. The equation $L_u(v) = 0$ is a linearization of the principle part of $F = 0$ and is a quasi-linear second order PDE in v. It can therefore be classified as before, yielding

$$\Delta = F_s^2 - 4F_rF_t, \tag{2.7}$$

and the classification (2.5). For the Monge-Ampère equation (2.1) we find

$$F_r = A_0t + A_1, \tag{2.8a}$$
$$F_s = -2A_0s + A_2, \tag{2.8b}$$
$$F_t = A_0r + A_3, \tag{2.8c}$$

and thus we have

$$\begin{aligned}\Delta &= (-2A_0 s + A_2)^2 - 4(A_0 t + A_1)(A_0 r + A_3) \\ &= 4A_0^2 s^2 + A_2^2 - 4A_0 A_2 s - 4(A_0^2 rt + A_0 A_1 r + A_0 A_3 t + A_1 A_3) \\ &= A_2^2 - 4A_1 A_3 + 4A_0 A_4,\end{aligned} \tag{2.9}$$

where we used (2.1) in the last equality. This classification is equivalent to the classification based on the method of characteristics to be presented in Section 3.1.4. The main theoretical results for the hyperbolic Monge-Ampère equation are found in the study of geometric structures on smooth manifolds. For the interested reader, a classification of the Monge-Ampère equation based on geometrical arguments is discussed in [49].

2.2 Null space of the Monge-Ampère operator

Unique and existence of solutions to the hyperbolic Monge-Ampère equation remains thus far an open problem. In this section we provide some background to this by introducing a range of functions in the null space of the Monge-Ampère operator:

$$L[u] = \det(H(u)) = u_{xx}u_{yy} - u_{xy}^2, \tag{2.10}$$

where $H(u)$ denotes the Hessian matrix of u. Note that the null space of L contains solutions to the Monge-Ampère equation for the case $A_0 = 1$ and $A_i = 0$ for $i = 1, \ldots, 4$; being the parabolic case. For brevity we use the notation $\partial^{\boldsymbol{\alpha}} = \partial^{(\alpha_1,\alpha_2)} = \frac{\partial^{|\boldsymbol{\alpha}|}}{\partial x^{\alpha_1}\partial y^{\alpha_2}}$ with $|\boldsymbol{\alpha}| = \|\boldsymbol{\alpha}\|_1 = \alpha_1 + \alpha_2$.

We start with a trivial case. Let $\mathcal{N}$ denote the null space. If a first derivative of a function g vanishes, i.e., if $g \in \mathcal{N}(\partial^\alpha)$ for a certain $|\boldsymbol{\alpha}| = 1$, then they also vanish under the Monge-Ampère operator, i.e., $g \in \mathcal{N}(L)$. By linearity of the derivative operators we can apply superposition and find that $\{ax + by + c \,|\, a, b, c \in \mathbb{R}\} \subset \mathcal{N}(L)$.

For a non-trivial case let $\phi \in C^1(\mathbb{R})$; if $g(x, y) = \phi(ax + by)$ for $a, b \in \mathbb{R}$ then $g \in \mathcal{N}(L)$, viz.

$$L[g] = \begin{vmatrix} a^2\phi''(ax+by) & ab\phi''(ax+by) \\ ab\phi''(ax+by) & b^2\phi''(ax+by) \end{vmatrix} = (a^2b^2 - (ab)^2)\phi''(ax+by) = 0. \tag{2.11}$$

Consequently we have a similar result for convoluted functions. To see this we first introduce the convolution and some basic properties. Let f and g

be functions such that their product is continuous and integrable. Then the convolution $(f * g)$ is given by:

$$(f * g)(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\tau_x, \tau_y) g(x - \tau_x, y - \tau_y) \, \mathrm{d}\tau_x \, \mathrm{d}\tau_y. \tag{2.12}$$

Let $\lambda \in \mathbb{R}$, then for convolutions the following basic properties hold:

$$(f * \lambda g)(x, y) = (\lambda f * g)(x, y) = \lambda (f * g)(x, y), \tag{2.13a}$$

$$(f * g)(x, y) = (g * f)(x, y), \tag{2.13b}$$

$$\frac{\partial}{\partial x_i}(f * g) = \frac{\partial f}{\partial x_i} * g = f * \frac{\partial g}{\partial x_i}. \tag{2.13c}$$

Consequently, if $g \in \mathcal{N}(\partial^{\boldsymbol{\alpha}})$ for any $\boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}| = 1$, then

$$L[f * g] = \begin{vmatrix} \frac{\partial^2}{\partial x^2}(f * g) & \frac{\partial^2}{\partial x \partial y}(f * g) \\ \frac{\partial^2}{\partial x \partial y}(f * g) & \frac{\partial^2}{\partial y^2}(f * g) \end{vmatrix} = \begin{vmatrix} f * \frac{\partial^2 g}{\partial x^2} & f * \frac{\partial^2 g}{\partial x \partial y} \\ f * \frac{\partial^2 g}{\partial x \partial y} & f * \frac{\partial^2 g}{\partial y^2} \end{vmatrix}. \tag{2.14}$$

Because either $\frac{\partial g}{\partial x} = 0$ or $\frac{\partial g}{\partial y} = 0$, we find $\frac{\partial^2 g}{\partial x \partial y} = 0$ and either $\frac{\partial^2 g}{\partial x^2} = 0$ or $\frac{\partial^2 g}{\partial y^2} = 0$, so $\frac{\partial^2 g}{\partial x^2} \frac{\partial^2 g}{\partial y^2} = 0$ and $L[f * g] = 0$ follows.

Analogously, if we let $\phi \in C^1(\mathbb{R})$, $a, b \in \mathbb{R}$, $g(x, y) = \phi(ax + by)$ and $\tilde{g}(x, y) = \phi''(ax + by)$, we have:

$$\begin{aligned} L[f * g] &= \begin{vmatrix} \frac{\partial^2}{\partial x^2}(f * g) & \frac{\partial^2}{\partial x \partial y}(f * g) \\ \frac{\partial^2}{\partial x \partial y}(f * g) & \frac{\partial^2}{\partial y^2}(f * g) \end{vmatrix} \\ &= \begin{vmatrix} f * \frac{\partial^2 g}{\partial x^2} & f * \frac{\partial^2 g}{\partial x \partial y} \\ f * \frac{\partial^2 g}{\partial x \partial y} & f * \frac{\partial^2 g}{\partial y^2} \end{vmatrix} \\ &= \begin{vmatrix} f * a^2 \tilde{g} & f * ab \tilde{g} \\ f * ab \tilde{g} & f * b^2 \tilde{g} \end{vmatrix} \\ &= (a^2 b^2 - (ab)^2)(f * \tilde{g})^2 \\ &= 0. \end{aligned} \tag{2.15}$$

There is a larger set of functions for which the Hessian is zero. To show this, let $r \in \mathbb{R}$, $r > 0$ and $a, b \in \mathbb{R}$. Then

$$g(x, y; r, a, b) := (a x^r + b y^r)^{\frac{1}{r}} \quad \Rightarrow \quad L[g] = 0. \tag{2.16}$$

To see this we calculate the derivatives of g, viz.

$$g_x(x,y;r,a,b) = ax^{r-1}\left(ax^r + by^r\right)^{\frac{1}{r}-1}, \tag{2.17a}$$

$$g_y(x,y;r,a,b) = by^{r-1}\left(ax^r + by^r\right)^{\frac{1}{r}-1}, \tag{2.17b}$$

$$\begin{aligned} g_{xx}(x,y;r,a,b) &= a^2(1-r)x^{2r-2}\left(ax^r + by^r\right)^{\frac{1}{r}-2} \\ &\quad + a(r-1)x^{r-2}\left(ax^r + by^r\right)^{\frac{1}{r}-1} \\ &= ab(r-1)x^{r-2}y^r\left(ax^r + by^r\right)^{\frac{1}{r}-2}, \end{aligned} \tag{2.17c}$$

$$\begin{aligned} g_{yy}(x,y;r,a,b) &= b^2(1-r)y^{2r-2}\left(ax^r + by^r\right)^{\frac{1}{r}-2} \\ &\quad + b(r-1)y^{r-2}\left(ax^r + by^r\right)^{\frac{1}{r}-1} \\ &= ab(r-1)x^r y^{r-2}\left(ax^r + by^r\right)^{\frac{1}{r}-2}, \end{aligned} \tag{2.17d}$$

$$g_{xy}(x,y;r,a,b) = ab(1-r)x^{r-1}y^{r-1}\left(ax^r + by^r\right)^{\frac{1}{r}-2}. \tag{2.17e}$$

Collecting powers of x and y then yields $g_{xx}g_{yy} - g_{xy}^2 = 0$.

This result can easily be generalized to sums of $g(x,y;r,a,b)$. To show this, let $\mathbf{r},\mathbf{a},\mathbf{b} \in \mathbb{R}^N$ with $N \in \mathbb{N}$, $N \geq 1$, such that for fixed $n = 1,\ldots,N$ we have $r_n > 0$. For $k = 2,\ldots,N$ define $G^k(x,y;r,a,b) = \sum_{n=1}^{k} g(x,y,r_n,a_n,b_n)$. It then holds that

$$L[G^N] = 0. \tag{2.18}$$

We prove the above statement by induction on N. To ease the reading, define $g_n = g(x,y;r_n,a_n,b_n)$ and the derivative of g_n w.r.t. x by $g_{n,x}$, and likewise for the other derivatives. For the case $N = 2$ we have

$$G^2(x,y;r,a,b) = g_1 + g_2 = g(x,y,r_1,a_1,b_1) + g(x,y,r_2,a_2,b_2), \tag{2.19}$$

and by applying the Monge-Ampère operator we find

$$\begin{aligned} L[G^2] &= (g_{n_1,xx} + g_{n_2,xx})(g_{n_1,yy} + g_{n_2,yy}) - (g_{n_1,xy} + g_{n_2,xy})^2 \\ &= g_{n_1,xx}g_{n_1,yy} - g_{n_1,xy}^2 + g_{n_2,xx}g_{n_2,yy} - g_{n_2,xy}^2 + \\ &\qquad g_{n_1,xx}g_{n_2,yy} - g_{n_1,xy}g_{n_2,xy} + g_{n_2,xx}g_{n_1,yy} - g_{n_1,xy}g_{n_2,xy} \\ &= L[g_1] + L[g_2] + g_{n_1,xx}g_{n_2,yy} - g_{n_1,xy}g_{n_2,xy} + g_{n_2,xx}g_{n_1,yy} - g_{n_1,xy}g_{n_2,xy}. \end{aligned} \tag{2.20}$$

By (2.16) we have $L[g_1] = L[g_2] = 0$. Evidently, if $g_{n_1,xx}g_{n_2,yy} - g_{n_1,xy}g_{n_2,xy} = 0$, then also $g_{n_2,xx}g_{n_1,yy} - g_{n_1,xy}g_{n_2,xy} = 0$, implying $L[G^2] = 0$. Explicitely writing out $g_{n_1,xx}g_{n_2,yy}$ and $g_{n_1,xy}g_{n_2,xy}$ using equations (2.17) yields the result, viz.

$$\begin{aligned} g_{n_1,xx}g_{n_2,yy} &= \left(a_1b_1(r_1-1)x^{r_1-2}y_1^r\left(a_1x_1^r + b_1y_1^r\right)^{\frac{1}{r_1}-2}\right)\cdot \\ &\qquad \left(a_2b_2(r_2-1)x^{r_2}y^{r_2-2}\left(a_2x_2^r + b_2y_2^r\right)^{\frac{1}{r_2}-2}\right) \\ &= a_1b_1a_2b_2(r_1-1)(r_2-1)x^{r_1+r_2-2}y^{r_1+r_2-2} \\ &\qquad \cdot \left(a_1x_1^r + b_1y_1^r\right)^{\frac{1}{r}-2}\left(a_2x_2^r + b_2y_2^r\right)^{\frac{1}{r_2}-2} \\ &= \left(a_1b_1(r_1-1)x^{r_1-1}y^{r_1-1}\left(a_1x_1^r + b_1y_1^r\right)^{\frac{1}{r}-2}\right)\cdot \\ &\qquad \left(a_2b_2(r_2-1)x^{r_2-1}y^{r_2-1}\left(a_2x_2^r + b_2y_2^r\right)^{\frac{1}{r_2}-2}\right) \\ &= g_{n_1,xy}g_{n_2,xy}. \end{aligned} \tag{2.21}$$

Note that this result implies that for fixed $n_1, n_2 \in \mathbb{N}$ with $n_1, n_2 \geq 1$ we have

$$g_{n_1,xx}g_{n_2,yy} - g_{n_1,xy}g_{n_2,xy} = g_{n_2,xx}g_{n_1,yy} - g_{n_1,xy}g_{n_2,xy} = 0, \tag{2.22}$$

with the second equality due to re-indexing. Assume $L[G^{k-1}] = 0$ holds for $k = 3, \ldots, N$, we show $L[G^k] = 0$ which finalizes the proof. We have

$$\begin{aligned} L[G^k] &= L[G^{k-1} + g_k] \\ &= \left(G_{xx}^{k-1} + g_{k,xx}\right)\left(G_{yy}^{k-1} + g_{k,yy}\right) - \left(G_{xy}^{k-1} + g_{k,xy}\right)^2 \\ &= G_{xx}^{k-1}G_{yy}^{k-1} - (G_{xy}^{k-1})^2 + g_{k,xx}g_{k,yy} - (g_{k,xy})^2 \\ &\qquad + \left(G_{xx}^{k-1}g_{k,yy} - G_{xy}^{k-1}g_{k,xy}\right) + \left(g_{k,xx}G_{yy}^{k-1} - G_{xy}^{k-1}g_{k,xy}\right) \\ &= L[G^{k-1}] + L[g_k] + \sum_{n=1}^{k-1}\left(g_{n,xx}g_{k,yy} - g_{n,xy}g_{k,xy}\right) \\ &\qquad + \sum_{n=1}^{k-1}\left(g_{k,xx}g_{n,yy} - g_{n,xy}g_{k,xy}\right) \\ &= 0. \end{aligned} \tag{2.23}$$

By the induction hypothesis we have $L[G^{k-1}] = 0$, by (2.16) we have $L[g_k] = 0$ and by (2.22) the two summations equal zero due to each of the terms equaling zero.

As a last example, let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^N$, $\phi_n \in C^2(\mathbb{R})$ for all $n = 1, \ldots, N$. Then functions of the form $\phi(x, y) = \sum_{n=1}^{N} \phi_n(a_n x + b_n y)$ satisfy $L[\phi] = 0$ if

$$a_{n_1} b_{n_2} - a_{n_2} b_{n_1} = 0, \quad \text{or} \quad \phi''_{n_1}(a_{n_1} x + b_{n_1} y)\phi''_{n_2}(a_{n_2} x + b_{n_2} y) = 0, \tag{2.24}$$

for all $n_1, n_2 = 1, \ldots, N$ with $n_1 \neq n_2$. More precisely, we have

$$L[\phi](x, y) = \sum_{n_1 \neq n_2} (a_{n_1} b_{n_2} - a_{n_2} b_{n_1})^2 \phi''_{n_1}(a_{n_1} x + b_{n_1} y)\phi''_{n_2}(a_{n_2} x + b_{n_2} y) \tag{2.25}$$

implying $L[\phi] = 0$ if (2.24) holds. The proof is analogous to that of (2.18) and will therefore be omitted.

2.3 Generating problem-solution pairs

Finding solutions to the Monge-Ampère equation can be problematic due to the nonlinear nature. Therefore we introduce a method* based on complex functions to quickly obtain a problem-solution pair (u, f) for the standard hyperbolic Monge-Ampère equation $u_{xx} u_{yy} - u_{xy}^2 = -f^2 < 0$ with $f \neq 0$. To this end let w be a complex analytical function and let $u(x, y) = \mathrm{Re}(w(x + iy))$, with i the imaginary unit. Differentiation then yields

$$\begin{aligned} u_{xx} &= \mathrm{Re}(w''), \\ u_{xy} &= \mathrm{Re}(iw'') = -\mathrm{Im}(w''), \\ u_{yy} &= -\mathrm{Re}(w''). \end{aligned} \tag{2.26}$$

It then follows that

$$f^2 = -u_{xx} u_{yy} + u_{xy}^2 = \Big(\mathrm{Re}(w'')\Big)^2 + \Big(\mathrm{Im}(w'')\Big)^2 = |w''|^2. \tag{2.27}$$

The implications of this are twofold. First, given w we can construct $u(x, y) = \mathrm{Re}(w(x + iy))$ and $f(x, y) = |w''(x + iy)|$, which form a solution. Secondly, if for given f there exists an analytical function w such that $f^2(x, y) = |w''|^2$, then $u(x, y) = \mathrm{Re}(w(x + iy))$ solves the Monge-Ampère equation.

Alternatively, $u(x, y) = -\mathrm{Re}(w(x + iy))$ and $u(x, y) = \pm\mathrm{Im}(w(x + iy))$ yield the same conclusions as above. We can generalize these results as follows: let w be a complex analytical function, $c_1, c_2 \in \mathbb{C}$, $\theta \in \mathbb{R}$ and define

$$u(x, y) := \mathrm{Re}(c_1 w(e^{i\theta}(x + iy) + c_2)), \tag{2.28a}$$

$$f^2(x, y) := |c_1|^2 |w''(e^{i\theta}(x + iy) + c_2)|^2, \tag{2.28b}$$

*Special thanks goes to J. de Graaf for originally contributing this method, as it has greatly simplified generating test examples for the Monge-Ampère equation.

then $u_{xx}u_{yy} - u_{xy}^2 = -f^2$. Proving this is straightforward and the proof is therefore omitted. Comparing (2.27) and (2.28b) shows how scaling, translating and rotating $w(x+iy)$ in the complex plane scales, translates and rotates f^2. The corresponding scaling, translation and rotation of the solution u is given by (2.28a).

The above methods can be extended to the general hyperbolic Monge-Ampère equation (2.1). We show the case for $u(x,y) = \mathrm{Re}(w(x+iy))$. In order to do so, define $R = \mathrm{Re}(w'')$ and $I = \mathrm{Im}(w'')$. The general Monge-Ampère equation with $A_0 = 1$ in terms of R and I is then given by

$$-R^2 - I^2 + (A_1 - A_3)R - A_2 I + A_4 = 0. \tag{2.29}$$

For the general Monge-Ampère equation we now have the additional freedom of choosing the coefficients A_i. To do so, we need to incorporate the hyperbolicity condition

$$\tfrac{1}{4}A_2^2 - A_3 A_1 + A_4 > 0. \tag{2.30}$$

First, we rewrite (2.29) as

$$A_4 + \tfrac{1}{4}A_2^2 + \tfrac{1}{4}(A_1 - A_3)^2 = (R - \tfrac{1}{2}(A_1 - A_3))^2 + (I + \tfrac{1}{2}A_2)^2. \tag{2.31}$$

Because the right-hand side is a sum of squares, it follows that the left-hand side of (2.31) should be nonnegative, hence the choice of the A_i's should satisfy both (2.30) and

$$A_4 + \tfrac{1}{4}A_2^2 + \tfrac{1}{4}(A_1 - A_3)^2 \geq 0. \tag{2.32}$$

We start by choosing $A_1 = R + I$, reducing inequality (2.32) to

$$A_4 + \tfrac{1}{4}A_2^2 + \tfrac{1}{4}(R + I - A_3)^2 \geq 0, \tag{2.33}$$

which we slightly simplify by choosing $A_3 = I$. Using $A_1 = R + I$ and $A_3 = I$, the hyperbolicity condition (2.30) reads

$$A_4 + \tfrac{1}{4}A_2^2 - I(R + I) > 0. \tag{2.34}$$

Choosing $A_4 = I(R + I)$ yields the hyperbolicity condition $A_2 \neq 0$. Solving equation (2.29) for A_2 finally yields $A_2 = R$. Inequality (2.33) then reads

$$\tfrac{1}{2}(R + I)^2 + \tfrac{1}{2}I^2 \geq 0, \tag{2.35}$$

which is automatically satisfied. By the hyperbolicity condition and the choice $A_2 = R$ we have $R \neq 0$. Hence we require $R \neq 0$ when choosing ω. So in total we have

$$\begin{array}{llll} u = \mathrm{Re}(w)), & R = \mathrm{Re}(w''), & I = \mathrm{Im}(w''), & A_0 = 1, \\ A_1 = R + I, & A_2 = R, & A_3 = I, & A_4 = I(R + I). \end{array} \tag{2.36}$$

What remains is to choose w and Ω such that $R \neq 0$. We choose $w(z) = \cos(z + i)$ and a straightforward calculation using (2.36) shows

$$u = \cos(x)\cosh(1 + y), \tag{2.37a}$$
$$R = -\cos(x)\cosh(1 + y), \tag{2.37b}$$
$$I = \sin(x)\sinh(1 + y). \tag{2.37c}$$

Exact conditions on x and y can be derived such that $R \neq 0$, here we simply choose $\Omega = [\frac{1}{5}, \frac{3}{4}] \times [-\frac{1}{2}, \frac{1}{3}]$, which is one of the possible domains satisfying this constraint.

A range of Monge-Ampère equations with solutions and appropriate domains is listed in Table 2.1 for the standard Monge-Ampère equation and in Tables 2.2 and 2.3 for the general case.

$u(x, y)$	$f(x, y)$	Ω
$\frac{1}{2}x^2 - \frac{1}{2}y^2$	1	$\mathbb{R}^2$
$\frac{1}{12}x^4 + \frac{1}{3}x^3 + \frac{1}{2}x^2 - \frac{1}{2}y^2$	$x + 1$	$[0, 1] \times [-\frac{1}{2}, \frac{1}{2}]$
$\frac{1}{12}x^4 + \frac{1}{3}x^3 + xy - \frac{1}{2}y^2$	$x + 1$	$[0, 1] \times [-\frac{1}{2}, \frac{1}{2}]$
$\frac{1}{12}\left((x-1)^4 - 6(x-1)^2y^2 + y^4\right)$	$(x-1)^2 + y^2$	$[-2, -\frac{1}{2}] \times [-1, 1]$
$\frac{1}{30}(x^6 - y^6) + \frac{1}{2}(x^2y^4 - x^4y^2) - x$	$(x^2 + y^2)^2$	$[-\frac{1}{4}, \frac{1}{4}] \times [\frac{3}{4}, \frac{3}{2}]$
$x^3y^2 + 1$	$2\sqrt{6}x^2y$	$[\frac{1}{2}, \frac{3}{2}] \times [-1, 1]$
$\frac{1}{4\sqrt{3}}(x^2 - y^2)^2$	$x^2 - y^2$	$[-\frac{1}{2}, \frac{1}{2}] \times [-1, -\frac{1}{3}]$
$\cos(x)\cosh(y)$	$\frac{1}{2}\sqrt{\cos(2x) + \cosh(2y)}$	$[-\frac{1}{2}, \frac{1}{2}] \times [-2, 2]$
$e^x\cos(y)$	e^x	$[-1, 1]^2$
$e^x\cos(x - y)$	e^x	$[-1, 1]^2$
$\exp(2\frac{y}{x})$	$\frac{2}{x^2}\exp(2\frac{y}{x})$	$[1, \frac{5}{2}] \times [-2, -\frac{3}{2}]$

Table 2.1: Problem-solution pairs for the standard Monge-Ampère equation and (a subset of) the domain on which it is hyperbolic.

$$\begin{aligned}
\Omega &= [-1,1] \times [1,\tfrac{3}{2}] \\
u(x,y) &= \cos(x)y^2 \\
A_1(x,y,u,p,q) &= 0 \\
A_2(x,y,u,p,q) &= 0 \\
A_3(x,y,u,p,q) &= 0 \\
A_4(x,y,u,p,q) &= \tfrac{1}{4}q^2 + p\sin(x) - 3y^2
\end{aligned} \tag{2.38}$$

$$\begin{aligned}
\Omega &= [-\tfrac{1}{4},\tfrac{1}{4}] \times [\tfrac{1}{2},1] \\
u(x,y) &= e^x\cos(x) \\
A_1(x,y,u,p,q) &= e^x\sin(y) + q + u - p \\
A_2(x,y,u,p,q) &= qe^x\sin(y) - up + e^{2x} \\
A_3(x,y,u,p,q) &= pe^x\sin(y) + qu \\
A_4(x,y,u,p,q) &= p^2 + q^2
\end{aligned} \tag{2.39}$$

$$\begin{aligned}
\Omega &= [-\tfrac{1}{4},\tfrac{1}{4}] \times [\tfrac{1}{2},1] \\
u(x,y) &= e^x\cos(x) \\
A_1(x,y,u,p,q) &= 1 + \frac{up}{\exp(x)\cos(y)} \\
A_2(x,y,u,p,q) &= -\frac{1+u+p}{q} - \exp x\sin(y) \\
A_3(x,y,u,p,q) &= 1 - q\frac{\cos(y)}{\sin(y)} \\
A_4(x,y,u,p,q) &= (1+\exp(x)\cos(y))^2
\end{aligned} \tag{2.40}$$

Table 2.2: Hyperbolic Monge-Ampère equation problems with solution and appropriate domain.

$$
\begin{aligned}
\Omega &= [\tfrac{1}{5}, \tfrac{3}{4}] \times [-\tfrac{1}{2}, \tfrac{1}{3}] \\
u(x,y) &= \cos(x)\cosh(1+y) \\
A_1(x,y,u,p,q) &= -u - \frac{pq}{\cos(x)\cosh(1+y)} \\
A_2(x,y,u,p,q) &= -\frac{q}{\tanh(1+y)} \\
A_3(x,y,u,p,q) &= -\cos(x)\cosh(1+y) \cdot \frac{\cosh(1+y)^2 - u^2 - sin(x)^2}{pq} \\
A_4(x,y,u,p,q) &= pq + p^2 - \sin(x)^2
\end{aligned}
\tag{2.41}
$$

$$
\begin{aligned}
\Omega &= [1, \tfrac{3}{2}] \times [2, \tfrac{5}{2}] \\
u(x,y) &= \frac{1}{2}\log(x^2+y^2) \\
A_1(x,y,u,p,q) &= (p+q)(x-y)e^{-2u} \\
A_2(x,y,u,p,q) &= 3\Big(2pq - \frac{1}{2xy}\Big) \\
A_3(x,y,u,p,q) &= q^2 - p^2 \\
A_4(x,y,u,p,q) &= 4xypqe^{-4u}
\end{aligned}
\tag{2.42}
$$

Table 2.3: Hyperbolic Monge-Ampère equation problems with solution and appropriate domain.

2.4 Integral formulation

Next we introduce an integral formulation of the standard Monge-Ampère equation. This allows us to transform a problem-solution pair (u, f) into a new pair $(\tilde{u}, \tilde{f})$, for which we give an example at the end of this section. Furthermore, the integral formulation allows for the construction of a residual method presented in Section 4.1.5.

The standard Monge-Ampère equation can be written as $p_x q_y - p_y q_x + f^2 = 0$, which only depends on f^2 and the derivatives of p and q. We can rewrite it in terms of ∇p and ∇q. To this end let $\mathbf{J}$ be the symplectic matrix $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$, from which the two equivalent formulations

$$-f^2 = \nabla \cdot (p\mathbf{J}\nabla q), \tag{2.43a}$$

$$f^2 = \nabla \cdot (q\mathbf{J}\nabla p), \tag{2.43b}$$

follow, where we used $\nabla \cdot (\mathbf{J}\nabla\phi) = 0$ for a scalar function ϕ, and $\nabla \cdot (\phi\mathbf{v}) = \phi\nabla \cdot \mathbf{v} + \mathbf{v} \cdot \nabla\phi$ with $\mathbf{v}$ a vector-valued function. Let $A \subseteq \mathbb{R}^2$ be an orientable domain and let $\hat{\mathbf{n}}$ be the outward unit normal on ∂A. By subsequently integrating the right-hand side of (2.43a) over A and applying Gauss's theorem we obtain

$$\iint_A \nabla \cdot (p\mathbf{J}\nabla q)\,\mathrm{d}A = \oint_{\partial A} p\mathbf{J}\nabla q \cdot \hat{\mathbf{n}}\,\mathrm{d}s = \oint_{\partial A} p\nabla q \cdot \hat{\boldsymbol{\tau}}\,\mathrm{d}s, \tag{2.44}$$

where we defined $\hat{\boldsymbol{\tau}} = \mathbf{J}^{\mathrm{T}}\hat{\mathbf{n}}$. Note that $\hat{\boldsymbol{\tau}}$ is the unit tangent vector to the domain taken in the counter clockwise direction. It follows from (2.43) that

$$\iint_A f^2\,\mathrm{d}A = -\oint_{\partial A} p\nabla q \cdot \hat{\boldsymbol{\tau}}\,\mathrm{d}s, \tag{2.45a}$$

and in a similar way we find

$$\iint_A f^2\,\mathrm{d}A = \oint_{\partial A} q\nabla p \cdot \hat{\boldsymbol{\tau}}\,\mathrm{d}s. \tag{2.45b}$$

Subtracting these equations yields

$$\oint_{\partial A} (p\nabla q + q\nabla p) \cdot \hat{\boldsymbol{\tau}}\,\mathrm{d}s = 0, \tag{2.46}$$

which trivially holds by Stokes' theorem since $p\nabla q + q\nabla p = \nabla(pq)$ and the gradient of a function is conservative.

For the general Monge-Ampère equation (2.1) with $A_0 = 1$, $A_i \neq 0$ $(i = 1, \ldots, 4)$ we can do something similar, but the area integral will depend on the solution itself, viz.

$$\iint_A \left(A_1 r + A_2 s + A_3 t + A_4\right) \mathrm{d}A = -\oint_{\partial A} p \nabla q \cdot \hat{\boldsymbol{\tau}} \, \mathrm{d}s, \tag{2.47}$$

or written in terms of p, q and the outward normal $\hat{\mathbf{n}} = \mathbf{J}\hat{\boldsymbol{\tau}}$:

$$\begin{aligned} \iint_A \left(p \nabla \cdot \begin{pmatrix} A_1 \\ \frac{1}{2}A_2 \end{pmatrix} + q \nabla \cdot \begin{pmatrix} \frac{1}{2}A_2 \\ A_3 \end{pmatrix} + A_4\right) \mathrm{d}A = \\ -\oint_{\partial A} \left(p \mathbf{J} \nabla q + \begin{pmatrix} A_1 \\ \frac{1}{2}A_2 \end{pmatrix} p + \begin{pmatrix} \frac{1}{2}A_2 \\ A_3 \end{pmatrix} q\right) \cdot \hat{\mathbf{n}} \, \mathrm{d}s. \end{aligned} \tag{2.48}$$

We conclude this section by generating a problem-solution pair $(\tilde{u}, \tilde{f})$ for the standard Monge-Ampère equation by means of an integral formulation. We do so by generalizing equations (2.45) further by introducing functions acting on p and q. Let $\Gamma, \Psi \in C^1(\mathbb{R}^2)$ with $\Gamma = \Gamma(p, q)$ and $\Psi = \Psi(p, q)$. By definition we have for the standard Monge-Ampère equation $\nabla p \cdot \mathbf{J} \nabla q = -f^2$. Multiplying both sides by $\Gamma_p \Psi_q - \Gamma_q \Psi_p$ and applying standard differential rules and $\mathbf{v} \cdot \mathbf{J}\mathbf{v} = 0$ we find

$$\begin{aligned} -\Big(\Gamma_p \Psi_q - \Gamma_q \Psi_p\Big) f^2 &= \Big(\Gamma_p \Psi_q - \Gamma_q \Psi_p\Big) \nabla p \cdot \mathbf{J} \nabla q \\ &= \Gamma_p \nabla p \cdot \mathbf{J} \Psi_q \nabla q + \Gamma_q \nabla q \cdot \mathbf{J} \Psi_p \nabla p \\ &= (\Gamma_p \nabla p + \Gamma_q \nabla q) \cdot \mathbf{J} (\Psi_p \nabla p + \Psi_q \nabla q) \\ &= \nabla \Gamma(p, q) \cdot \mathbf{J} \nabla \Psi(p, q). \end{aligned} \tag{2.49}$$

As before, integrating (2.49) over an orientable domain A, applying integration by parts and Gauss's theorem, we obtain

$$\iint_A (\Gamma_p \Psi_q - \Gamma_q \Psi_p) f^2 \, \mathrm{d}A = -\oint_{\partial A} \Gamma(p, q) \nabla \Psi(p, q) \cdot \hat{\boldsymbol{\tau}} \, \mathrm{d}s. \tag{2.50}$$

This equation is a more general version of (2.45). Indeed, by choosing $\Gamma(p, q) = p, \Psi(p, q) = q$ we obtain (2.45a) and by choosing $\Gamma(p, q) = q, \Psi(p, q) = p$ we obtain (2.45b).

Furthermore, solutions to (2.50) are equivalent to the standard Monge-Ampère equation if the integrability condition

$$\int \Gamma(p(x, y), q(x, y)) \, \mathrm{d}x = \int \Psi(p(x, y), q(x, y)) \, \mathrm{d}y, \tag{2.51}$$

is satisfied. To see this, let $\tilde{f}^2 = (\Gamma_p \Psi_q - \Gamma_q \Psi_p) f^2$, $\tilde{p} = \Gamma(p,q)$ and $\tilde{q} = \Psi(p,q)$. Then, if

$$\int \tilde{p}(x,y)\,\mathrm{d}x = \int \tilde{q}(x,y)\,\mathrm{d}y := \tilde{u}, \tag{2.52}$$

we find that $\tilde{u}$ solves the standard Monge-Ampère equation $\det(H(\tilde{u})) = -\tilde{f}^2$, which follows from

$$\begin{aligned}
\det(H(\tilde{u})) &= \begin{vmatrix} \frac{\partial}{\partial x}\tilde{p}(x,y) & \frac{\partial}{\partial y}\tilde{p}(x,y) \\ \frac{\partial}{\partial x}\tilde{q}(x,y) & \frac{\partial}{\partial y}\tilde{q}(x,y) \end{vmatrix} \\
&= \begin{vmatrix} \frac{\partial}{\partial x}\Gamma(p,q) & \frac{\partial}{\partial y}\Gamma(p,q) \\ \frac{\partial}{\partial x}\Psi(p,q) & \frac{\partial}{\partial y}\Psi(p,q) \end{vmatrix} \\
&= \begin{vmatrix} \Gamma_p p_x + \Gamma_q q_x & \Gamma_p p_y + \Gamma_q q_y \\ \Psi_p p_x + \Psi_q q_x & \Psi_p p_y + \Psi_q q_y \end{vmatrix} \\
&= \Gamma_p \Psi_p p_x p_y + \Gamma_q \Psi_q q_x q_y + \Gamma_p \Psi_q p_x q_y + \Gamma_q \Psi_p p_y q_x \\
&\quad - \Gamma_p \Psi_p p_x p_y - \Gamma_p \Psi_q p_x q_y - \Gamma_q \Psi_p p_y q_x - \Gamma_q \Psi_q q_x q_y \\
&= \Gamma_p \Psi_q (p_x q_y - q_x p y) + \Gamma_q \Psi_p (p_y q_x - p_x q_y) \\
&= (\Gamma_p \Psi_q - \Gamma_q \Psi_p)(p_x q_y - q_x p y) \\
&= -(\Gamma_p \Psi_q - \Gamma_q \Psi_p) f^2,
\end{aligned} \tag{2.53}$$

where we used $p_x q_y - p_y q_x = -f^2$ and selectively omitted the arguments (x,y) and (p,q) for readability.

In order to construct a problem solution pair $(\tilde{u}, \tilde{f})$, we choose $u = x^3 y^2 + 1$ and $f^2 = 24 x^4 y^2$, as given by Table 2.1. Furthermore, we choose second order polynomials in p and q for Γ and Ψ according to:

$$\Gamma(p,q) = \gamma_1 p + \gamma_2 q + \gamma_3 p^2 + \gamma_4 q^2 + \gamma_5 pq, \tag{2.54a}$$

$$\Psi(p,q) = \psi_1 p + \psi_2 q + \psi_3 p^2 + \psi_4 q^2 + \psi_5 pq, \tag{2.54b}$$

with $\gamma_i, \psi_i \in \mathbb{R}$ for $i = 1, \ldots, 5$. A straightforward calculation shows

$$\int \tilde{p}(x,y)\,\mathrm{d}x = \gamma_1 x^3 y^2 + \frac{1}{2}\gamma_2 x^4 y + \frac{9}{5}\gamma_3 x^5 y^4 + \gamma_5 x^6 y^3 + \frac{4}{7}\gamma_4 x^7 y^2 + C_1(y), \tag{2.55a}$$

$$\int \tilde{q}(x,y)\,\mathrm{d}y = \psi_1 x^2 y^3 + \psi_2 x^3 y^2 + \frac{9}{5}\psi_3 x^4 y^5 + \frac{3}{2}\psi_5 x^5 y^4 + \frac{4}{3}\psi_4 x^6 y^3 + C_2(x). \tag{2.55b}$$

Matching terms of (2.55a) and (2.55b) by their polynomial order, we find $\gamma_2 = 0$, $\gamma_4 = 0$, $\psi_1 = 0$, $\psi_2 = \gamma_1$, $\psi_3 = 0$, $\psi_4 = \frac{3}{4}\gamma_5$, $\psi_5 = \frac{6}{5}\gamma_3$, $C_1(y) = C_2(x) = C$ with $C \in \mathbb{R}$. Consequently, the pair

$$\tilde{u}(x,y) = \gamma_1 x^3 y^2 + \gamma_5 x^6 y^3 + \frac{9}{5}\gamma_3 x^5 y^4 + C, \tag{2.56a}$$

$$\begin{aligned}\tilde{f}^2(x,y) = 24\gamma_1^2 x^4 y^2 + 120\gamma_1\gamma_5 x^7 y^3 + \frac{1152}{5}\gamma_1\gamma_3 x^6 y^4 \\ + 144\gamma_5^2 x^{10} y^4 + 432\gamma_3\gamma_5 x^9 y^5 + \frac{2592}{5}\gamma_3^2 x^8 y^6\end{aligned} \tag{2.56b}$$

solves the Monge-Ampère equation $\det(H(\tilde{u})) = -\tilde{f}^2$ for all $\gamma_1, \gamma_3, \gamma_5, C \in \mathbb{R}$.

2.5 Symmetries

In this section we briefly discuss symmetries of the Monge-Ampère equation. As we will see, the symmetries do not only depend on the Monge-Ampère equation, but also on the domain and the range of ∇u.

In this thesis, we encounter two different boundary conditions for the Monge-Ampère equation, Cauchy and transport boundary conditions. The former will be properly introduced in Section 3.4 in relation to the method of characteristics. The latter we encounter in Chapters 5 and 6. Here we mainly focus on the transport boundary condition and the symmetries it allows for.

Let us consider the standard elliptic and hyperbolic Monge-Ampère equation. We introduce the mapping $\mathbf{m} = \nabla u$ with $\mathbf{m} \in C^1(\mathcal{X})$, $\mathbf{m} : \mathcal{X} \mapsto \mathcal{Y}$ and $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^2$. Let $\mathrm{D}\mathbf{m}$ denote the Jacobi matrix of $\mathbf{m}$, then the Monge-Ampère equation reads $\det(\mathrm{D}\mathbf{m}) + A_4 = 0$. The transport boundary condition is given by

$$\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}, \tag{2.57}$$

meaning each point from the boundary $\partial\mathcal{X}$ is mapped to $\partial\mathcal{Y}$ by the mapping $\mathbf{m}$. First, let us consider a simple case, viz. $\mathcal{X} = [-1,1]^2$ and $\mathcal{Y}$ a square domain. Let $\mathbf{x}(s)$ with $s \in [0,1)$ be a bijective parametrization of $\partial\mathcal{X}$. Clearly, there exist distinct s_i for $i = 1,\dots,4$ such that $\mathbf{x}(s_i)$ coincides with the corners of $\mathcal{X}$. Consequently, $\mathbf{x}$ is not differentiable at $s = s_i$ and $\mathbf{m}(s) := \mathbf{m}(\mathbf{x}(s))$ is not differentiable at $s = s_i$. By the transport boundary condition and an analogous bijective parametrization $\mathbf{y}(s)$ of $\partial\mathcal{Y}$, we have four non differential points on $\partial\mathcal{X}$ and four on $\partial\mathcal{Y}$. By identifying $\mathbf{y}(s) = \mathbf{m}(s)$ we have that the corners of $\mathcal{X}$ are mapped to corners of $\mathcal{Y}$.

Consider the case $f \equiv 1$ with $\mathcal{Y} = \mathcal{X} = [-1,1]^2$. By (2.9) and (2.5) the Monge-Ampère equation is elliptic in case $A_4 = -f^2$ and hyperbolic

when $A_4 = f^2$. Starting with the elliptic variant, we have that $\mathbf{m} = \nabla u$ and $\mathbf{m}(x, y) = (x, y)^{\mathrm{T}}$ is a solution to the standard Monge-Ampère equation. By rotational symmetry of f, rotational symmetry of both $\mathcal{X}$ and $\mathcal{Y}$ over $k\pi/2$ ($k \in \mathbb{Z}$) radians, it makes sense to consider mappings of the form

$$\mathbf{m}(x, y; \theta) = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}, \tag{2.58}$$

as solutions to the standard MA equation described above. It follows that $H(u) = \mathrm{D}\mathbf{m}$ and

$$\begin{aligned} \det\left(\mathrm{D}\left[\begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \mathbf{m}\right]\right) &= \det\left(\begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \mathrm{D}\mathbf{m}\right) \\ &= \det\begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \det(\mathrm{D}\mathbf{m}) \\ &= \det(\mathrm{D}\mathbf{m}), \end{aligned} \tag{2.59}$$

meaning rotations of $\mathbf{m}$ automatically satisfy the Monge-Ampère equation if $\mathbf{m}$ does. Because of $\mathbf{m} = \nabla u$, i.e., $m_1 = u_x$ and $m_2 = u_y$, we have

$$u(x, y; \theta) = \int m_1(x, y; \theta)\, \mathrm{d}x = \tfrac{1}{2}\cos(\theta)x^2 - \sin(\theta)xy + C_1(x), \tag{2.60a}$$

$$u(x, y; \theta) = \int m_2(x, y; \theta)\, \mathrm{d}y = \sin(\theta)xy + \tfrac{1}{2}\cos(\theta)y^2 + C_2(x), \tag{2.60b}$$

with scalar functions C_1 and C_2. Note that to uniquely determine u from $\mathbf{m}$, additional boundary conditions are required. We come back to this in Section 5.1.1. For now we just remark that if u solves the Monge-Ampère equation, then so do $c_1 u$ and $c_2 + u$ with $c_1, c_2 \in \mathbb{R}$. By matching the polynomial terms of (2.60a) and (2.60b), we obtain the solution u, given by

$$\begin{cases} u = \frac{1}{2}\cos(\theta)(x^2 + y^2) + c, \\ \sin(\theta) = -\sin(\theta), \end{cases} \tag{2.61}$$

with $c \in \mathbb{R}$. From the latter equation we have $\theta = k\pi$ with $k \in \mathbb{Z}$, thus we obtain the two solutions

$$u_1(x, y; \theta) = \tfrac{1}{2}(x^2 + y^2) + c, \tag{2.62a}$$

$$u_2(x, y; \theta) = -\tfrac{1}{2}(x^2 + y^2) + c, \tag{2.62b}$$

and, given in terms of the mapping

$$\mathbf{m}_1(x, y; \theta) = (x, y)^{\mathrm{T}}, \tag{2.63a}$$
$$\mathbf{m}_2(x, y; \theta) = (-x, -y)^{\mathrm{T}}. \tag{2.63b}$$

For $A_4 = f^2$ we have that $\mathbf{m} = (x, -y)^{\mathrm{T}}$ solves the hyperbolic Monge-Ampère equation. A similar approach as for the elliptic MA yields

$$u(x, y; \theta) = \tfrac{1}{2}\cos(\theta)(x^2 - y^2) + \sin(\theta)xy + c. \tag{2.64}$$

Because the corner of the source needs to be mapped to the corner of the target, we have $\theta = k\pi/2$ with $k \in \mathbb{Z}$. Substituting these values for θ in (2.64) yields the four distinct solutions

$$u_1(x, y; \theta) = \tfrac{1}{2}(x^2 - y^2) + c, \tag{2.65a}$$
$$u_2(x, y; \theta) = xy + c, \tag{2.65b}$$
$$u_3(x, y; \theta) = -\tfrac{1}{2}(x^2 - y^2) + c, \tag{2.65c}$$
$$u_4(x, y; \theta) = -xy + c. \tag{2.65d}$$

We thus obtain two elliptic and four hyperbolic solutions to the Monge-Ampère equation. Note that if we choose

$$w(z; \theta) = \tfrac{1}{2}\exp(\theta i)z^2, \tag{2.66}$$

with $\theta \in \mathbb{R}$, then in accordance with Section 2.3 we have

$$u(x, y) = \mathrm{Re}(w(x + iy; \theta)) = \tfrac{1}{2}\cos(\theta)(x^2 - y^2) - \sin(\theta)xy, \tag{2.67}$$

and $|w''|^2 = f^2 = 1$, where the prime denotes differentiation w.r.t. z, yielding the same result as in (2.65) for the hyperbolic problem. To slightly expand upon the previous example, consider $\mathcal{Y}$ to be the square $[-1, 1]^2$ rotated over $\phi \in (0, \pi) \cup (\pi, 2\pi)$ radians. Then,

$$u(x, y) = \tfrac{1}{2}\cos(k\tfrac{\pi}{2} + \phi)(x^2 - y^2) + \sin(k\tfrac{\pi}{2} + \phi)xy + c, \quad k = 0, 1, 2, 3, \tag{2.68}$$

solves the hyperbolic problem, but according to (2.61), no elliptic solution exists.

Without considering boundary conditions, (2.67) generates an uncountable set of solutions, parametrized by $\theta \in \mathbb{R}$. In case $\mathcal{X} = \mathcal{Y}$ are balls of fixed radius with their center at the origin, (2.67) also satisfies the transport boundary condition.

Let us now focus on the hyperbolic Monge-Ampère equation. Let u satisfy $u_{xx}u_{yy} - u_{xy}^2 = -f^2$, with mapping $\mathbf{m} = \nabla u : \mathcal{X} \subset \mathcal{Y}$ for $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^2$.

Furthermore, we assume $w \in C^2(\mathbb{R})$ and $f^2(x,y) = |w''(x+iy)|^2$ such that $u(x,y) = \mathrm{Re}(w(x+iy))$. We scale $\mathcal{Y}$ by a factor c_1, rotate it over ϕ radians, and translate it over $(c_2, c_3)^{\mathrm{T}}$ in $\mathbb{R}^2$. We denote the result by $\widetilde{\mathcal{Y}}$ for $c_1, c_2, c_3, \phi \in \mathbb{R}$. It follows that

$$\tilde{u}(x,y) = Re(\tilde{w}(x+iy)), \tag{2.69a}$$

$$\tilde{w}(z) = (c_2 - c_3 i)z + c_1 \exp(i\phi) w(z), \tag{2.69b}$$

is the solution to $\tilde{u}_{xx}\tilde{u}_{yy} - \tilde{u}_{xy}^2 = -c_1^2 f^2$ such that $\tilde{\mathbf{m}} = \nabla\tilde{u} : \mathcal{X} \mapsto \widetilde{\mathcal{Y}}$. The proof relies on basic (complex) calculus. The calculations are straightforward and therefore omitted.

2.6 Literature review on numerical methods

Next we present a brief overview of the current literature on numerical methods for the Monge-Ampère equation. In Chapter 6.2 we will consider both elliptic and hyperbolic problems for the design of optical systems. Henceforth, we present references to numerical methods for both types here.

2.6.1 The elliptic Monge-Ampère Equation

We start with the elliptic variant. Due to its connection with optimal transport theory [77, Ch. 12] and optics [29], the solution u can be obtained directly, by solving the PDE itself, or indirectly by solving an equivalent problem, e.g., the Monge-Kantorovich transport problem. These solution methods are therefore either called direct or indirect solvers. Here we only focus on the direct solvers. For a more detailed survey including indirect solvers we refer to [68, Ch. 5] and [26]. Furthermore, solutions methods for the elliptic variant with Dirichlet and transport boundary conditions exist.

Notable authors who directly solve the standard Monge-Ampère equation are:

- **Oliker, Prussner.** The authors [61, 62] developed a finite difference method to solve the Monge-Ampère equation with Dirichlet boundary conditions on a convex domain. Furthermore, proof of existence and uniqueness of solutions are provided.
- **Feng, Neilan.** Feng and Neilan developed a generalization to viscosity solutions called moment solutions [26–28]. They furthermore developed a vanishing moment method to solve second-order nonlinear PDEs, among others the standard Monge-Ampère equation.

- **Lakkis, Pryer.** Lakkis and Pryer introduced a Galerkin finite element method for nonlinear elliptic PDE [47, 50], which can be used for the standard Monge-Ampère equation with Dirichlet boundary conditions [51].

- **Froese, Oberman, Benamou, Prins.** Using the theory of viscosity solutions [32], the standard Monge-Ampère equation can be solved using a monotone finite difference scheme yielding a convex solution [5, 32–34, 59, 60]. The transport boundary condition is treated in [6, 31]. A convergence proof for the scheme and treatment of the boundary condition is also given. Prins [65, p. 109] expands on [6, 34] to develop a wide-stencil algorithm for the interior domain and introduces a signed-distance function for the boundary.

- **Dean, Glowinski, Caboussat, Prins, Beltman, Glowinski, Yadav.** The authors solve the standard Monge-Ampère equation equation in a least-squares sense [22, 23]. The method solves the MA equation with Dirichlet boundary conditions and was improved upon by including a relaxation method in [16] and was extended to three dimensions in [15]. Prins et al. [66] extended the approach to incorporate transport boundary conditions. Beltman et al. generalized the method to arbitrary orthogonal coordinate systems [4]. Yadav et al. [161, 162, 164] extend the least-squares approach by Prins et al. to the general Monge-Ampère equation. Later it was further generalized to generating functions by Romijn et al. [68, Ch. 6]. In Chapter 5 we both generalize and improve upon it to encompass the hyperbolic Monge-Ampère equation.

- **Loeper, Rapetti, Chang et al., Wu et al.** In [17, 54, 81, 85] the Monge-Ampère equation is solved using finite differences and Newton iteration. Wu et al. derive the Monge-Ampère equation [81] for a lens surface with a point source and near-field target, yielding a generalized Monge-Ampère equation and include transport boundary conditions. Similarly, Chang et al. [17] derive the Monge-Ampère equation for a collimated light source, spherical output wavefront using a double freeform lens. Loeper and Rapetti incorperate periodic boundary conditions in their solver for the standard Monge-Ampère equation.

- **Bonnet, Mirebeau** An application of finite difference schemes to non-imaging optics is found in [10]. Here the Monge-Ampère equation with transport boundary conditions is solved using a monotone finite differences scheme whose solutions are stable by addition of a constant. Convergence of the scheme is proven in the setting of quadratic optimal transport. The proposed numerical scheme is based on a reformulation

of the Monge-Ampère operator as a maximum of semilinear operators. By using Selling's formula in order to choose the parameters of the discretization for the 2D Monge-Ampère equation, the scheme can be solved efficiently.

- **Benamou, Duval** For the Monge-Ampère equation with transport boundary conditions and convex target set, the work of Benamou and Duval [8] presents an adaption of a lattice basis reduction scheme for the Monge-Ampère equation with Dirichlet boundary conditions [7]. Given two absolutely continuous measures, where the measure on the target set has convex support, the method demonstrated in [8] yields a fast adaptive method to numerically solve the optimal transport problem. Furthermore, convergence of the method as the grid step size goes to zero is shown and numerical experiments are demonstrated.

2.6.2 The hyperbolic Monge-Ampère Equation

For the hyperbolic Monge-Ampère equation we can be more concise than for the elliptic problem, as there is little known literature on the subject.

- **Tuy.** In [76] the Cauchy problem for $u_{xy} - A_4 = 0$ is treated. The author constructs a two-stage Runge-Kutta method on a triangular domain to solve the equation. Because the hyperbolic Monge-Ampère equation can be written as a system of five equations, where each equation is of the form $u_{\alpha\beta} - f = 0$, it is posited that the method therefore can solve the Monge-Ampère equation.

- **Brickell, Westcott.** In the work of Brickell and Westcott, [12] the Monge-Ampère equation is derived for the use of reflector design. Subsequently, the equation is written as a system of quasi-linear first order PDEs. In a follow up paper [79], the authors use standard finite differences to solve the quasi-linear PDE system by stepping from one horizontal grid line to the next, such that the grid lines shorten with each step, finally forming an isosceles triangle as domain. All dependent function values are prescribed on the initial grid line.

- **Howard.** The author formulates a system of equations equivalent to the Monge-Ampère equation [43]. Two objective functions using this system are established and minimized subsequently by using a (Sobolev) gradient descent method.

2.7 Discussion

From the literature research it has become clear that very few numerical methods for the hyperbolic Monge-Ampère equation exist. The existing methods furthermore do not handle the boundary conditions well. In [76, 79] the MA equation is only solved on part of the domain and in [43] four different configurations of Neumann boundary conditions have been tried for two different examples. Neither of the papers produce any results of numerical convergence. In the remaining of this thesis one of our aims is to provide numerical methods which adequately handle the boundary conditions and to demonstrate convergence of these schemes.

Chapter 3

The Method of Characteristics

The hyperbolic Monge-Ampère equation has proven to be more difficult to solve than its elliptic counterpart. This is due to the existence of two mutually coupled families of characteristics. Neglecting or even mishandling these characteristics, e.g., by using standard finite difference methods, generally yields unstable algorithms for the hyperbolic problem, while the (complex) characteristics of the elliptic variant may be safely ignored. To see why this poses an issue, we consider the domain of dependence. The domain of dependence of an interior point (x_0, y_0) is the region enclosed by the two characteristics through (x_0, y_0) facing back to the boundary. The solution $u(x_0, y_0)$ depends on all function values $u(x, y)$ with (x, y) in this domain. Conversely, the characteristics emanating from (x_0, y_0) bound the domain of influence of (x_0, y_0), which is the region where the solution is determined by $u(x_0, y_0)$. Figure 3.1 shows one such example where the blue and black lines indicate (a few of the) characteristics, the point (x_0, y_0) equals $(0.363, -0.167)$, and is denoted by the black dot and where the red and yellow parts indicate the domain of dependence and the domain of influence, respectively. Hence, boundary data determine the solution in the interior domain, and the interior determines the number of boundary conditions that should be imposed.

To start with, we introduce a general framework for second-order nonlinear hyperbolic PDEs and subsequently restrict ourselves to the general Monge-Ampère equation and later the special case of the standard hyperbolic Monge-Ampère equation. The method of characteristics gives rise to two mutually coupled ODE systems which we solve numerically in the next chapter. We classify the characteristics at the boundary as entering or leaving characteristics and fix a so-called initial strip on which we prescribe Cauchy boundary conditions. The remaining boundary conditions then follow from the course of the characteristics by considering the domain of dependence.

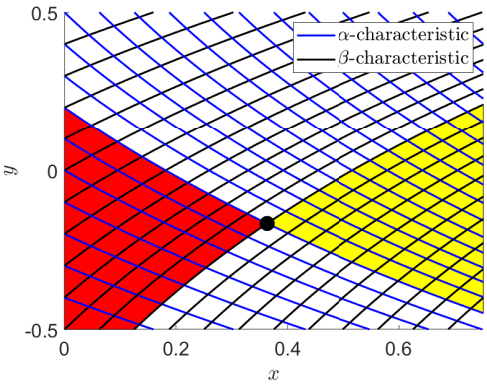


Figure 3.1: Schematic representation of the domain of dependence (red) and domain of influence (yellow) of (x_0, y_0) (black dot).

We have organized the chapter as follows. Section 3.1 introduces the method of characteristics. We start by giving an overview of the fundamental and necessary concepts of the method of characteristics (MOC), and expand on the work by Courant and Hilbert [20] by deriving ODE systems for the solution of a general second-order hyperbolic PDE. In Sections 3.2 and 3.3 we derive the ODE systems for the general and standard hyperbolic Monge-Ampère equation, respectively, and discuss how to determine the necessary boundary conditions and the resulting existence and uniqueness of the solution.

3.1 Method of characteristics for a second-order nonlinear hyperbolic PDE

We start by introducing the method of characteristics for a general nonlinear second-order PDE in two variables. To this end we assume, unless explicitly stated otherwise, that all functions are continuous and have continuous derivatives of all orders involved. Let the PDE of interest be given by

$$F(x, y, u, p, q, r, s, t) = 0, \quad (x, y) \in \Omega, \tag{3.1}$$

where $u = u(x, y)$, $p = u_x$, $q = u_y$, $r = u_{xx}$, $s = u_{xy}$, $t = u_{yy}$ and $\Omega \subseteq \mathbb{R}^2$ the domain of interest.

3.1.1 An introduction to the method of characteristics

In this section we give a brief introduction to the method of characteristics.

Let C_b be a curve in the (x, y)-plane, parameterized by $\lambda \in I$, for an interval $I \subset \mathbb{R}$, i.e., $C_b = \{(X(\lambda), Y(\lambda)) \mid \lambda \in I\}$ with $X, Y : I \to \mathbb{R}$. Let C_0 be a corresponding curve in (x, y, z)-space, which we also parameterize by $\lambda \in I$, i.e., $C_0 = \{(X(\lambda), Y(\lambda), U(\lambda)) \mid \lambda \in I\}$ where $U : I \to \mathbb{R}$. The projection of C_0 on the (x, y)-plane yields the curve C_b, see Figure 3.2. We call C_b the base curve of C_0, or simply the base curve.

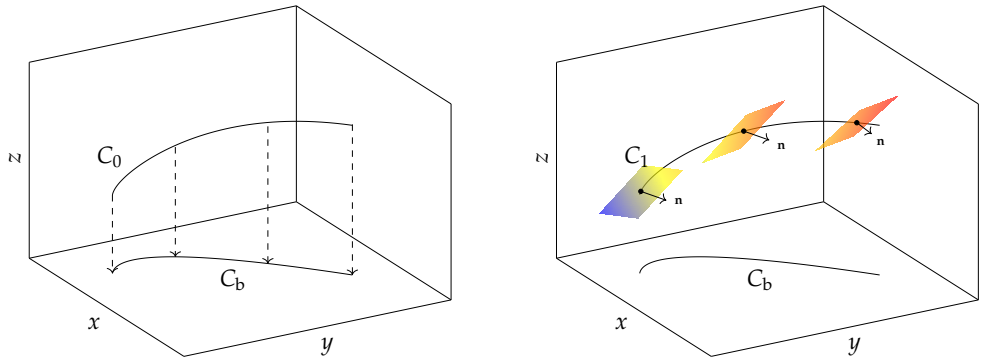


Figure 3.2: Schematic representation of the curves C_b, C_0 and the C_1-strip. Three tangent planes and their normal vectors $\mathbf{n}$ are drawn.

A base curve C_b is said to be differentiable if the corresponding map $\lambda \mapsto (X(\lambda), Y(\lambda))$ is differentiable for every $\lambda \in I$. A curve is regular if it is differentiable and the tangent vector has non-zero length for all $\lambda \in I$. We generally assume C_b to be regular, implying that $\frac{\mathrm{d}}{\mathrm{d}\lambda}(X(\lambda), Y(\lambda))^{\mathrm{T}} \neq \mathbf{0}$, or equivalently $X_\lambda^2 + Y_\lambda^2 \neq 0$ for all $\lambda \in I$, where a subscript denotes differentiation.

Let $\mathbf{v} = (X_\lambda, Y_\lambda, U_\lambda)^{\mathrm{T}}$ be the tangent vector to $(X(\lambda), Y(\lambda), U(\lambda)) \in C_0$ with $\lambda \in I$. A plane through the point $(X(\lambda), Y(\lambda), U(\lambda))$ with normal vector $\mathbf{n}$ is tangent to the curve C_0 if $\mathbf{v} \cdot \mathbf{n} = 0$. To identify those planes let $\mathbf{n} = (P, Q, -1)^{\mathrm{T}}$ with $P, Q : I \to \mathbb{R}$. Note that if the third component $n_3 \neq 0$, $\mathbf{n}$ can always be reduced to such a form by scaling the components. If for each point on C_0 we fix the tangent plane, then the collection of C_0 together with said tangent planes forms a so-called C_1-strip, i.e.,

$$C_1 = \{(X(\lambda), Y(\lambda), U(\lambda), P(\lambda), Q(\lambda)) \mid \lambda \in I\}, \tag{3.2}$$

sometimes referred to as a strip of first order. Figure 3.2 shows the C_1-strip for three tangent planes with corresponding normals. We use the notation C_1 interchangeably to denote either the strip's order (e.g. a C_1-strip), or the strip itself (e.g. equation (3.2)). From $\mathbf{v} \cdot \mathbf{n} = 0$ it follows that

$$PX_\lambda + QY_\lambda - U_\lambda = 0, \tag{3.3}$$

which is the strip condition of first order, in short, the strip condition. Note that so far the C_1-strip and the strip condition have no connection to the PDE (3.1).

Let $u(x,y)$ be a solution of (3.1), then $z = u(x,y)$ is called an integral surface of (3.1). An integral surface $z = u(x,y)$ naturally induces a C_1-strip. Given a base curve C_b, let $u(\lambda) := u(X(\lambda), Y(\lambda))$, $p(\lambda) := p(X(\lambda), Y(\lambda))$ and $q(\lambda) := q(X(\lambda), Y(\lambda))$. The normal of the integral surface $u(x,y) - z = 0$ is given by $\mathbf{n} = (u_x, u_y, -1)^{\mathrm{T}}$ in (x,y,z)-space. Hence the strip $C_1 = \{(x(\lambda), y(\lambda), u(\lambda), p(\lambda), q(\lambda)) \mid \lambda \in I\}$ is obtained. From the chain rule we conclude

$$u_\lambda = u_x X_\lambda + u_y Y_\lambda = pX_\lambda + qY_\lambda. \tag{3.4}$$

which is identical to the strip condition (3.3) with $P = p = u_x$, $Q = q = u_y$ and $U = u$, the solution of (3.1).

One can naturally generalize first-order strips to higher-order strips. A C_2-strip, a second order strip, consists of the C_1-strip together with the tangent planes of the curves
$(X(\lambda), Y(\lambda), P(\lambda))$ and $(X(\lambda), Y(\lambda), Q(\lambda))$. Higher-order strip conditions are also found naturally in the following way: with $(X(\lambda), Y(\lambda), P(\lambda))$ we can associate two functions $R, S : I \to \mathbb{R}$ such that the normal vector of a tangent plane is $(R, S, -1)$. The tangent vector of $(X(\lambda), Y(\lambda), P(\lambda))$ equals $(X_\lambda, Y_\lambda, P_\lambda)$. The same reasoning as before applies and we find

$$P_\lambda = RX_\lambda + SY_\lambda. \tag{3.5}$$

Analogously, for $(X(\lambda), Y(\lambda), Q(\lambda))$ let the normal vector of a tangent plane be $(\tilde{S}, T, -1)$, it then follows that

$$Q_\lambda = \tilde{S}X_\lambda + TY_\lambda. \tag{3.6}$$

Note that the functions S and $\tilde{S}$ are not necessarily equal as $(R, S, -1)^{\mathrm{T}}$ should be perpendicular to the curve $(X_\lambda, Y_\lambda, P_\lambda)$ and $(\tilde{S}, T, -1)^{\mathrm{T}}$ should be perpendicular to the curve $(X_\lambda, Y_\lambda, Q_\lambda)$. As before, an integral surface $z = u(x,y)$ induces a C_2-strip where we identify Q, R, S, $\tilde{S}$ and T with the values r, s, t via $R = r = u_{xx}(X(\lambda), Y(\lambda))$, $S = s = u_{xy}(X(\lambda), Y(\lambda))$, $\tilde{S} = s = u_{yx}(X(\lambda), Y(\lambda))$ and $T = t = u_{yy}(X(\lambda), Y(\lambda))$. We assume u is twice continuously differentiable, and therefore $S = u_{xy} = u_{yx} = \tilde{S}$. Henceforth strip conditions of second order for u become

$$p_\lambda = rX_\lambda + sY_\lambda, \tag{3.7a}$$
$$q_\lambda = sX_\lambda + tY_\lambda. \tag{3.7b}$$

For completeness we give the strip conditions of third order, viz.,

$$r_\lambda = u_{xxx}X_\lambda + u_{xxy}Y_\lambda = r_xX_\lambda + r_yY_\lambda = r_xX_\lambda + s_xY_\lambda, \tag{3.8a}$$
$$s_\lambda = u_{xyx}X_\lambda + u_{xyy}Y_\lambda = s_xX_\lambda + s_yY_\lambda = s_xX_\lambda + t_xY_\lambda = r_yX_\lambda + s_yY_\lambda, \tag{3.8b}$$
$$t_\lambda = u_{yyx}X_\lambda + u_{yyy}Y_\lambda = t_xX_\lambda + t_yY_\lambda = s_yX_\lambda + t_yY_\lambda. \tag{3.8c}$$

The process of finding higher-order strips is called extending. To clarify, the curve $C_0 = \{(X(\lambda), Y(\lambda), U(\lambda)) \mid \lambda \in I\}$ (a strip of zeroth order) is extended to a strip C_1, given by (3.2). Similarily C_1 is extended to a strip of second order, given by

$$C_2 = \{(X(\lambda), Y(\lambda), U(\lambda), P(\lambda), Q(\lambda), R(\lambda), S(\lambda), \tilde{S}(\lambda), T(\lambda)) \mid \lambda \in I\}. \tag{3.9}$$

We define C_2 to be an integral strip if there exists a C_1-strip which can be extended to the C_2-strip uniquely, solely using the PDE (3.1) and the strip conditions (3.7). In this case C_1 is called a *free strip*. If C_1 is not a free strip, additional requirements should be prescribed in order for C_1 to be extendable to an integral strip C_2. In this case we call C_1 a *characteristic strip* which implies that not all second-order derivatives of u can be determined uniquely from C_1, the PDE (3.1) and the strip conditions.

To put into context, let $C_\mathrm{b} = \{(x(\lambda), y(\lambda)) \mid \lambda \in I\}$ be a base curve, $z = u(x, y)$ be an integral surface of (3.1) and let C_0 be a corresponding zeroth order strip. Furthermore, we supplement $p = u_x$ and $q = u_y$ to obtain a first order strip C_1. If by using the PDE (3.1) and the strip conditions (3.7) we are able to determine $r = u_{xx}, s = u_{xy}$ and $t = u_{yy}$ uniquely, then the strip $C_2 = \{(x(\lambda), y(\lambda), u(\lambda), p(\lambda), q(\lambda), r(\lambda), s(\lambda), t(\lambda)) \mid \lambda \in I\}$ is an integral strip, and C_1 is a free strip, otherwise C_1 is a characteristic strip.

Note that along a free strip, but not along a characteristic strip, the derivatives u_{xx}, u_{xy} and u_{yy} can all be determined along the strip, either by being interior derivatives with respect to C_1, or by combining the PDE (3.1) with the remaining interior derivatives. To illustrate, given u, $p = u_x$ and $q = u_y$, on a vertical line segment, i.e., $X_\lambda = 0$, by differentiation with respect to y one can obtain u_{xy} and u_{yy}, and u_{xx} follows from the PDE, as will be shown in Section 3.4.

For completeness, if C_1 is a characteristic strip, its carrier C_0 will be called a characteristic curve in (x, y, z)-space, and the base curve C_b, will be called a characteristic base curve. Generally we refer to a characteristic strip, characteristic curve and characteristic base curve simply as 'the characteristic'.

Note that thus far we considered an entire curve/strip to be either free or characteristic. Formally this should be evaluated pointwise, which introduces the notion of a characteristic base point, a characteristic point and a

characteristic element for a 2-, 3- and 5-dimensional point on C_b, C_0 and C_1, respectively. This distinction is often not necessary due to the fact that every strip, which has one point in common with the integral surface and all its tangent planes equal to that of the integral surface, lies entirely on said surface. To see this consider the strip γ parameterized by $\lambda \in I$, given by $\gamma(\lambda) = (X(\lambda), Y(\lambda), U(\lambda), u_x(\lambda), u_y(\lambda))$ with strip condition $U_\lambda = u_x X_\lambda + u_y Y_\lambda$. Let (x_0, y_0, u_0) lie on the integral surface $z = u(x, y)$, so $u_0 = u(x_0, y_0)$. Furthermore, let γ pass through (x_0, y_0, u_0), i.e., there exists a λ_0 such that $(x_0, y_0, u_0) = (X(\lambda_0), Y(\lambda_0), U(\lambda_0))$ and $u_0 = u(X_0, Y_0) = u(X(\lambda_0), Y(\lambda_0))$. Let $d(\lambda) = U(\lambda) - u(X(\lambda), Y(\lambda))$ be the pointwise signed vertical distance between the integral surface $z = u(x, y)$ and the strip γ at the point $(X(\lambda), Y(\lambda), u(X(\lambda), Y(\lambda)))$. If $d \equiv 0$ then clearly γ lies on the integral surface $z = u(x, y)$. Obviously it holds that $d(\lambda_0) = 0$. Furthermore, the change in the signed distance d for $\lambda \in I$ can be found by

$$\frac{\mathrm{d}d}{\mathrm{d}\lambda} = \frac{\mathrm{d}U}{\mathrm{d}\lambda} - u_x \frac{\mathrm{d}X}{\mathrm{d}\lambda} - u_y \frac{\mathrm{d}Y}{\mathrm{d}\lambda} = \frac{\mathrm{d}U}{\mathrm{d}\lambda} - (u_x X_\lambda + u_y Y_\lambda) = \frac{\mathrm{d}U}{\mathrm{d}\lambda} - U_\lambda = 0, \tag{3.10}$$

where we applied the strip condition. Because $\frac{\mathrm{d}d}{\mathrm{d}\lambda} = 0$ for all $\lambda \in I$ and $d(\lambda_0) = 0$, $d \equiv 0$, and hence the strip lies entirely on the integral surface u.

3.1.2 The characteristic condition

An equivalent definition of a characteristic can be phrased as follows [20, p. 407]: if the differential equation $F = 0$ represents an interior differential operator along a strip C_1, then C_1 is a characteristic strip. The term interior differential operator here means that along C_1 the second order differential operator F can be expressed solely in terms of derivatives of u with respect to the parameter describing the base curve of C_1. Therefore, deriving the evolution of F along characteristics yields systems of ODEs equivalent to $F = 0$. These systems of ODEs are in general easier to solve than $F = 0$, therefore we derive and discuss the conditions under which a strip C_1 is a characteristic strip next. These conditions will be called the characteristic conditions. We will impose conditions on C_1, based on our starting equation (3.1), such that (at least one of the) second- and higher-order derivatives cannot be determined uniquely. We will discuss three different approaches to obtain the characteristic conditions.

3.1.2.1 The characteristic condition by the implicit function theorem

Fundamentally, we are looking for conditions on the solvability for the second-order derivatives r, s and t. One way to derive the characteristic condition is to apply the implicit function theorem. To this end, let the C_1-strip be parameterized by λ as before. Define

$$\mathbf{f}(x, y, u, p, q, x_\lambda, y_\lambda, p_\lambda, q_\lambda \mid r, s, t) = \begin{pmatrix} F(x, y, u, p, q, r, s, t) \\ x_\lambda r + y_\lambda s - p_\lambda \\ x_\lambda s + y_\lambda t - q_\lambda \end{pmatrix}. \tag{3.11}$$

The components of the vector-valued function $\mathbf{f}$ are formed by our PDE (3.1) and the two strip conditions (3.7) for p_λ and q_λ. The implicit function theorem [2, p. 731] states that if there exists a λ_0 such that

$$\begin{aligned} \mathbf{f}(x(\lambda_0), y(\lambda_0), u(\lambda_0), p(\lambda_0), q(\lambda_0), x_\lambda(\lambda_0), y_\lambda(\lambda_0), \\ p_\lambda(\lambda_0), q_\lambda(\lambda_0) \mid r(\lambda_0), s(\lambda_0), t(\lambda_0)) = \mathbf{0}, \end{aligned} \tag{3.12}$$

and the Jacobi matrix

$$\mathbf{A} := \frac{\partial \mathbf{f}}{\partial(r, s, t)} = \begin{pmatrix} F_r & F_s & F_t \\ x_\lambda & y_\lambda & 0 \\ 0 & x_\lambda & y_\lambda \end{pmatrix}, \tag{3.13}$$

is nonsingular, then there is an open set $\Lambda \subset \mathbb{R}$ containing λ_0 and a unique continuously differentiable function $\mathbf{g} : \Lambda \to \mathbb{R}^3$ with $\mathbf{g}(\lambda_0) = (r(\lambda_0), s(\lambda_0), t(\lambda_0))^{\mathrm{T}}$ such that

$$\mathbf{f}(x(\lambda), y(\lambda), u(\lambda), p(\lambda), q(\lambda), x_\lambda(\lambda), y_\lambda(\lambda), p_\lambda(\lambda), q_\lambda(\lambda) \mid \mathbf{g}(\lambda)) = \mathbf{0}, \tag{3.14}$$

for all $\lambda \in \Lambda$. If $D := \det(\mathbf{A}) \neq 0$, then (r, s, t) can be found uniquely along C_1, i.e., we have a free strip. Alternatively, if $D = 0$, then (r, s, t) cannot be determined uniquely, hence we have a characteristic strip. The case $D = 0$ is therefore called the characteristic condition and it can be written as

$$D = F_r y_\lambda^2 - F_s x_\lambda y_\lambda + F_t x_\lambda^2 = 0. \tag{3.15}$$

3.1.2.2 The characteristic condition by a coordinate transformation

The characteristic condition can also be derived by means of a coordinate transformation [20, p. 419]. This can be achieved with the following equivalent definition of a characteristic. If the differential equation $F = 0$ represents an interior differential operator along a strip C_1, then C_1 is a characteristic strip.

The term interior differential operator here means that along C_1 the second-order differential operator F can be expressed solely in terms of derivatives of u with respect to the parameter describing the base curve of C_1.

Let C_1 be the strip of interest with corresponding base curve C_b as shown in Figure 3.3. We introduce the coordinate transformation

$$(x, y) \to (\phi(x, y), \lambda(x, y)), \tag{3.16}$$

where λ is the parameter along the curve C_b and ϕ leads away from C_b. Recall

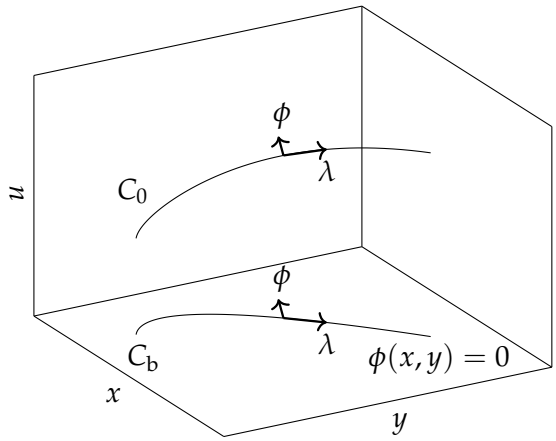


Figure 3.3: Schematic representation of a curve C_0, with its base curve C_b and the accompanying parameters along and perpendicular to the curves.

that we require C_b to be a regular curve, i.e., $x_\lambda^2 + y_\lambda^2 \neq 0$. Adopting the new coordinates, the derivatives of u are:

$$u_x = u_\phi \phi_x + u_\lambda \lambda_x, \tag{3.17a}$$
$$u_y = u_\phi \phi_y + u_\lambda \lambda_y, \tag{3.17b}$$
$$u_{xx} = u_{\phi\phi}(\phi_x)^2 + 2u_{\phi\lambda}\phi_x\lambda_x + u_{\lambda\lambda}(\lambda_x)^2 + u_\phi \phi_{xx} + u_\lambda \lambda_{xx}, \tag{3.17c}$$
$$u_{xy} = u_{\phi\phi}\phi_x\phi_y + u_{\phi\lambda}(\phi_x\lambda_y + \phi_y\lambda_x) + u_{\lambda\lambda}\lambda_x\lambda_y + u_\phi \phi_{xy} + u_\lambda \lambda_{xy}, \tag{3.17d}$$
$$u_{yy} = u_{\phi\phi}(\phi_y)^2 + 2u_{\phi\lambda}\phi_y\lambda_y + u_{\lambda\lambda}(\lambda_y)^2 + u_\phi \phi_{yy} + u_\lambda \lambda_{yy}. \tag{3.17e}$$

Using these relations, one can construct a function G such that

$$F(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy}) = G(\phi, \lambda, u, u_\phi, u_\lambda, u_{\phi\phi}, u_{\phi\lambda}, u_{\lambda\lambda}) = 0. \tag{3.18}$$

Recall that a C_1-strip is a characteristic if not all higher order derivatives can be determined uniquely along the strip. As p and q are known on a C_1-strip, both u_ϕ and u_λ can be obtained provided the Jacobian of the coordinate

transformation (3.16) $\phi_x\lambda_y - \phi_y\lambda_x \neq 0$. Because λ is the parameter along the strip, $(u_\phi)_\lambda$ and $(u_\lambda)_\lambda$ can naturally be found by differentiating along the strip. Therefore, for C_1 to be characteristic, $u_{\phi\phi}$ should be undetermined. If $G_{u_{\phi\phi}} = 0$ then $u_{\phi\phi}$ cannot be determined from $G = 0$. Hence differentiating (3.18) and applying (3.17c)-(3.17e) we find

$$\begin{aligned} G_{u_{\phi\phi}}(\phi, \lambda, u, u_\phi, u_\lambda, u_{\phi\phi}, u_{\phi\lambda}, u_{\lambda\lambda}) &= F_{u_{xx}}\frac{\partial u_{xx}}{\partial u_{\phi\phi}} + F_{u_{xy}}\frac{\partial u_{xy}}{\partial u_{\phi\phi}} + F_{u_{yy}}\frac{\partial u_{yy}}{\partial u_{\phi\phi}} \\ &= F_r\phi_x^2 + F_s\phi_x\phi_y + F_t\phi_y^2 \\ &= 0, \end{aligned} \tag{3.19}$$

which is the characteristic condition. Assuming $\phi_y \neq 0$, this can further be rewritten as

$$F_r\left(\frac{\phi_x}{\phi_y}\right)^2 + F_s\frac{\phi_x}{\phi_y} + F_t = 0. \tag{3.20}$$

To see that this is equivalent to (3.15) consider the following: ϕ is constant along C_b, therefore for fixed x, y it should not depend on λ, i.e.,

$$\phi_\lambda = \phi_x x_\lambda + \phi_y y_\lambda = 0, \tag{3.21}$$

which is equivalent to

$$\frac{\phi_x}{\phi_y} = -\frac{y_\lambda}{x_\lambda}. \tag{3.22}$$

Substituting this in (3.20) yields the previously found characteristic condition (3.15).

3.1.2.3 The characteristic condition for second-order strips

In the previous sections we established relations such that C_1 is uniquely defined, while C_2-strips are not. Although the characteristic condition $D = 0$ is fundamental, it yields no practical means to determine the evolution of the solution along a C_1-strip. In this section we show that the characteristic condition obtained also holds for second-order strips, which does provide insights on how to determine the evolution of a C_1-strip. This evolution will be further discussed in Section 3.1.3 and Section 3.1.4.

Not all second-order derivatives can be determined along a characteristic C_1-strip. Likewise, not all third-order derivatives can be determined uniquely either. To determine relations for the derivatives of r, s, t we apply

strip conditions (3.8), together with additional relations, which are found by differentiating the PDE (3.1) with respect to x and y, viz.,

$$\frac{\mathrm{d}F}{\mathrm{d}x} = F_x + F_u p + F_p r + F_q s + F_r r_x + F_s s_x + F_t t_x = 0, \tag{3.23a}$$

$$\frac{\mathrm{d}F}{\mathrm{d}y} = F_y + F_u q + F_p s + F_q t + F_r r_y + F_s s_y + F_t t_y = 0. \tag{3.23b}$$

Combining these with strip conditions (3.8) with $X(\lambda) = x(\lambda)$ etc., yields two systems of equations

$$\mathbf{A}\begin{pmatrix} r_x \\ s_x \\ t_x \end{pmatrix} = \begin{pmatrix} -F^x \\ r_\lambda \\ s_\lambda \end{pmatrix}, \qquad \mathbf{A}\begin{pmatrix} r_y \\ s_y \\ t_y \end{pmatrix} = \begin{pmatrix} -F^y \\ s_\lambda \\ t_\lambda \end{pmatrix}, \tag{3.24}$$

where the equations are formed by collecting the x- and y-derivatives, respectively, and where

$$F^x := F_x + F_u p + F_p r + F_q s, \tag{3.25a}$$

$$F^y := F_y + F_u q + F_p s + F_q t, \tag{3.25b}$$

and $\mathbf{A}$ is given by (3.13). Because F becomes an interior operator along a characteristic strip, r_λ, s_λ and t_λ can be determined along a C_2-strip while not all of r_x, s_x, t_x, r_y, s_y and t_y can. Therefore $\mathbf{A}$ should be singular. Hence we obtain once more the characteristic condition (3.15).

3.1.3 Compatibility conditions

Because $D = \det(\mathbf{A}) = 0$, the systems in (3.24) may not have solutions. In this section we will derive compatibility conditions such that solutions do exist. Consider the rank of $\mathbf{A}$. The rank of a matrix equals the order of the largest non-vanishing minor, which is known as the determinantal rank. The matrix $\mathbf{A}$ has 9 minors of order 2, the minors formed by the lower right and lower left 2×2 submatrices are

$$M_{1,1} = \begin{vmatrix} y_\lambda & 0 \\ x_\lambda & y_\lambda \end{vmatrix} = y_\lambda^2, \qquad M_{1,3} = \begin{vmatrix} x_\lambda & y_\lambda \\ 0 & x_\lambda \end{vmatrix} = x_\lambda^2, \tag{3.26}$$

where M_{ij} denotes the minor formed by deleting the ith row and jth column. Because $x_\lambda^2 + y_\lambda^2 \neq 0$, $M_{1,1}$ and $M_{1,3}$ cannot be 0 simultaneously, hence $D = 0$ implies $\mathrm{rank}(\mathbf{A}) = 2$. From (3.24) we conclude that solutions $(r_x, s_x, t_x)^{\mathrm{T}}$ and $(r_y, s_y, t_y)^{\mathrm{T}}$ do not always exist if $\mathbf{A}$ is singular. Therefore we require the left- and right-hand sides of (3.24) to be compatible, i.e., the vectors $(-F^x, r_\lambda, s_\lambda)^{\mathrm{T}}$

and $(-F^y, s_\lambda, t_\lambda)^{\mathrm{T}}$ should be in the column-space of any two columns of $\mathbf{A}$. Hence the matrix of any two column vectors of $\mathbf{A}$ with either $(-F^x, r_\lambda, s_\lambda)^{\mathrm{T}}$ or $(-F^y, s_\lambda, t_\lambda)^{\mathrm{T}}$ should be singular. To this end we introduce the matrices

$$\mathbf{A}^x = \left(\begin{array}{ccc|c} F_r & F_s & F_t & -F^x \\ x_\lambda & y_\lambda & 0 & r_\lambda \\ 0 & x_\lambda & y_\lambda & s_\lambda \end{array}\right), \tag{3.27a}$$

$$\mathbf{A}^y = \left(\begin{array}{ccc|c} F_r & F_s & F_t & -F^y \\ x_\lambda & y_\lambda & 0 & s_\lambda \\ 0 & x_\lambda & y_\lambda & t_\lambda \end{array}\right). \tag{3.27b}$$

Let D^x_{klm} denote the determinant formed by selecting the columns k, l and m of $\mathbf{A}^x$ and similarly, we introduce D^y_{klm}. We find by including the first, third and fourth column

$$D^x_{134} = \begin{vmatrix} F_r & F_t & -F^x \\ x_\lambda & 0 & r_\lambda \\ 0 & y_\lambda & s_\lambda \end{vmatrix} = -(F^x x_\lambda y_\lambda + F_r y_\lambda r_\lambda + F_t x_\lambda s_\lambda), \tag{3.28a}$$

$$D^y_{134} = \begin{vmatrix} F_r & F_t & -F^y \\ x_\lambda & 0 & s_\lambda \\ 0 & y_\lambda & t_\lambda \end{vmatrix} = -(F^y x_\lambda y_\lambda + F_r y_\lambda s_\lambda + F_t x_\lambda t_\lambda), \tag{3.28b}$$

which both should equal zero due to $\mathbf{A}$ being singular. Trivially, since $D = 0$ similar calculations yield $D^x_{124} = D^x_{134}\, x_\lambda / y_\lambda$, $D^x_{234} = D^x_{134}\, y_\lambda / x_\lambda$, $D^y_{124} = D^y_{134}\, x_\lambda / y_\lambda$, $D^y_{234} = D^y_{134}\, y_\lambda / x_\lambda$ for $x_\lambda, y_\lambda \neq 0$.

Condition (3.15) turns out to be enough to determine the evolution of x, y, u, p and q along the characteristics. Conditions (3.28) are needed for the evolutions of other variables which are introduced in the next section.

3.1.4 Evolution along the characteristics

To derive the evolution of the solution along the characteristics we rewrite (3.15) as a second-order polynomial equation, viz.

$$D(\mu) = F_r \mu^2 - F_s \mu + F_t = 0, \tag{3.29}$$

where we introduced $\mu = y_\lambda / x_\lambda$ assuming $x_\lambda \neq 0$. In case $x_\lambda = 0$, we can use $\tilde{\mu} = x_\lambda / y_\lambda$ instead. Solving (3.29) for μ yields two roots a and b, viz.

$$a = \frac{F_s + \sqrt{\Delta}}{2F_r}, \qquad b = \frac{F_s - \sqrt{\Delta}}{2F_r}, \tag{3.30}$$

where $\Delta = F_s^2 - 4F_rF_t$ is the discriminant. The discriminant allows us to classify the differential operator F. If at a point $\mathbf{x}_0 = (x_0, y_0) \in \Omega$ the discriminant $\Delta(\mathbf{x}_0) > 0$, then the PDE (3.1) is hyperbolic in that point [20, p. 420]. Naturally there exists a (small) neighborhood of $\mathbf{x}_0$ for which the PDE is hyperbolic. Similarly we call the PDE parabolic in $\mathbf{x}_0$ if $\Delta(\mathbf{x}_0) = 0$ and elliptic if $\Delta(\mathbf{x}_0) < 0$. If for all $\mathbf{x}_0 \in \Omega$ we have $\Delta(\mathbf{x}_0) > 0$, then F is called hyperbolic, or hyperbolic in the entire domain.

In the following we restrict ourselves to the hyperbolic case. By definition, we have two separate families of characteristic curves defined by $a(x(\lambda), y(\lambda))x_\lambda = y_\lambda$ and $b(x(\lambda), y(\lambda))x_\lambda = y_\lambda$, respectively, passing through the point $(x(\lambda), y(\lambda))$.

In the above discussion we restricted ourselves to the case $x_\lambda \neq 0$. For completeness, an equivalent formulation of the method is outlined below where $x_\lambda = 0$ may occur but $y_\lambda \neq 0$ is required. We first introduce $\tilde{\mu} = x_\lambda / y_\lambda$ and write (3.15) as

$$D(\tilde{\mu}) = F_r - F_s\tilde{\mu} + F_t\tilde{\mu}^2 = 0, \tag{3.31}$$

with roots

$$\tilde{a} = \frac{F_s + \sqrt{\Delta}}{2F_t}, \qquad \tilde{b} = \frac{F_s - \sqrt{\Delta}}{2F_t}, \tag{3.32}$$

with Δ as before. The classification of the PDE, as would be expected, does not change. For simplicity of notation we assume $F_r \neq 0$ and only use (3.30), i.e., we continue with the roots a, b in the remaining of the chapter.

We can express F_r, F_s and F_t in terms of a, b and Δ using (3.30) as

$$F_t = \frac{ab}{a-b}\sqrt{\Delta}, \qquad F_s = \frac{a+b}{a-b}\sqrt{\Delta}, \qquad F_r = \frac{1}{a-b}\sqrt{\Delta}. \tag{3.33}$$

Alternatively, we can express a and b in terms of F_r, F_s and F_t as

$$a + b = \frac{F_s}{F_r}, \qquad ab = \frac{F_t}{F_r}. \tag{3.34}$$

Using the definition of μ we find $y_\lambda / x_\lambda = a$ or $y_\lambda / x_\lambda = b$, implying we have two distinct families of characteristics, one induced by a, and the other induced by b. To distinguish the characteristics, we write $x = x(\alpha)$, $y = y(\alpha)$ and $x = x(\beta)$, $y = y(\beta)$ for the characteristic induced by a and b, respectively. Henceforth α and β effectively take over the role of λ. As such, instead of x_λ we write $\frac{\mathrm{d}x}{\mathrm{d}\alpha} = x_\alpha$ for the derivative of x with respect to α, and similarly for the other variables and for differentiation with respect to β.

The matrix $\mathbf{A}$, given in (3.13) actually represents two distinct matrices, $\mathbf{A}^\alpha$ and $\mathbf{A}^\beta$, because the derivatives w.r.t. λ can be associated with both α and β. Because further derivations for either characteristic is done analogously for the other, we will only treat the characteristic induced by a, the α-characteristic, and postulate the results for the β-characteristic. Note that for fixed $(x_0, y_0) \in \Omega$, two characteristics pass through (x_0, y_0), i.e., both the α- and β-characteristic. Because the matrix $\mathbf{A}^\alpha$ has rank 2, the rows are linearly dependent and therefore κ_1^α and κ_2^α exist such that

$$\begin{pmatrix} F_r \\ F_s \\ F_t \end{pmatrix} = \kappa_1^\alpha \begin{pmatrix} x_\alpha \\ y_\alpha \\ 0 \end{pmatrix} + \kappa_2^\alpha \begin{pmatrix} 0 \\ x_\alpha \\ y_\alpha \end{pmatrix}. \tag{3.35}$$

The first row gives $x_\alpha = F_r/\kappa_1^\alpha$. By definition we have $y_\alpha = a x_\alpha$ and hence $y_\alpha = a F_r/\kappa_1^\alpha$. The third row then yields $\kappa_2^\alpha = F_t/y_\alpha = \kappa_1^\alpha F_t/(aF_r)$ which yields $\kappa_2^\alpha = b\kappa_1^\alpha$ by (3.34). For the sake of brevity we write $\kappa^\alpha = \kappa_1^\alpha$. Then (3.35) reduces to

$$F_r = \kappa^\alpha x_\alpha, \tag{3.36a}$$
$$F_s = \kappa^\alpha y_\alpha + b\kappa^\alpha x_\alpha, \tag{3.36b}$$
$$F_t = b\kappa^\alpha y_\alpha. \tag{3.36c}$$

The evolution of x, y, u, p and q along the characteristics can be determined from (3.36), the strip conditions (3.4) and (3.7), respectively, giving

$$x_\alpha = \frac{F_r}{\kappa^\alpha}, \tag{3.37a}$$
$$y_\alpha = a\frac{F_r}{\kappa^\alpha}, \tag{3.37b}$$
$$u_\alpha = (p + aq)\frac{F_r}{\kappa^\alpha}, \tag{3.37c}$$
$$p_\alpha = (r + as)\frac{F_r}{\kappa^\alpha}, \tag{3.37d}$$
$$q_\alpha = (s + at)\frac{F_r}{\kappa^\alpha}, \tag{3.37e}$$

where the choice of κ^α determines the parametric scaling of the base curve. The evolution of r, s and t can be obtained using the compatibility conditions (3.28). To that purpose we rewrite (3.28) as the underdetermined linear system

$$\begin{pmatrix} F_r y_\alpha & F_t x_\alpha & 0 \\ 0 & F_r y_\alpha & F_t x_\alpha \end{pmatrix} \begin{pmatrix} r_\alpha \\ s_\alpha \\ t_\alpha \end{pmatrix} = -x_\alpha y_\alpha \begin{pmatrix} F^x \\ F^y \end{pmatrix}. \tag{3.38}$$

By the rank-nullity theorem [53, p. 175] the general solution of (3.38) reads

$$\begin{pmatrix} r_\alpha \\ s_\alpha \\ t_\alpha \end{pmatrix} = - \begin{pmatrix} \frac{F^x x_\alpha}{F_r} \\ 0 \\ \frac{F^y y_\alpha}{F_t} \end{pmatrix} + \theta^\alpha \begin{pmatrix} \frac{F_t x_\alpha}{F_r y_\alpha} \\ -1 \\ \frac{F_r y_\alpha}{F_t x_\alpha} \end{pmatrix}, \tag{3.39}$$

where the first term is the particular solution with $s_\alpha = 0$ and the second term is an element of the null space of the matrix for arbitrary θ^α. Rewriting this, using (3.36), yields

$$\begin{pmatrix} r_\alpha \\ s_\alpha \\ t_\alpha \end{pmatrix} = -\frac{1}{\kappa^\alpha} \begin{pmatrix} F^x \\ 0 \\ \frac{F^y}{b} \end{pmatrix} + \theta^\alpha \begin{pmatrix} b \\ -1 \\ \frac{1}{b} \end{pmatrix}. \tag{3.40}$$

Because the ODE system (3.37) depends on a and b, what remains is to determine the evolution of a and b along the characteristics. These are straightforwardly calculated by taking the derivative of (3.30) w.r.t. α. At this point we do not further evaluate a_α or b_α as it yields no meaningful insight. Treating the β-characteristic analogously to the α-characteristic, we similarly obtain κ^β and θ^β, and the ODE systems read

$$\begin{aligned}
x_\alpha &= \frac{F_r}{\kappa^\alpha}, & x_\beta &= \frac{F_r}{\kappa^\beta}, \\
y_\alpha &= a\frac{F_r}{\kappa^\alpha}, & y_\beta &= b\frac{F_r}{\kappa^\beta}, \\
u_\alpha &= (p + aq)\frac{F_r}{\kappa^\alpha}, & u_\beta &= (p + bq)\frac{F_r}{\kappa^\beta}, \\
p_\alpha &= (r + as)\frac{F_r}{\kappa^\alpha}, & p_\beta &= (r + bs)\frac{F_r}{\kappa^\beta}, \\
q_\alpha &= (s + at)\frac{F_r}{\kappa^\alpha}, & q_\beta &= (s + bt)\frac{F_r}{\kappa^\beta}, \\
r_\alpha &= \theta^\alpha b - \frac{1}{\kappa^\alpha}F^x, & r_\beta &= \theta^\beta a - \frac{1}{\kappa^\beta}F^x, \\
s_\alpha &= -\theta^\alpha, & s_\beta &= -\theta^\beta, \\
t_\alpha &= \frac{1}{b}\left(\theta^\alpha - \frac{1}{\kappa^\alpha}F^y\right), & t_\beta &= \frac{1}{a}\left(\theta^\beta - \frac{1}{\kappa^\beta}F^y\right), \\
a_\alpha &= \left(\frac{F_s + \sqrt{\Delta}}{2F_r}\right)_\alpha, & a_\beta &= \left(\frac{F_s + \sqrt{\Delta}}{2F_r}\right)_\beta, \\
b_\alpha &= \left(\frac{F_s - \sqrt{\Delta}}{2F_r}\right)_\alpha, & b_\beta &= \left(\frac{F_s - \sqrt{\Delta}}{2F_r}\right)_\beta.
\end{aligned} \tag{3.41}$$

Note the coupling between the two ODE systems, for example the evolution of r_α depends on b, which forms the direction of the other characteristic via $bx_\beta = y_\beta$. Also note that the variables θ^α and θ^β cannot be determined, else the evolution of r, s and t would be determined, violating the definition of a characteristic.

3.2 MOC for the general hyperbolic Monge-Ampère equation

Consider the general Monge-Ampère equation

$$F(x,y,u,p,q,r,s,t) = (rt - s^2) + A_1 r + A_2 s + A_3 t + A_4 = 0, \tag{3.42}$$

where $u = u(x,y)$, $p = u_x$, $q = u_y$, $r = u_{xx}$, $s = u_{xy}$, $t = u_{yy}$ and $A_i = A_i(x,y,u,p,q)$ for $i = 1,\dots,4$. We apply the method of characteristics to (3.42), which is a special case of the problem considered in the previous section. We start by calculating the discriminant of the characteristic condition, viz.

$$\begin{aligned}
\Delta &= F_s^2 - 4F_rF_t \\
&= (-2s + A_2)^2 - 4(t + A_1)(r + A_3) \\
&= 4s^2 - 4A_2 s + A_2^2 - 4rt - 4A_3 t - 4A_1 r - 4A_1 A_3 \\
&= 4((s^2 - rt) - A_1 r - A_2 s - A_3 t) - 4A_1 A_3 + A_2^2 \\
&= 4A_4 + A_2^2 - 4A_1 A_3 \\
&= 4A_4 + \Delta_{\tilde{F}},
\end{aligned} \tag{3.43a}$$

$$\Delta_{\tilde{F}} = \tilde{F}_s^2 - 4\tilde{F}_r\tilde{F}_t, \tag{3.43b}$$

where we introduced $\tilde{F} = A_1 r + A_2 s + A_3 t$, which is the quasi-linear part of the general Monge-Ampère equation. It follows that the Monge-Ampère equation (3.42) is hyperbolic if $\Delta > 0$. Because $A_i(x,y,u,p,q)$ may depend on the solution itself, the condition may in general not be checked a priori.

For the above case, classification of the Monge-Ampère equation (3.42) based on Δ, is equivalent to the classification by linearization introduced in Section 2.1 given by (2.9) and (2.5).

Using (3.30), we find the roots of the characteristic condition to be

$$a, b = \frac{B_1 \pm \sqrt{\Delta}}{B_2}, \tag{3.44a}$$

$$B_1 = -2s + A_2 = F_s, \tag{3.44b}$$

$$B_2 = 2(t + A_1) = 2F_r. \tag{3.44c}$$

Note that $B_2 = 0$ can occur, and when it does, one should consider $\tilde{a}$ and $\tilde{b}$ instead of a and b, given by (3.32). The derivations are similar, henceforth we will not explicitly treat this case for brevity. In the following we will calculate the evolution of each of the relevant variables along the characteristics. To do so we use (3.41) and we make use of the sum, chain and product rule along characteristics, e.g.,

$$(A+B)_\alpha = A_\alpha + B_\alpha, \tag{3.45a}$$

$$(AB)_\alpha = A_\alpha B + AB_\alpha, \tag{3.45b}$$

$$\Big(\phi(A)\Big)_\alpha = \phi'(A)A_\alpha, \tag{3.45c}$$

for the α-characteristic (and similar for the β-characteristic) for general functions $\phi : \mathbb{R} \to \mathbb{R}$ and $A, B : (g_1, g_2, \dots, g_n) \mapsto \mathbb{R}$ with $g_i = g_i(x, y)$ and $i = 1, \dots, n$, for $n \in \mathbb{N}$ and $n > 0$. Before we can efficiently elaborate on the evolutions, we will require a few auxiliary relations between the second derivatives of u and a, b. By applying (3.33) to (3.42) we find

$$r = \frac{ab\sqrt{\Delta}}{a-b} - A_3, \qquad s = -\frac{(a+b)\sqrt{\Delta}}{2(a-b)} + \frac{A_2}{2}, \qquad t = \frac{\sqrt{\Delta}}{a-b} - A_1, \tag{3.46}$$

which by (3.44b) and (3.44c) imply

$$B_1 = \frac{(a+b)\sqrt{\Delta}}{a-b}, \tag{3.47a}$$

$$B_2 = \frac{2\sqrt{\Delta}}{a-b}. \tag{3.47b}$$

Starting with the evolution of u, given by (3.41);

$$u_\alpha = (p+aq)\frac{F_r}{\kappa^\alpha}, \qquad u_\beta = (p+bq)\frac{F_r}{\kappa^\beta}, \tag{3.48}$$

we find using (3.44c) that

$$u_\alpha = (p+aq)\frac{B_2}{2\kappa^\alpha}, \qquad u_\beta = (p+bq)\frac{B_2}{2\kappa^\beta}. \tag{3.49}$$

The above expressions cannot be simplified further. Using (3.46) and (3.44c) ,

we find for the the evolution of p, given by (3.41), that

$$\begin{aligned} p_\alpha &= (r+as)\frac{F_r}{\kappa^\alpha} \\ &= \left(\frac{ab\sqrt{\Delta}}{a-b} - A_3 - \frac{a(a+b)\sqrt{\Delta}}{2(a-b)} + a\frac{A_2}{2}\right)\frac{B_2}{2\kappa^\alpha} \\ &= \left(\frac{a(b-a)\sqrt{\Delta}}{2(a-b)} + \frac{aA_2-2A_3}{2}\right)\frac{B_2}{2\kappa^\alpha} \\ &= \left(\frac{a}{2}(A_2-\sqrt{\Delta}) - A_3\right)\frac{B_2}{2\kappa^\alpha}, \end{aligned} \tag{3.50a}$$

and

$$\begin{aligned} p_\beta &= (r+bs)\frac{F_r}{\kappa^\beta} \\ &= \left(\frac{ab\sqrt{\Delta}}{a-b} - A_3 - \frac{b(a+b)\sqrt{\Delta}}{2(a-b)} + b\frac{A_2}{2}\right)\frac{B_2}{2\kappa^\beta} \\ &= \left(\frac{b(a-b)\sqrt{\Delta}}{2(a-b)} + \frac{bA_2-2A_3}{2}\right)\frac{B_2}{2\kappa^\beta} \\ &= \left(\frac{b}{2}(A_2+\sqrt{\Delta}) - A_3\right)\frac{B_2}{2\kappa^\beta}. \end{aligned} \tag{3.50b}$$

Similarily for the evolution of q we obtain

$$\begin{aligned} q_\alpha &= (s+at)\frac{F_r}{\kappa^\alpha} \\ &= \left(-\frac{(a+b)\sqrt{\Delta}}{2(a-b)} + \frac{A_2}{2} + \frac{a\sqrt{\Delta}}{a-b} - aA_1\right)\frac{B_2}{2\kappa^\alpha} \\ &= \left(\frac{(a-b)\sqrt{\Delta}}{2(a-b)} + \frac{A_2-2aA_1}{2}\right)\frac{B_2}{2\kappa^\alpha} \\ &= \left(\frac{1}{2}(A_2+\sqrt{\Delta}) - aA_1\right)\frac{B_2}{2\kappa^\alpha}, \end{aligned} \tag{3.51a}$$

and

$$\begin{aligned} q_\beta &= (s+bt)\frac{F_r}{\kappa^\beta} \\ &= \left(-\frac{(a+b)\sqrt{\Delta}}{2(a-b)} + \frac{A_2}{2} + \frac{b\sqrt{\Delta}}{a-b} - bA_1\right)\frac{B_2}{2\kappa^\beta} \\ &= \left(\frac{(b-a)\sqrt{\Delta}}{2(a-b)} + \frac{A_2-2bA_1}{2}\right)\frac{B_2}{2\kappa^\beta} \\ &= \left(\frac{1}{2}(A_2-\sqrt{\Delta}) - bA_1\right)\frac{B_2}{2\kappa^\beta}. \end{aligned} \tag{3.51b}$$

Because not all of the second derivatives of u can be determined along the characteristics, we cannot find the evolutions of r, s and t. This is also represented in equations (3.41) by the factors θ^α and θ^β, which cannot be determined. For the evolution of a and b we will make use of the evolution of r, s and t by using (3.41) such that the undetermined factors θ^α and θ^β cancel out. Furthermore, because the evolutions of r, s and t are undetermined, we aim to get rid of r, s and t in the expressions for the evolution of a and b. In Section 3.3 we show that a_β can be determined for the standard Monge-Ampère equation, see equations (3.70) and the surrounding discussion, but a_α can not. Because the general Monge-Ampère equation can be reduced to the standard Monge-Ampère equation by choosing the coefficients A_i accordingly, we know that only a_β can be determined for the general Monge-Ampère equation. Here we therefore present an expression for a_β and not for a_α. Using (3.44a), we henceforth obtain

$$\begin{aligned} a_\beta &= \frac{B_{1,\beta} + (\sqrt{\Delta})_\beta}{B_2} - \frac{B_1+\sqrt{\Delta}}{B_2}\cdot\frac{B_{2,\beta}}{B_2} \\ &= \frac{B_{1,\beta} + (\sqrt{\Delta})_\beta - aB_{2,\beta}}{B_2}. \end{aligned} \tag{3.52}$$

We simplify the expression for a_β using the definitions (3.44b), (3.44c) and (3.41), viz.

$$\begin{aligned} B_{1,\beta} - aB_{2,\beta} &= -2s_\beta + A_{2,\beta} - 2a(t_\beta + A_{1,\beta}) \\ &= A_{2,\beta} - 2aA_{1,\beta} - 2(s_\beta + at_\beta) \\ &= A_{2,\beta} - 2aA_{1,\beta} + \frac{2}{\kappa^\beta}F^y, \end{aligned} \tag{3.53}$$

where F^y is defined by (3.25b) and a comma separated subscript implies differentiation w.r.t. the second argument, e.g. $A_{1,\beta}$ is the derivative of A_1 w.r.t.

β. Note, if one had chosen to derive the evolution along the characteristics according to $\tilde{a}, \tilde{b}$ instead of a, b, the term F^x would have appeared instead of F^y, and A_3 instead of A_1. For the expression for a_β, we furthermore require the evolution of $\sqrt{\Delta}$, which reads

$$(\sqrt{\Delta})_\beta = \frac{1}{2\sqrt{\Delta}}\Delta_\beta = \frac{1}{2\sqrt{\Delta}}\Big(4A_{4,\beta} + (\Delta_{\tilde{F}})_\beta\Big), \tag{3.54a}$$

$$(\Delta_{\tilde{F}})_\beta = 2A_2A_{2,\beta} - 4(A_{1,\beta}A_3 - A_1A_{3,\beta}). \tag{3.54b}$$

Henceforth we obtain

$$a_\beta = \left(\frac{A_{2,\beta} - 2aA_{1,\beta} + \frac{2}{\kappa^\beta}F^y}{2\sqrt{\Delta}} + \frac{4A_{4,\beta} + (\Delta_{\tilde{F}})_\beta}{4\Delta}\right)(a-b). \tag{3.55a}$$

Analogously we obtain the evolution of b along the α-characteristic, viz.

$$b_\alpha = \left(\frac{A_{2,\alpha} - 2bA_{1,\alpha} + \frac{2}{\kappa^\alpha}F^y}{2\sqrt{\Delta}} - \frac{4A_{4,\alpha} + (\Delta_{\tilde{F}})_\alpha}{4\Delta}\right)(a-b). \tag{3.55b}$$

What remains is to express $F^y = F_y + F_u q + F_p s + F_q t$ in terms of the coefficients A_i and the variables u, p, q, a and b. To do so, let

$$A_i^y = A_{i,y} + A_{i,u}q + A_{i,p}s + A_{i,q}t, \text{ for } i = 1, \dots, 4. \tag{3.56}$$

For the general Monge-Ampère equation we find using the definition of F, given by (3.42), that

$$F^y = A_1^y r + A_2^y s + A_3^y t + A_4^y, \tag{3.57}$$

where we grouped the terms involving r, s and t. Using equations (3.46) one can eliminate r, s and t. The process is straightforward but tedious, and barely any simplifications can be made. Therefore we express F^y as a product viz.

$$\begin{aligned} F^y &= \begin{pmatrix} r \\ s \\ t \\ 1 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} A_{1,y} & A_{1,u} & A_{1,p} & A_{1,q} \\ A_{2,y} & A_{2,u} & A_{2,p} & A_{2,q} \\ A_{3,y} & A_{3,u} & A_{3,p} & A_{3,q} \\ A_{4,y} & A_{4,u} & A_{4,p} & A_{4,q} \end{pmatrix} \begin{pmatrix} 1 \\ q \\ s \\ t \end{pmatrix} \\ &= \begin{pmatrix} \frac{ab\sqrt{\Delta}}{a-b} - A_3 \\ -\frac{(a+b)\sqrt{\Delta}}{2(a-b)} + \frac{A_2}{2} \\ \frac{\sqrt{\Delta}}{a-b} - A_1 \\ 1 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} A_{1,y} & A_{1,u} & A_{1,p} & A_{1,q} \\ A_{2,y} & A_{2,u} & A_{2,p} & A_{2,q} \\ A_{3,y} & A_{3,u} & A_{3,p} & A_{3,q} \\ A_{4,y} & A_{4,u} & A_{4,p} & A_{4,q} \end{pmatrix} \begin{pmatrix} 1 \\ q \\ -\frac{(a+b)\sqrt{\Delta}}{2(a-b)} + \frac{A_2}{2} \\ \frac{\sqrt{\Delta}}{a-b} - A_1 \end{pmatrix}. \end{aligned} \tag{3.58}$$

In total we obtain two coupled ODE systems of which the α-system is given by $y_\lambda / x_\lambda = a$ and equations (3.49), (3.50a), (3.51a), (3.55b). Similarily, the β-system is given by $y_\lambda / x_\lambda = b$ and equations (3.49), (3.50b), (3.51b), (3.55a). All ODEs depend only on x, y, u, p, q, a, b and the functions A_i and their derivatives via

$$A_{i,\beta} = A_{i,x} x_\beta + A_{i,y} y_\beta + A_{i,u} u_\beta + A_{i,p} p_\beta + A_{i,q} q_\beta, \tag{3.59}$$

and similar for $A_{i,\alpha}$.

3.3 MOC for the standard hyperbolic Monge-Ampère equation

In the remaining of the chapter we will consider the standard hyperbolic Monge-Ampère equation. Recall that the hyperbolic Monge-Ampère equation is given by

$$F(x, y, u, p, q, r, s, t) = rt - s^2 + f^2 = 0 \quad \text{for} \quad (x, y) \in \Omega, \tag{3.60}$$

for $\Omega \subseteq \mathbb{R}^2$, the unknown function $u = u(x, y) \in C^3(\Omega)$ and the known function $f = f(x, y) \in C^1(\Omega)$, with $f \neq 0$ on $\overline{\Omega}$. The derivatives of F are

$$F_r = t, \qquad F_s = -2s, \qquad F_t = r. \tag{3.61}$$

Furthermore, the characteristic equation is given by

$$D(\mu) = F_r \mu^2 - F_s \mu + F_t = t\mu^2 + 2s\mu + r = 0, \tag{3.62}$$

and the corresponding discriminant is

$$\Delta = F_s^2 - 4F_r F_t = 4s^2 - 4rt = 4f^2, \tag{3.63}$$

which is positive, and hence (3.60) is hyperbolic. The two real and distinct roots of (3.62) are given by

$$a = \frac{F_s + \sqrt{\Delta}}{2F_r} = \frac{-s + f}{t}, \qquad b = \frac{F_s - \sqrt{\Delta}}{2F_r} = \frac{-s - f}{t}. \tag{3.64}$$

for $t \neq 0$. In case $t = 0$, we express a and b as

$$a = \frac{r}{-s - f}, \qquad b = \frac{r}{-s + f}, \tag{3.65}$$

instead. Furthermore, the auxiliary functions F^x and F^y are given by (3.25) and read

$$\begin{aligned} F^x &= F_x + F_u p + F_p r + F_q s = F_x = (f^2)_x = 2ff_x, \\ F^y &= F_y + F_u q + F_p s + F_q t = F_y = (f^2)_y = 2ff_y. \end{aligned} \tag{3.66}$$

From (3.60) and (3.64), it follows that

$$a + b = \frac{-2s}{t}, \qquad a - b = \frac{2f}{t}, \qquad ab = \frac{s^2 - f^2}{t^2} = \frac{r}{t}. \tag{3.67}$$

From these relations we can express the second derivatives in terms of a, b and f as follows

$$r = \frac{2ab}{a-b} f, \qquad s = -\frac{a+b}{a-b} f, \qquad t = \frac{2f}{a-b}. \tag{3.68}$$

The systems of ODEs (3.41) then read

$$\begin{aligned} x_\alpha &= \frac{t}{\kappa^\alpha}, & x_\beta &= \frac{t}{\kappa^\beta}, \\ y_\alpha &= a\frac{t}{\kappa^\alpha}, & y_\beta &= b\frac{t}{\kappa^\beta}, \\ u_\alpha &= (p + aq)\frac{t}{\kappa^\alpha}, & u_\beta &= (p + bq)\frac{t}{\kappa^\beta}, \\ p_\alpha &= (r + as)\frac{t}{\kappa^\alpha}, & p_\beta &= (r + bs)\frac{t}{\kappa^\beta}, \\ q_\alpha &= (s + at)\frac{t}{\kappa^\alpha}, & q_\beta &= (s + bt)\frac{t}{\kappa^\beta}, \\ r_\alpha &= \theta^\alpha b - \frac{1}{\kappa^\alpha}\left(f^2\right)_x, & r_\beta &= \theta^\beta a - \frac{1}{\kappa^\beta}\left(f^2\right)_x, \\ s_\alpha &= -\theta^\alpha, & s_\beta &= -\theta^\beta, \\ t_\alpha &= \frac{1}{b}\Big(\theta^\alpha - \frac{1}{\kappa^\alpha}\left(f^2\right)_y\Big), & t_\beta &= \frac{1}{a}\Big(\theta^\beta - \frac{1}{\kappa^\beta}\left(f^2\right)_y\Big). \end{aligned} \tag{3.69}$$

Note that the ODE systems in (3.69) contain four parameters, viz. $\kappa^\alpha, \kappa^\beta$, which are determined by an appropriate scaling, and θ^α and θ^β, which are free parameters. Consequently, the derivatives of r, s and t cannot be rewritten such that they no longer depend on θ^α as this would uniquely determine r, s, t along the characteristic strip which contradicts the definition of a characteristic strip. By differentiating the expressions for a and b and using (3.69) we find

the evolution along the α-characteristic, viz.

$$\begin{aligned} a_\alpha &= \frac{1}{t}\left(1-\frac{a}{b}\right)\theta^\alpha + \frac{1}{\kappa^\alpha}\left(f_x + a f_y\right) + \frac{a}{b}\frac{1}{\kappa^\alpha t}\left(f^2\right)_y \\ &= \frac{1}{t}\left(1-\frac{a}{b}\right)\theta^\alpha + \frac{1}{\kappa^\alpha}\left(f_x + \frac{a^2}{b} f_y\right), \end{aligned} \tag{3.70a}$$

$$\begin{aligned} b_\alpha &= \frac{\theta^\alpha - f_x\frac{t}{\kappa^\alpha} - f_y a\frac{t}{\kappa^\alpha}}{t} - \frac{b}{t}\left(\theta^\alpha - \frac{2 f f_y}{\kappa^\alpha}\right)\frac{1}{b} \\ &= -\frac{1}{\kappa^\alpha}(f_x + b f_y). \end{aligned} \tag{3.70b}$$

The expression for a_α could be rewritten, for example by using (3.68), but due to $a \neq b$, it will include the unknown θ^α. More fundamentally, we cannot determine a_α explicitly as then both a and b can be uniquely determined along the α-characteristic, from which r, s, t would follow by (3.68), which contradicts the definition of a characteristic.

Note that we are free to choose κ^α and κ^β due to the freedom in parameterization of the base curve. In the following we conveniently choose $\kappa^\alpha = \kappa^\beta = F_r = t$. Using relations (3.68) the ODE system (3.69) reduces to

$$\begin{aligned}
x_\alpha &= 1, & x_\beta &= 1, \\
y_\alpha &= a, & y_\beta &= b, \\
u_\alpha &= p + aq, & u_\beta &= p + bq, \\
p_\alpha &= -af, & p_\beta &= bf, \\
q_\alpha &= f, & q_\beta &= -f, \\
r_\alpha &= \theta^\alpha b - (a-b) f_x, & r_\beta &= \theta^\beta a - (a-b) f_x, \\
s_\alpha &= -\theta^\alpha, & s_\beta &= -\theta^\beta, \\
t_\alpha &= \frac{1}{b}\Big(\theta^\alpha - (a-b) f_y\Big), & t_\beta &= \frac{1}{a}\Big(\theta^\beta - (a-b) f_y\Big), \\
a_\alpha &= \frac{a-b}{2f}\Big[\left(1-\frac{a}{b}\right)\theta^\alpha + f_x + \frac{a^2}{b} f_y\Big], & a_\beta &= \frac{a-b}{2f}(f_x + a f_y), \\
b_\alpha &= \frac{b-a}{2f}(f_x + b f_y), & b_\beta &= -\frac{a-b}{2f}\Big[\left(\frac{b}{a}-1\right)\theta^\beta + f_x + \frac{b^2}{a} f_y\Big].
\end{aligned} \tag{3.71}$$

There is a lot of redundancy in these equations which directly follows from (3.64), (3.67) and (3.68). We therefore reduce (3.71) by omitting the equations

involving θ_α and θ_β. What remains are the ODE systems

$$\begin{aligned} x_\alpha &= 1, & x_\beta &= 1, \\ y_\alpha &= a, & y_\beta &= b, \\ u_\alpha &= p + aq, & u_\beta &= p + bq, \\ p_\alpha &= -af, & p_\beta &= bf, \\ q_\alpha &= f, & q_\beta &= -f, \\ b_\alpha &= \frac{b-a}{2f}(f_x + bf_y), & a_\beta &= \frac{a-b}{2f}(f_x + af_y), \end{aligned} \tag{3.72}$$

which we integrate to obtain the solution u. Recall that: "every strip, which has one point in common with the integral surface and all its tangent planes equal to that of the integral surface, lies entirely on said surface", which furthermore justifies the reduction of (3.71) to (3.72), as only u, p, q and the characteristics, so u, p, q, a and b, need to be known.

3.4 Boundary conditions

Next we discuss the appropriate boundary conditions to prescribe for solving the hyperbolic Monge-Ampère equation. For brevity and simplicity of notation, we only treat the standard Monge-Ampère equation here. The results are straightforwardly extended to the general Monge-Ampère equation and one such instance will be discussed in Section 4.3 and, more specifically, demonstrated by equations (4.40).

We solve (3.60) on a rectangular domain $\Omega = [x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}]$, for $x_{\min}, x_{\max}, y_{\min}, y_{\max} \in \mathbb{R}$, with $x_{\min} < x_{\max}$, $y_{\min} < y_{\max}$ and we choose $\{x_{\min}\} \times [y_{\min}, y_{\max}]$ as the initial base curve. We extend the initial base curve to an initial C_1-strip by supplementing it with $u_{\mathrm{W}}, p_{\mathrm{W}} : [y_{\min}, y_{\max}] \to \mathbb{R}$, where we prescribe $u(x_{\min}, y) = u_{\mathrm{W}}(y)$ and $p(x_{\min}, y) = p_{\mathrm{W}}(y)$ for some u_{W}, p_{W}. The subscript 'W' is used as the values are prescribed on the western part of $\partial\Omega$. The conditions imposed on the initial base curve are also known as Cauchy boundary conditions.

If the initial strip is a free strip, then prescribing u_{W}, p_{W} uniquely determines the C_2-strip as an extension of the initial base curve. To verify this, we check whether the characteristic condition (3.29) holds. Therefore we parameterize the initial base curve as $x(\lambda) = x_{\min}$, $y(\lambda) = y_{\min} + \lambda(y_{\max} - y_{\min})$, $\lambda \in [0, 1]$, then $x_\lambda = 0$, $y_\lambda = y_{\max} - y_{\min} \neq 0$ and the characteristic condition yields $ty_\lambda^2 + 2sx_\lambda y_\lambda + rx_\lambda^2 = ty_\lambda^2 = 0$. Hence if $t \neq 0$ on the initial strip, i.e., if $u''_{\mathrm{W}}(y) \neq 0$, the initial strip is a free strip. Henceforth we assume $u''_{\mathrm{W}}(y) \neq 0$,

which then implies the initial base curve uniquely extends to a C_2-strip, and we can uniquely determine q, a, b, r, s, t on the initial strip via

$$q(x_{\min}, y) = u'_{\mathrm{W}}(y), \tag{3.73a}$$
$$t(x_{\min}, y) = u''_{\mathrm{W}}(y), \tag{3.73b}$$
$$s(x_{\min}, y) = p'_{\mathrm{W}}(y), \tag{3.73c}$$
$$a(x_{\min}, y) = \frac{-s(x_{\min}, y) + f_{\mathrm{W}}(y)}{t(x_{\min}, y)}, \tag{3.73d}$$
$$b(x_{\min}, y) = \frac{-s(x_{\min}, y) - f_{\mathrm{W}}(y)}{t(x_{\min}, y)}, \tag{3.73e}$$
$$r(x_{\min}, y) = \frac{2a(x_{\min}, y)b(x_{\min}, y)}{a(x_{\min}, y) - b(x_{\min}, y)} f_{\mathrm{W}}(y), \tag{3.73f}$$

where $f_{\mathrm{W}}(y) := f(x_{\min}, y)$. For the lower and upper boundary, i.e., for $y \in \{y_{\min}, y_{\max}\}$, the required boundary conditions are more delicate. To understand this let $\mathbf{x}_\alpha = (x_\alpha, y_\alpha)$ and $\mathbf{x}_\beta = (x_\beta, y_\beta)$ denote the tangent vectors of the characteristics. Let $(x_{\mathrm{b}}, y_{\mathrm{b}}) = \mathbf{b} \in \partial\Omega$ and $\hat{\mathbf{n}}$ the outward unit normal vector on the boundary. We classify the α-characteristics, and similarly the β-characteristic, based on whether they are entering or leaving the domain as follows

- Leaving characteristic if $\mathbf{x}_\alpha(\mathbf{b}) \cdot \hat{\mathbf{n}} \geq 0$,
- Entering characteristic if $\mathbf{x}_\alpha(\mathbf{b}) \cdot \hat{\mathbf{n}} < 0$,

which is schematically shown in Figure 3.4 for three possible α-characteristics and $y = y_{\min}$.

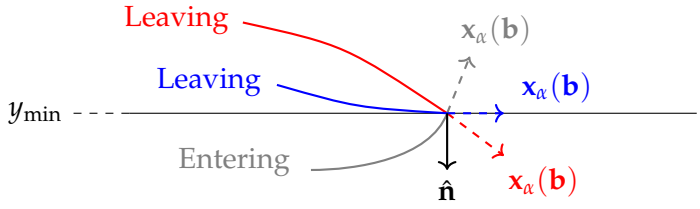


Figure 3.4: Schematic classification of two leaving and one entering characteristics.

We assume $a(x, y)$ and $b(x, y)$ to be well defined and distinct, i.e., $a(x, y) \neq b(x, y)$ for all $(x, y) \in \Omega$, else, by (3.68), the Monge-Ampère equation would not be well defined or not hyperbolic. Furthermore, the assumptions imply there are two characteristics, one of each family, passing through each point

$(x, y) \in \Omega$. By classifying the characteristics as entering or leaving, one can determine if and how many boundary conditions need to be prescribed at each boundary point.

We distinguish three possible cases: one characteristic entering, two characteristics entering and zero characteristics entering at the point $\mathbf{b}$ on the boundary.

Case 1. One characteristic leaving and one entering the domain, either

$$\mathbf{x}_\alpha \cdot \hat{\mathbf{n}} \geq 0, \quad \mathbf{x}_\beta \cdot \hat{\mathbf{n}} < 0 \qquad \text{or} \qquad \mathbf{x}_\alpha \cdot \hat{\mathbf{n}} < 0, \quad \mathbf{x}_\beta \cdot \hat{\mathbf{n}} \geq 0. \tag{3.74}$$

An example of this case is shown in Figure 3.5 at $y = y_{\min}$, with N, W, S, E denoting the western, northern, southern and eastern boundary segments of the domain, respectively. The curves denote one α- (dashed) and one β-characteristic (dotted). At $\mathbf{b}$, the values u, p, q and b can be computed from the ODE system for the α-characteristic. However, a cannot be determined because the evolution of a along the α-characteristic is unknown. Therefore we should impose one boundary condition, which is the initial condition for the entering β-characteristic, such that a can be computed. We can either prescribe a directly, or prescribe either r, s or t and compute a from (3.68). We prescribe one of the functions $a_S(x) = a(x, y_{\min})$, $r_S(x) = u_{xx}(x, y_{\min})$, $s_S(x) = u_{xy}(x, y_{\min})$ or $t_S(x) = u_{yy}(x, y_{\min})$, such that a is given by either of the following expressions:

$$a(\mathbf{b}) = a_S(x_b), \tag{3.75a}$$

$$a(\mathbf{b}) = \frac{b(\mathbf{b}) r_S(x_b)}{r_S(x_b) - 2b(\mathbf{b}) f(\mathbf{b})}, \tag{3.75b}$$

$$a(\mathbf{b}) = \frac{s_S(x_b) - f(\mathbf{b})}{s_S(x_b) + f(\mathbf{b})}, \tag{3.75c}$$

$$a(\mathbf{b}) = b(\mathbf{b}) + \frac{2f(\mathbf{b})}{t_S(x_b)}. \tag{3.75d}$$

Depending on the function prescribed, the remaining three follow from (3.68). A similar approach is taken in case the β-characteristic leaves the domain and the α-characteristic enters.

Case 2. Two characteristics entering the domain, i.e.,

$$\mathbf{x}_\alpha \cdot \hat{\mathbf{n}} < 0, \qquad \mathbf{x}_\beta \cdot \hat{\mathbf{n}} < 0. \tag{3.76}$$

In this situation the values of u, p, q, a and b cannot be determined (see Figure 3.6). Therefore we prescribe u and its normal derivative, which is q if the boundary is horizontal (as in Figure 3.6), or p if it is vertical. The calculation of the relevant variables at $y = y_{\min}$ and $y = y_{\max}$ follow analogously

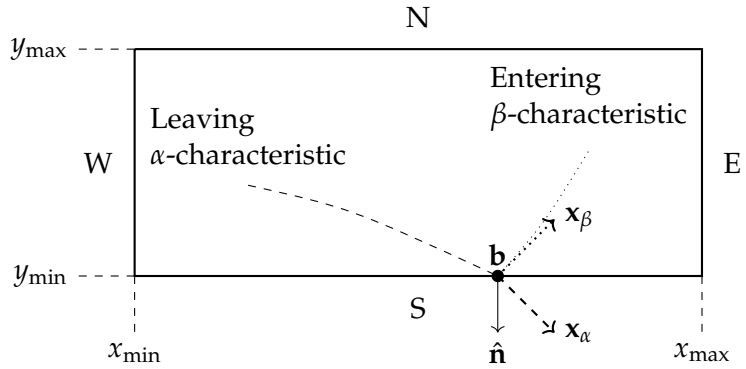


Figure 3.5: Schematic overview of a rectangular domain where the α-characteristic leaves, and the β-characteristic enters the domain.

to (3.73). As an example we consider the boundary at $y = y_{\min}$. Let $u_S, q_S : [x_{\min}, x_{\max}] \to \mathbb{R}$ be given and let $u(x, y_{\min}) = u_S(x)$ and $q(x, y_{\min}) = q_S(x)$. We obtain u, p, a, b, r, s and t at $y = y_b$ via

$$p(x, y_{\min}) = u_S'(x), \tag{3.77a}$$
$$r(x, y_{\min}) = u_S''(x), \tag{3.77b}$$
$$s(x, y_{\min}) = q_S'(x), \tag{3.77c}$$
$$a(x, y_{\min}) = -\frac{r(x, y_{\min})}{s(x, y_{\min}) + f_S(x)}, \tag{3.77d}$$
$$b(x, y_{\min}) = -\frac{r(x, y_{\min})}{s(x, y_{\min}) - f_S(x)}, \tag{3.77e}$$
$$t(x, y_{\min}) = \frac{2 f_S(x)}{a(x, y_{\min}) - b(x, y_{\min})}. \tag{3.77f}$$

We require $u_S''(x) \neq 0$ in this case, such that the hyperbolicity condition $a(x, y_b) \neq b(x, y_b)$ is satisfied. Note that the initial strip is one example of *Case 2*, where two characteristics enter. To see this, note that

$$\begin{aligned} \mathbf{x}_\alpha \cdot \hat{\mathbf{n}} &= -x_\alpha = -1 < 0, \\ \mathbf{x}_\beta \cdot \hat{\mathbf{n}} &= -x_\beta = -1 < 0, \end{aligned} \tag{3.78}$$

thus classifying both as entering characteristics.

Case 3. Two characteristics leaving the domain, i.e.,

$$\mathbf{x}_\alpha \cdot \hat{\mathbf{n}} \geq 0, \qquad \mathbf{x}_\beta \cdot \hat{\mathbf{n}} \geq 0. \tag{3.79}$$

Here, we should not prescribe anything at all as all values are known, or can be determined by integrating the ODE systems of the α- and β-characteristics, see Figure 3.7. Note that this situation is identical to that of an interior point.

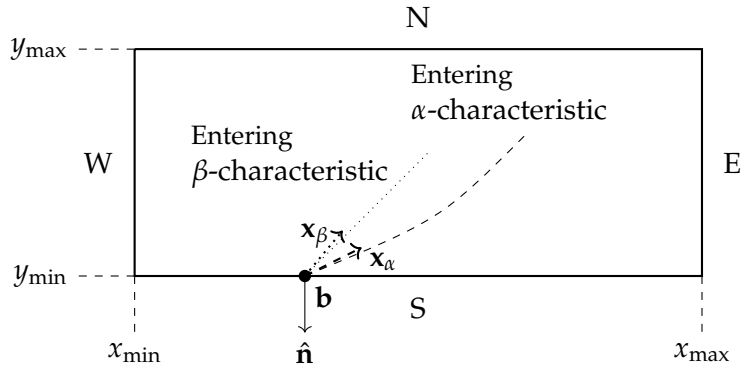


Figure 3.6: Schematic overview of a rectangular domain where both an α- and β-characteristic enter the domain at $y = y_{\min}$.

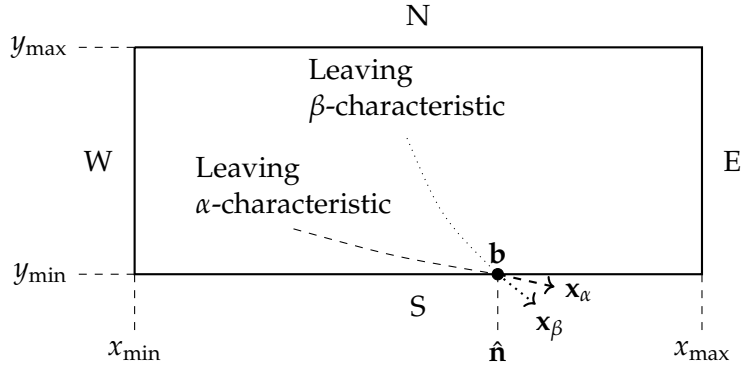


Figure 3.7: Schematic overview of a rectangular domain where both an α- and β-characteristic leave the domain $y = y_{\min}$.

Note that at $x = x_{\max}$, $x_\alpha = 1$ and $x_\beta = 1$ such that $\mathbf{x}_\alpha \cdot \hat{\mathbf{n}} \geq 0$ and $\mathbf{x}_\beta \cdot \hat{\mathbf{n}} \geq 0$, hence this boundary segment coincides with Case 3. The possible exception being the corner points $(x_{\max}, y_{\min})$ and $(x_{\max}, y_{\max})$ because the normal is not uniquely defined. In this case the point should be treated as in the Cases 1 or 2. Classifying the boundary segment $x = x_{\max}$ as Case 3 is a direct consequence of the choice $\kappa^\alpha = \kappa^\beta = t$, such that $x_\alpha = 1$ and $x_\beta = 1$ for the two characteristic families.

3.5 Summary and discussion

We presented the general formulation of the method of characteristics for a nonlinear second-order hyperbolic PDE in two variables. We derived condi-

tions which determine whether a curve is characteristic. Along these characteristics the PDE reduces to two systems of ODEs which are mutually coupled. We applied the MOC to the Monge-Ampère equation, which explicitly shows the coupling of the characteristics. We discussed how, for the (general) Monge-Ampère equation, the direction of the characteristics at the boundary determines the number of boundary conditions which should be prescribed. What remains is to show existence, uniqueness and regularity of solutions to the Monge-Ampère equation given these boundary conditions. Currently little is known about these topics and we refer the reader to the only known relevant literature, to the best of the authors knowledge, viz. [42] and [75]. The former deals with an initial base curve of infinite length, which is not applicable to our purposes. The latter shows existence and uniqueness results for solutions in a generalized sense, though application of the results to our case have proven to be cumbersome and beyond the scope of this thesis.

The ODE systems derived in this chapter present a means to compute the solution to the PDE. In the following chapter we introduce numerical methods based on these ODEs.

Chapter 4

MOC: Numerical Methods and Results

In the previous chapter we introduced mutually coupled systems of ODEs which are equivalent to the hyperbolic Monge-Ampère equation. Next we introduce numerical methods to solve the hyperbolic Monge-Ampère equation by solving the systems of ODEs.

These systems describe the evolution of the solution along two families of characteristics, while the ODEs themselves also depend on the solution. For the standard Monge-Ampère equation, the ODEs are parameterized with respect to the x-coordinate, which we utilize by employing explicit one-step methods, e.g., forward Euler and Runge-Kutta, w.r.t. the x-direction. For these one-step methods we introduce a rectangular grid. Because the evolution of the characteristics is determined by the systems of ODEs and the solution, the direction of the characteristics is a priori unknown and hence the numerical approximation of a characteristic passing through a grid point will generally not pass through another grid point. We apply one-step methods to all grid points on one grid line, i.e., a line $x_i = \text{const}$, at once. This step results in an approximation of the solution on the next grid line at x_{i+1}. By subsequently using spline interpolation and applying boundary conditions, we approximate the solution in the grid points at x_{i+1}. This process is repeated until an approximation for the solution is obtained for the whole of the computational domain.

Because the system of ODEs are mutually coupled, each characteristic depends on neighboring characteristics of the other family. This is of no issue in the interior of the domain due to interpolation between the two families of characteristics. At the boundary we determine which characteristics are entering or leaving the domain. This consideration yields which variables

cannot be determined at the boundary, and hence which boundary conditions need to be prescribed.

Three explicit numerical one-step methods are elaborated for the systems of ODEs, viz. forward Euler, modified Euler and Runge-Kutta. Furthermore, because analytical solutions are not always known, we utilize an integral formulation of the Monge-Ampère equation to evaluate the error of the numerical approximation of the solution. The numerical errors, that is both the error obtained by the integral formulation and the global discretization error, depend on both the one-step method and the spline interpolant used. We introduce a method to determine the distance between grid lines such that neither of those errors dominates the other. This allows us to estimate and test the rate of convergence for the numerical method for integration and interpolation combined. We present various examples, among which an example where all boundary segments require different boundary conditions, one example where the number of required boundary conditions varies along the boundary, and one example with discontinuous third derivatives, for which no analytical solution is known. Lastly, we also show three cases for the general Monge-Ampère equation where we not only consider integration in the positive x-direction, but integration in all four cardinal directions.

4.1 Numerical methods

In the previous chapter we found that the standard Monge-Ampère equation corresponds to the coupled ODE systems

$$\begin{aligned}
x_\alpha &= 1, & x_\beta &= 1,\\
y_\alpha &= a, & y_\beta &= b,\\
u_\alpha &= p + aq, & u_\beta &= p + bq,\\
p_\alpha &= -af, & p_\beta &= bf,\\
q_\alpha &= f, & q_\beta &= -f,\\
b_\alpha &= \frac{b-a}{2f}(f_x + bf_y), & a_\beta &= \frac{a-b}{2f}(f_x + af_y).
\end{aligned} \tag{4.1}$$

In order to solve (4.1) we rewrite it as

$$\frac{\mathrm{d}\mathbf{v}^\alpha}{\mathrm{d}\alpha} = \mathbf{g}^\alpha(\mathbf{v}^\alpha, a), \qquad \frac{\mathrm{d}\mathbf{v}^\beta}{\mathrm{d}\beta} = \mathbf{g}^\beta(\mathbf{v}^\beta, b), \tag{4.2a}$$

$$\mathbf{v}^{\alpha} = \begin{pmatrix} x \\ y \\ u \\ p \\ q \\ b \end{pmatrix}, \ \mathbf{v}^{\beta} = \begin{pmatrix} x \\ y \\ u \\ p \\ q \\ a \end{pmatrix}, \ \mathbf{g}^{\alpha} = \begin{pmatrix} 1 \\ a \\ p + aq \\ -af \\ f \\ \frac{b-a}{2f}(f_x + bf_y) \end{pmatrix}, \ \mathbf{g}^{\beta} = \begin{pmatrix} 1 \\ b \\ p + bq \\ bf \\ -f \\ \frac{a-b}{2f}(f_x + af_y) \end{pmatrix}. \tag{4.2b}$$

Equations (4.2) are two mutually coupled systems because the evolution of a and b are determined on the other characteristic. By supplying initial conditions, the problem can be treated as a Cauchy problem which we solve by numerical integration.

For our numerical grid we choose N_x points in the x-direction and N_y in the y-direction. Let the grid points $\mathbf{x}_{i,j}$ be given by $\mathbf{x}_{i,j} = (x_i, y_j)$ for $i = 1, \ldots, N_x, j = 1, \ldots, N_y$. We choose the grid to be equidistant in the y-direction with spacing $h_y = (y_{\max} - y_{\min})/(N_y - 1)$. The grid spacing in the x-direction does not need to be equidistant, i.e., we write $(h_x)_i = x_{i+1} - x_i$. This adaptive step size will be detailed in Section 4.1.4. We denote the numerical approximation of u in a grid point as $u_{i,j} \approx u(\mathbf{x}_{i,j})$, and likewise for the other variables.

When discussing numerical methods we generally consider one step at a time, i.e., we consider the evolution from the grid line $x = x_i$ to the line $x = x_{i+1}$. Therefore it is convenient to write $h_x = (h_x)_i$ when no ambiguity arises.

4.1.1 Numerical method based on forward Euler

In this section we will introduce a numerical scheme based on the forward Euler method to calculate $\mathbf{v}^{\alpha}_{i+1,j}$ and $\mathbf{v}^{\beta}_{i+1,j}$ given $\mathbf{v}^{\alpha}_{i,j}$ and $\mathbf{v}^{\beta}_{i,j}$.

The numerical stencil is schematically shown in Figure 4.1. The black dots represent the grid points. The solid blue and dashed red arrows correspond to the numerical approximations of the α- and β-characteristics, respectively. At the grid points these characteristics are approximated using forward Euler, i.e., we approximate the characteristics as tangent lines passing through grid points $\mathbf{x}_{i,j}$ and having slope $a_{i,j}$ for the α-characteristic and $b_{i,j}$ for the β-characteristic. This implies that for a step size h_x, the two characteristics departing from $\mathbf{x}_{i,j}$ arrive at $(x_{i+1}, \tilde{y}^{\alpha}_{i+1}(j))$ and $(x_{i+1}, \tilde{y}^{\beta}_{i+1}(j))$ for the α- and β-characteristic

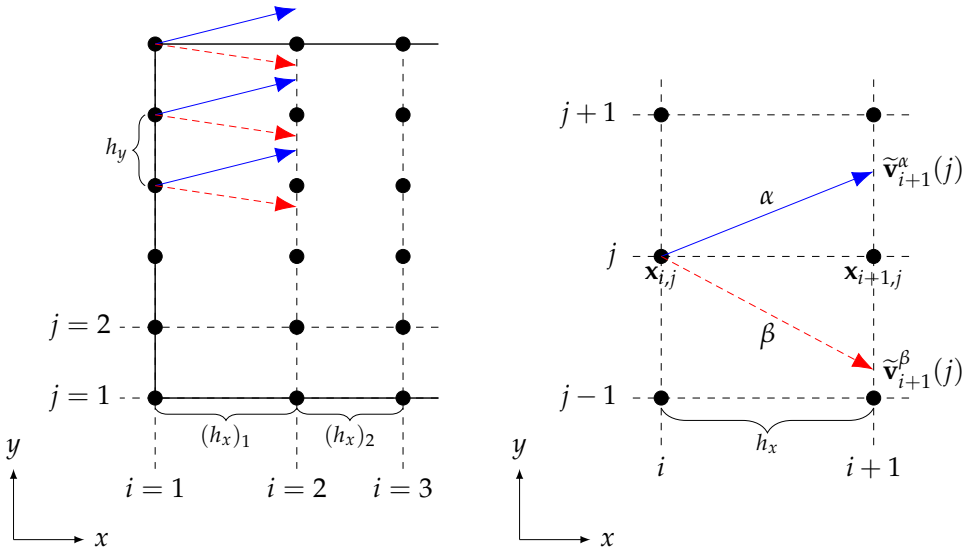


Figure 4.1: Schematic representation of the numerical method using forward Euler.

respectively, where

$$x_{i+1} = x_i + h_x, \tag{4.3a}$$
$$\tilde{y}^{\alpha}_{i+1}(j) = y_j + h_x a_{i,j}, \tag{4.3b}$$
$$\tilde{y}^{\beta}_{i+1}(j) = y_j + h_x b_{i,j}. \tag{4.3c}$$

The points $(x_{i+1}, \tilde{y}^{\alpha}_{i+1}(j))$ and $(x_{i+1}, \tilde{y}^{\beta}_{i+1}(j))$ do generally not coincide with any grid point. Similarly we can calculate u, p, q on both characteristics, while a and b can only be determined on one characteristic. More compactly written, we have the following forward Euler step:

$$\begin{aligned} \widetilde{\mathbf{v}}^{\alpha}_{i+1}(j) &= \mathbf{v}^{\alpha}_{i,j} + h_x \mathbf{g}^{\alpha}(\mathbf{v}^{\alpha}_{i,j}, a_{i,j}), \\ \widetilde{\mathbf{v}}^{\beta}_{i+1}(j) &= \mathbf{v}^{\beta}_{i,j} + h_x \mathbf{g}^{\beta}(\mathbf{v}^{\beta}_{i,j}, b_{i,j}). \end{aligned} \tag{4.4}$$

Here $\widetilde{\mathbf{v}}^{\alpha}_{i+1}(j)$ denotes the new values of $\mathbf{v}^{\alpha}$ at $(x_{i+1}, \tilde{y}^{\alpha}_{i+1}(j))$ for which the corresponding characteristic passes through the grid point $\mathbf{x}_{i,j}$ as shown in Figure 4.1. Analogously we define $\widetilde{\mathbf{v}}^{\beta}_{i+1}(j)$. Because we are interested in obtaining $\mathbf{v}^{\alpha}(x_{i+1}, y_j)$ and $\mathbf{v}^{\beta}(x_{i+1}, y_j)$, we interpolate $\widetilde{\mathbf{v}}^{\alpha}_{i+1}(j)$ and $\widetilde{\mathbf{v}}^{\beta}_{i+1}(j)$. Let the y, u, p, q-components of $\widetilde{\mathbf{v}}^{\alpha}_{i+1}(j)$ be denoted by $\tilde{y}^{\alpha}_{i+1}(j)$, $\tilde{u}^{\alpha}_{i+1}(j)$, $\tilde{p}^{\alpha}_{i+1}(j)$ and $\tilde{q}^{\alpha}_{i+1}(j)$, respectively, and similarly for $\widetilde{\mathbf{v}}^{\beta}_{i+1}(j)$. Two approaches to carry out the interpolation are shown in Figure 4.2, using linear interpolation.

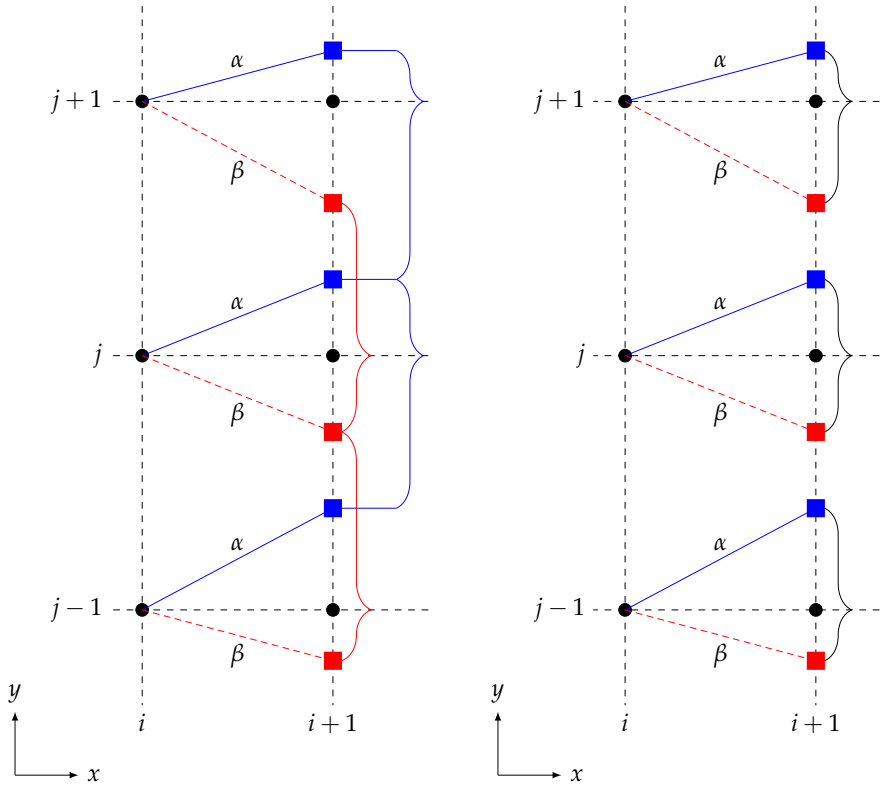


Figure 4.2: Schematic representation of linear interpolation where a curly bracket denotes the values used for interpolation and the grid point it influences. Interpolation for u, p and q can be done either using two distinct sets, formed by the two characteristic families (Approach 1, left), or using both sets combined (Approach 2, right).

Approach 1: We consider all values computed for each characteristic family, e.g. $\tilde{u}^{\alpha}_{i+1}(j)$, as one set and interpolate using local B-Splines ([41, p. 90-97]; see App. 4.1.1) within that set, which yields the approximations $u^{\alpha}_{i+1,j}$ and $u^{\beta}_{i+1,j}$. Because we have no a priori preference whether $u^{\alpha}_{i+1,j}$ or $u^{\beta}_{i+1,j}$ approximates $u_{i+1,j}$ better, we average the results and set

$$u_{i+1,j} = \frac{u^{\alpha}_{i+1,j} + u^{\beta}_{i+1,j}}{2}. \tag{4.5}$$

Similarly we obtain $p_{i+1,j}$ and $q_{i+1,j}$.

Approach 2: We collect $\tilde{u}^{\alpha}_{i+1}(j)$ and $\tilde{u}^{\beta}_{i+1}(j)$ for all $j = 1, \ldots, N_y$ in one set and interpolate using local B-Splines over that set to obtain $u_{i+1,j}$. Similarly we obtain $p_{i+1,j}$ and $q_{i+1,j}$.

Numerical results show that neither of these methods outperforms the other significantly or consistently. Computationally both approaches are approximately equally expensive [73]. Approach 1 will be used throughout.

Various subtleties arise using the approach above. For one, a and b are known along one family of characteristics only, so we use one set of values for interpolation of these variables. If a grid point is not located between two numerically estimated characteristics, which may occur for a grid point on the boundary, then we supplement the missing boundary value as detailed in Section 3.4.

To determine an appropriate spline interpolant we consider both the error associated with one forward Euler step and that of the interpolation. The local truncation error of the forward Euler method is $\mathcal{O}(h_x^2)$ [35, p. 335]. In Section 4.1.4 we introduce a method to control the step size h_x, such that the local truncation error behaves as $\mathcal{O}(h_y^2)$. The error of a spline interpolant of order n, also called a spline interpolant of degree $n-1$, is $\mathcal{O}(h_y^n)$ [41, p. 95]. In general we choose the degree of the spline interpolant such that its order matches the order of the integration method. For the forward Euler method this implies a spline interpolant of order 2 which is standard local linear interpolation.

4.1.2 Numerical method based on modified Euler

In this section we will discuss a numerical scheme based on the modified Euler scheme. The local truncation error of the modified Euler method is $\mathcal{O}(h_x^3)$.

First we calculate a predictor by doing a forward Euler step of step size $\frac{h_x}{2}$ for (4.2), viz.

$$\begin{aligned} \widetilde{\mathbf{v}}^{\alpha}_{i+\frac{1}{2}}(j) &= \mathbf{v}^{\alpha}_{i,j} + \frac{h_x}{2}\mathbf{g}^{\alpha}(\mathbf{v}^{\alpha}_{i,j}, a_{i,j}), \\ \widetilde{\mathbf{v}}^{\beta}_{i+\frac{1}{2}}(j) &= \mathbf{v}^{\beta}_{i,j} + \frac{h_x}{2}\mathbf{g}^{\beta}(\mathbf{v}^{\beta}_{i,j}, b_{i,j}). \end{aligned} \tag{4.6}$$

We adhere to the same notation as in the previous section, where a tilde denotes the function value on the corresponding characteristic which is not necessarily located at a grid point.

Because a and b are not known but are needed at $x_{i+1/2}$ along the α- and β-characteristic, respectively, we approximate them using spline interpolation. In general we choose third-order splines for this method. We interpolate the a-component of $\widetilde{\mathbf{v}}^{\beta}_{i+\frac{1}{2}}(j)$, known at the points $\widetilde{y}^{\beta}_{i+\frac{1}{2}}(j)$ to the points $\widetilde{y}^{\alpha}_{i+\frac{1}{2}}(j)$, which approximate $a(x_{i+\frac{1}{2}}, \widetilde{y}^{\alpha}_{i+\frac{1}{2}}(j))$. We denote this approximation by $a_{i+\frac{1}{2},j}$. In the same way we approximate $b_{i+\frac{1}{2},j}$ from $\widetilde{\mathbf{v}}^{\alpha}_{i+\frac{1}{2}}(j)$. The modified Euler step

is then given by

$$\begin{aligned}\widetilde{\mathbf{v}}^{\alpha}_{i+1}(j) &= \mathbf{v}^{\alpha}_{i,j} + h_x \mathbf{g}^{\alpha}(\widetilde{\mathbf{v}}^{\alpha}_{i+\frac{1}{2}}(j), a_{i+\frac{1}{2},j}),\\ \widetilde{\mathbf{v}}^{\beta}_{i+1}(j) &= \mathbf{v}^{\beta}_{i,j} + h_x \mathbf{g}^{\beta}(\widetilde{\mathbf{v}}^{\beta}_{i+\frac{1}{2}}(j), b_{i+\frac{1}{2},j}).\end{aligned} \tag{4.7}$$

We conclude the modified Euler step by interpolating to the grid points using Approach 1, as discussed in Section 4.1.1. We choose third-order spline interpolation for the modified Euler method, as it corresponds to its local truncation error.

4.1.3 Numerical method based on classic Runge-Kutta

In this section we will introduce the classic Runge-Kutta method. Following a similar approach, we can generalize our integration methods to other higher-order explicit Runge-Kutta methods.

First we make a forward Euler step of size $\frac{h_x}{2}$ for (4.2) viz.

$$\begin{aligned}\widetilde{\mathbf{v}}^{\alpha}_{i+\frac{1}{2}}(j) &= \mathbf{v}^{\alpha}_{i,j} + \frac{h_x}{2}\mathbf{g}^{\alpha}(\mathbf{v}^{\alpha}_{i,j}, a_{i,j}),\\ \widetilde{\mathbf{v}}^{\beta}_{i+\frac{1}{2}}(j) &= \mathbf{v}^{\beta}_{i,j} + \frac{h_x}{2}\mathbf{g}^{\beta}(\mathbf{v}^{\beta}_{i,j}, b_{i,j}).\end{aligned} \tag{4.8}$$

As in the case of the modified Euler based method, we interpolate the a- and b-components of $\widetilde{\mathbf{v}}^{\beta}_{i+\frac{1}{2}}(j)$ and $\widetilde{\mathbf{v}}^{\alpha}_{i+\frac{1}{2}}(j)$ to approximate $a_{i+\frac{1}{2},j}$ and $b_{i+\frac{1}{2},j}$, respectively. In general we use fifth-order splines for this method. Second, we do a step of size $\frac{h_x}{2}$ with the slope based on the previously found values of $\widetilde{\mathbf{v}}^{\alpha}$ and $\widetilde{\mathbf{v}}^{\beta}$, viz.

$$\begin{aligned}\widehat{\mathbf{v}}^{\alpha}_{i+\frac{1}{2}}(j) &= \mathbf{v}^{\alpha}_{i,j} + \frac{h_x}{2}\mathbf{g}^{\alpha}(\widetilde{\mathbf{v}}^{\alpha}_{i+\frac{1}{2}}(j), a_{i+\frac{1}{2},j}),\\ \widehat{\mathbf{v}}^{\beta}_{i+\frac{1}{2}}(j) &= \mathbf{v}^{\beta}_{i,j} + \frac{h_x}{2}\mathbf{g}^{\beta}(\widetilde{\mathbf{v}}^{\beta}_{i+\frac{1}{2}}(j), b_{i+\frac{1}{2},j}),\end{aligned} \tag{4.9}$$

where we used a hat to distinguish between the different stages. Similar as before we interpolate a and b from the β- and α-characteristics to the α- and β-characteristics to obtain $\hat{a}_{i+\frac{1}{2},j}$ and $\hat{b}_{i+\frac{1}{2},j}$, respectively. Using these slopes we do a step of size h_x which yields

$$\begin{aligned}\widehat{\mathbf{v}}^{\alpha}_{i+1}(j) &= \mathbf{v}^{\alpha}_{i,j} + h_x \mathbf{g}^{\alpha}(\widehat{\mathbf{v}}^{\alpha}_{i+\frac{1}{2}}(j), \hat{a}_{i+\frac{1}{2},j}),\\ \widehat{\mathbf{v}}^{\beta}_{i+1}(j) &= \mathbf{v}^{\beta}_{i,j} + h_x \mathbf{g}^{\beta}(\widehat{\mathbf{v}}^{\beta}_{i+\frac{1}{2}}(j), \hat{b}_{i+\frac{1}{2},j}).\end{aligned} \tag{4.10}$$

Interpolating the a- and b-components of $\widehat{\mathbf{v}}^{\beta}_{i+1}(j)$ and $\widehat{\mathbf{v}}^{\alpha}_{i+1}(j)$ yields approximations for $\hat{a}^{\alpha}_{i+1}(j)$ and $\hat{b}^{\beta}_{i+1}(j)$, respectively. Finally the full Runge-Kutta step is given by

$$\begin{aligned}\widetilde{\mathbf{v}}^{\alpha}_{i+1}(j) = \mathbf{v}^{\alpha}_{i,j} + \frac{h_x}{6}\Big(&\mathbf{g}^{\alpha}(\mathbf{v}^{\alpha}_{i,j}, a_{i,j}) + 2\mathbf{g}^{\alpha}(\widetilde{\mathbf{v}}^{\alpha}_{i+\frac{1}{2}}(j), a_{i+\frac{1}{2},j}) + \\ &2\mathbf{g}^{\alpha}(\widehat{\mathbf{v}}^{\alpha}_{i+\frac{1}{2}}(j), \hat{a}_{i+\frac{1}{2},j}) + \mathbf{g}^{\alpha}(\widehat{\mathbf{v}}^{\alpha}_{i+1}(j), \hat{a}^{\alpha}_{i+1}(j))\Big),\end{aligned} \tag{4.11a}$$

$$\begin{aligned}\widetilde{\mathbf{v}}^{\beta}_{i+1}(j) = \mathbf{v}^{\beta}_{i,j} + \frac{h_x}{6}\Big(&\mathbf{g}^{\beta}(\mathbf{v}^{\beta}_{i,j}, b_{i,j}) + 2\mathbf{g}^{\beta}(\widetilde{\mathbf{v}}^{\beta}_{i+\frac{1}{2}}(j), b_{i+\frac{1}{2},j}) + \\ &2\mathbf{g}^{\beta}(\widehat{\mathbf{v}}^{\beta}_{i+\frac{1}{2}}(j), \hat{b}_{i+\frac{1}{2},j}) + \mathbf{g}^{\beta}(\widehat{\mathbf{v}}^{\beta}_{i+1}(j), \hat{b}^{\beta}_{i+1}(j))\Big).\end{aligned} \tag{4.11b}$$

We conclude the Runge-Kutta step by interpolating $\widetilde{\mathbf{v}}^{\alpha}_{i+1}(j)$ and $\widetilde{\mathbf{v}}^{\beta}_{i+1}(j)$ using fifth-order spline interpolation to the grid points using Approach 1, as discussed in Section 4.1.1.

4.1.4 Adaptive step size control

In this section we introduce a procedure to choose the step size h_x adaptively. We aim to reduce the computational error while maintaining convergence, which depends both on the integration method and the interpolation methods.

Because integration is done in the positive x-direction, the corresponding discretization error is a function of h_x. On the other hand, we interpolate in the y-direction, but the interpolation error is not solely a function of h_y as we will see. Ideally we want both the integration and interpolation error to be of the same order, such that neither of them dominates. Asymptotically, this is obtained most easily by using both an integration and interpolation method of the same order, and choosing the discretization steps in the x- and y-direction of the same order of magnitude. In the x-direction the discretization step size is h_x, though the distance between the interpolation nodes, those points that are the intersection between a grid line and the numerical approximation of the characteristics, is not equidistant, but rather follows from both the evolution along the characteristics (3.72) and the integration method used. Without loss of generality, we solely consider the β-characteristic. Let $\Delta y(j) = |\tilde{y}^{\beta}_{i+1}(j) - y_j|$ denote the distance between, on the one hand, the intersection point of (the approximation of) the β-characteristic at $x = x_{i+1}$, and on the other hand, the point (x_{i+1}, y_j), as shown in Figure 4.3 for $j = 1$. Approximating $\Delta y(j)$ by a forward Euler step yields

$$\Delta y(j) = |y_j + (h_x)_i\, b_{i,j} - y_j| = |b_{i,j}|(h_x)_i. \tag{4.12}$$

Hence $\Delta y(j) \leq h_y$ is obtained if we choose

$$(h_x)_i < h_y \min_{j\in\{1,\ldots,N_y\}} (1, 1/|b_{i,j}|), \tag{4.13}$$

where the constant "1" is chosen such that $(h_x)_i < h_y$, regardless of the slope of the characteristics. This allows us to control the error of the numerical methods by solely controlling h_y. Similarly we would like to have $(h_x)_i < h_y \min(1, 1/|a_{i,j}|)$ for all $j = 1, \ldots, N_y$. Therefore we choose

$$(h_x)_i = \gamma\, h_y \cdot \min_{j\in\{1,\ldots,N_y\}} \left\{1, \left|\frac{1}{a_{i,j}}\right|, \left|\frac{1}{b_{i,j}}\right|\right\}, \tag{4.14}$$

where $0 \leq \gamma \leq 1$ is a tuning parameter. Generally we choose $\gamma = 0.95$, as this implies strict inequality.

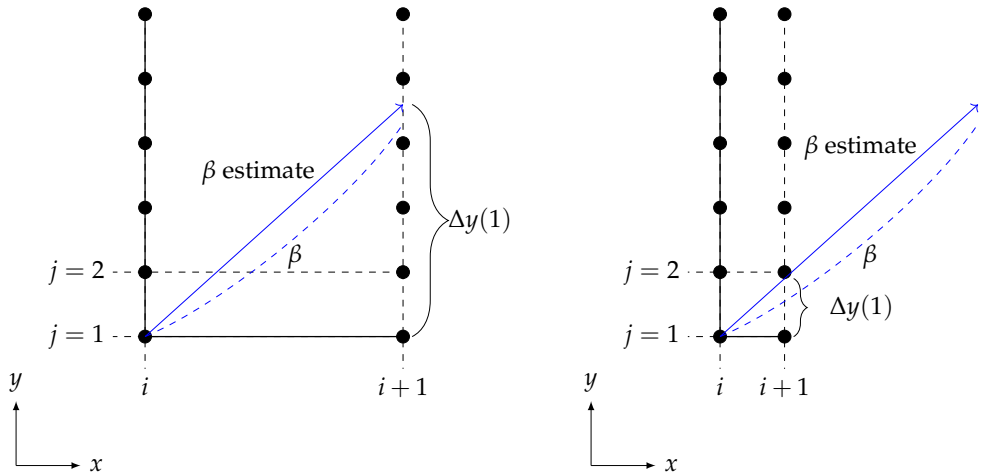


Figure 4.3: Two schematic situations of Δy, as function of $(h_x)_i$.

4.1.5 Residual of the Monge-Ampère equation

We solve the mutually coupled ODE systems (3.72) without calculating r, s or t, nor by calculating any numerical derivatives. As discontinuities may arise in the solution, as will be shown in Section 4.2.5, numerical errors of (standard) numerical differentiation methods are a priori unknown. Furthermore, due to the adaptive step size, we have a non-uniform grid in the x-direction for which we require non-standard numerical differentiation methods. These methods are themselves often prone to errors (especially for higher-order schemes). For example, inversion of the Vandermonde matrix for the calculation of

finite-difference weigths, may be ill-conditioned as the matrix can be close to singular [84]. Addressing such methods will overly complicate the discussion. We therefore introduce a method which does not require differentiations in order to measure the residual. For this we rely on the integral formulation of the Monge-Ampère equation as given in Section 2.4, viz.

$$\iint_A f^2 \,\mathrm{d}A = -\oint_{\partial A} p\nabla q \cdot \hat{\boldsymbol{\tau}} \,\mathrm{d}s, \tag{4.15a}$$

$$\iint_A f^2 \,\mathrm{d}A = \oint_{\partial A} q\nabla p \cdot \hat{\boldsymbol{\tau}} \,\mathrm{d}s. \tag{4.15b}$$

Because the integral formulations (4.15) are both equivalent to the Monge-Ampère equation, we use numerical approximations of (4.15) as a measure for the residual of the numerical solution to the Monge-Ampère equation. To this purpose, let

$$\mathbf{H}_1 = -p\nabla q = -p\begin{pmatrix} s \\ t \end{pmatrix} = \frac{pf}{a-b}\begin{pmatrix} a+b \\ -2 \end{pmatrix}, \tag{4.16a}$$

$$\mathbf{H}_2 = \quad q\nabla p = \quad q\begin{pmatrix} r \\ s \end{pmatrix} = \frac{qf}{a-b}\begin{pmatrix} 2ab \\ -a-b \end{pmatrix}. \tag{4.16b}$$

Equations (2.45) are equivalent to

$$\iint_A f^2 \,\mathrm{d}A = \oint_{\partial A} \mathbf{H}_k \cdot \hat{\boldsymbol{\tau}} \,\mathrm{d}s, \tag{4.17}$$

for $k = 1, 2$. Choosing an appropriate domain A, this can therefore be used to determine the numerical residual. We choose the control volume $A = A_{i,j} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$, and write (4.17) as

$$I_k^N + I_k^S + I_k^W + I_k^E - \iint_{A_{i,j}} f^2 \,\mathrm{d}A = 0, \tag{4.18}$$

where

$$\begin{aligned} I_k^N &= -\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} H_{k,x}(x, y_{j+\frac{1}{2}}) \,\mathrm{d}x, & I_k^S &= \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} H_{k,x}(x, y_{j-\frac{1}{2}}) \,\mathrm{d}x, \\ I_k^W &= -\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} H_{k,y}(x_{i-\frac{1}{2}}, y) \,\mathrm{d}y, & I_k^E &= \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} H_{k,y}(x_{i+\frac{1}{2}}, y) \,\mathrm{d}y, \end{aligned} \tag{4.19}$$

and $H_{k,x}$ denotes the x-component of $\mathbf{H}_k$, and the line integrals are carried out over the North, South, West and East part of the control volume $A_{i,j}$, as shown in Figure 4.4.

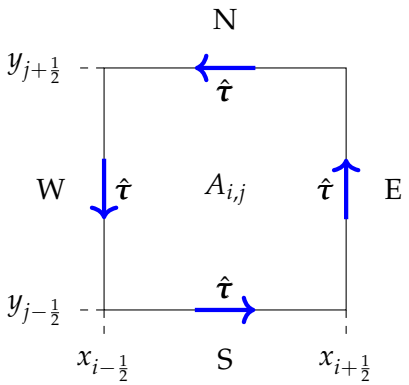


Figure 4.4: Schematic overview of domain $A_{i,j}$ with corresponding tangent vectors.

We approximate the integrals I_k^N, I_k^S, I_k^W and I_k^E using Gauss-Legendre quadrature rules. To this end, let $z_1, z_2 \in \mathbb{R}$, $z_1 < z_2$ and $g \in C([z_1, z_2])$. Given the n points $\xi_i \in [-1, 1]$ and n weights w_i (see [70] for example), we approximate

$$\int_{z_1}^{z_2} g(x)\,\mathrm{d}x \approx \frac{z_2 - z_1}{2} \sum_{i=1}^{n} w_i g\Big(\frac{z_2 - z_1}{2}\xi_i + \frac{z_1 + z_2}{2}\Big). \tag{4.20}$$

In case g is $2n$ times continuously differentiable, the error associated with (4.20) is $\mathcal{O}((z_2 - z_1)^{2n+1})$ [46, §5.2]. Therefore, we choose $n = 3$ such that the calculation of the residual is at least sixth order, one order higher than the classic Runge-Kutta scheme. Note that we generally do not know $\mathbf{H}_k$ in the points $(\frac{z_2-z_1}{2}\xi_i + \frac{z_1+z_2}{2})$, where either $(z_1, z_2) = (x_{i-1/2}, x_{i+1/2})$ or $(z_1, z_2) = (y_{i-1/2}, y_{i+1/2})$. Therefore, we interpolate $\mathbf{H}_k$ using splines of order 5, such that the interpolation is at least as accurate as the one-step method. Similarly, we approximate $\iint_{A_{i,j}} f^2\,\mathrm{d}A$ by subsequently integrating over x and y using the Gauss-Legendre quadrature rule. We normalize the absolute residual of (4.18) over $A_{i,j}$ by dividing it by the area $|A_{i,j}|$, and denote the result by $\epsilon_k(i, j)$. Lastly, we measure the residual over the whole grid by

$$\epsilon_k = \max_{\substack{i\in\{2,\dots,N_x-1\}\\ j\in\{2,\dots,N_y-1\}}} |\epsilon_k(i, j)|. \tag{4.21}$$

4.2 Numerical results for the standard MA equation

In this section we present numerical results for the Monge-Ampère equation. We will present results for the forward Euler, modified Euler and classic Runge-

Kutta based methods for a default test case (Section 4.2.1), an example with known analytical solution. For the modified Euler method we furthermore consider 2nd, 3rd and 5th order splines. In Section 4.2.2, we compare the methods for a second example. Additionally, we show numerical results for the Runge-Kutta based method, for which we prescribe either two or zero boundary conditions per boundary segment (Section 4.2.3), a case where the number of boundary conditions varies along the boundary (Section 4.2.4) and a case where one boundary value is nonsmooth (Section 4.2.5).

4.2.1 A smooth default test case

To validate the numerical methods we design a default test case. To this end we let $\Omega = [0,1] \times [-0.5, 0.5]$ and we calculate a right-hand side-solution pair (f,u) using the method outlined in Section 2.3 with $w(z) = \cos(iz)$, giving

$$\begin{aligned} u(x,y) &= \cos(y)\cosh(x), \\ f(x,y) &= \sqrt{\frac{\cos(2y)+\cosh(2x)}{2}}. \end{aligned} \tag{4.22}$$

From the exact solution (4.22) and (3.64) we find a and b on the whole domain, viz.

$$a(x,y) = -\frac{\sin(y)\sinh(x) + f(x,y)}{\cos(y)\cosh(x)}, \tag{4.23a}$$

$$b(x,y) = \frac{-\sin(y)\sinh(x) + f(x,y)}{\cos(y)\cosh(x)}. \tag{4.23b}$$

We impose the corresponding initial conditions

$$u(0,y) = \cos(y), \qquad p(0,y) = 0. \tag{4.24}$$

By (3.73) we find

$$\begin{aligned} q(0,y) &= -\sin(y), & t(0,y) &= -\cos(y), & s(0,y) &= 0, \\ a(0,y) &= -1, & b(0,y) &= 1, & r(0,y) &= \cos(y). \end{aligned} \tag{4.25}$$

To justify (4.24), note that the outward unit normal vector $\hat{\mathbf{n}}$ on the initial strip is $\hat{\mathbf{n}}(0,y) = (-1,0)^{\mathrm{T}}$, $\mathbf{x}_\alpha(0,y) = (1,-1)^{\mathrm{T}}$ and $\mathbf{x}_\beta(0,y) = (1,1)^{\mathrm{T}}$. Therefore $\mathbf{x}_\alpha(0,y)\cdot\hat{\mathbf{n}}(0,y) = \mathbf{x}_\beta(0,y)\cdot\hat{\mathbf{n}}(0,y) = -1$, and hence we should prescribe u and p on the initial strip according to Case 2 in Section 3.4. A similar calculation shows that at $x = 1$ no boundary conditions need to be prescribed.

On the upper boundary, we have $\hat{\mathbf{n}} = (0,1)^{\mathrm{T}}$, $\mathbf{x}_\alpha = (1,a)^{\mathrm{T}}$, $\mathbf{x}_\beta = (1,b)^{\mathrm{T}}$, $a < 0$ and $b > 0$, so that the α-characteristic is entering the domain. Hence we need to prescribe b and a is known. In the same way, we need to prescribe a at the lower boundary and b is known.

4.2.1.1 Forward Euler based method

We present the results for the forward Euler based method for which we use splines of first degree, i.e., linear interpolants. The convergence of the forward Euler scheme is shown in Figure 4.5 as function of h_y, which controls $(h_x)_i$; see Section 4.1.4. In the left figure the maximum absolute differences at $x = 1$ between the function value and its numerical approximation for several variables are shown. More precisely, the figure shows

$$E[a] := \max_{j=1,\dots,N_y} \tfrac{1}{N_y} \left| a(x_{N_x}, y_j) - a_{N_x,j} \right|, \tag{4.26}$$

and similar errors for b, u, p and q for varying h_y. The adaptive step size control implies $(h_x)_i \approx \mathcal{O}(h_y)$, which allows us to quantify the error solely in terms of h_y.

It is well known that the forward Euler method is locally second order and globally first order accurate, if the solutions are sufficiently smooth. Because the interpolation error and the local discretization error are both second-order accurate, we expect the global convergence to be that of the forward Euler method, i.e., first order, which is also seen in the figure.

The residuals are also shown and show first-order convergence for ϵ_1, ϵ_2, as defined by (4.21). To understand this, note that we divide by the area of the control volume, that is, we divide by $|A_{i,j}|$, effectively normalizing ϵ_1, ϵ_2 such that they converge as the integrand does, which in this case is first order (for $\mathbf{H}_1$ and $\mathbf{H}_2$).

Figure 4.6 shows the solution surface u (left), and a color map of the residual ϵ_1 (right) for the case $N_y = 1000$. The surface u is clearly both smooth and a saddle surface. The right image shows the residual along with some characteristics. The shown characteristics are calculated after the simulation is done, and chosen such that they enter at seven equidistant points on the initial strip and five on the upper and lower boundary. The direction of the characteristics clearly shows that both the blue characteristic, i.e., the α-characteristic, and the black characteristic, i.e., the β-characteristic, enter at the initial strip. Hence, both u and p need to be prescribed at the initial strip. This is in agreement with (4.24). Furthermore, it shows that the α-characteristics and β-characteristics leave the domain at the lower and upper boundary, respectively. Therefore a and b should be prescribed at the

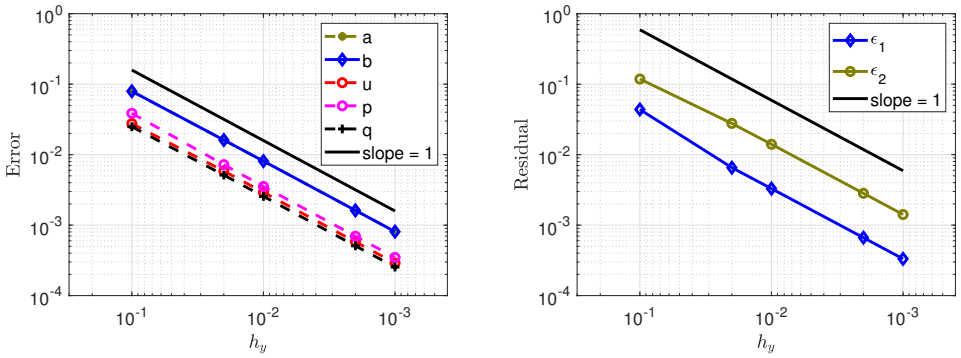


Figure 4.5: Convergence of the global error (left) and the residual (right) for the forward Euler based method.

lower and upper boundary, respectively, which agrees with the discussion on boundary conditions above.

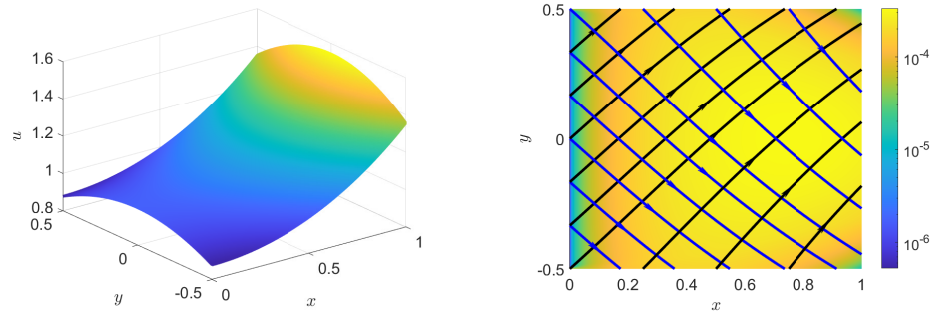


Figure 4.6: Surface u (left) and a color map of the residual ϵ_1 with characteristics (right).

4.2.1.2 Modified Euler based method

In this section we will discuss a few results for the modified Euler based method. We demonstrate the importance of choosing an appropriate interpolation routine and show accompanying convergence results. Generally we use an interpolant which is as accurate as the integration routine because a more accurate interpolant will not increase the overall accuracy while being computationally more expensive, and a lower-order interpolant will lower the convergence. In Figures 4.7, 4.8 and 4.9 the convergence is shown for splines of order 2, 3, and 5, respectively. Using splines of second order yields first-order convergence for both the global error and the residual. This is what we expect because the local discretization error after interpolation is

second-order. Henceforth, second-order splines, i.e., linear B-splines, reduce the rate of convergence, and higher-order splines are preferred.

Figures 4.8 and 4.9 show the same order of convergence because the interpolants are at least as accurate as the local integration error. For a spline of order 3, i.e., quadratic B-splines, the interpolation error is as accurate as the local error of the modified Euler method. The accuracy of the global error is one order lower than the local error and equal to that of the residual. For the spline of order 5, i.e., for polynomials of fourth degree as B-splines, the local error of the modified Euler method is the limiting factor and the global error and the residual are second-order accurate.

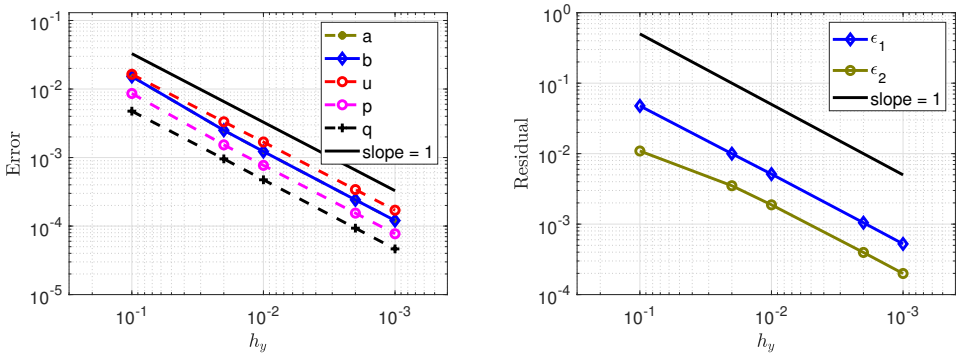


Figure 4.7: Convergence for the global error (left) and the residual (right) for the modified Euler based method with a second-order accurate interpolant.

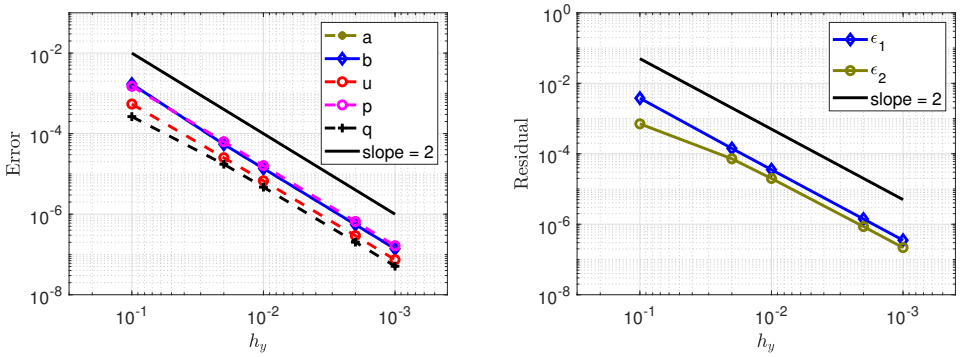


Figure 4.8: Convergence for the global error (left) and the residual (right) for the modified Euler based method with a third-order accurate interpolant.

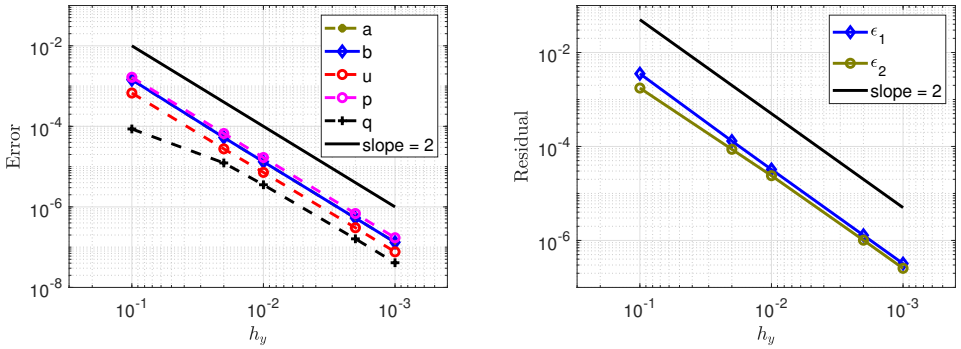


Figure 4.9: Convergence for the global error (left) and the residual (right) for the modified Euler based method with a fifth-order accurate interpolant.

4.2.1.3 Classic Runge-Kutta based method

Analogously to the previous sections, we consider the default test case. Figure 4.10 shows the results for using the Runge-Kutta method with spline interpolants of order 5. Because the Runge-Kutta method is locally fifth-order accurate, which coincides with the accuracy of the splines, we expect a fourth-order global convergence. This is indeed shown in the figure. The convergence of the residuals is also expected to be of order 4, as also seen in the figure. The convergence seems to slow down for $h_y \approx 1/1000$, which is due to round-off errors as the solutions reach computer precision.

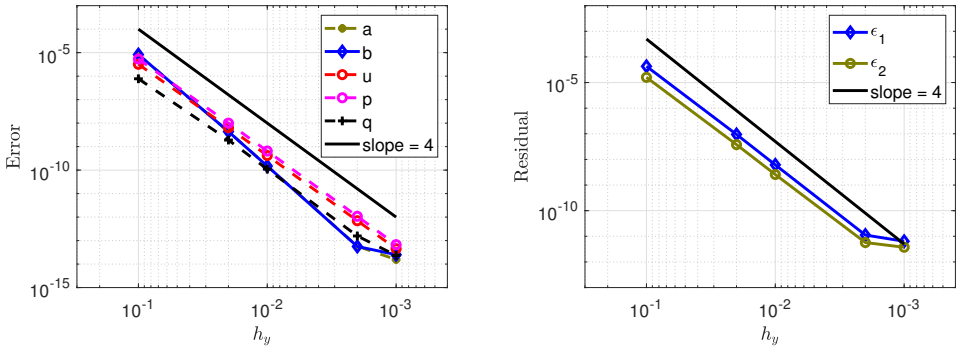


Figure 4.10: Convergence for the global error (left) and the residual (right) for the classic Runge-Kutta based method with a fifth-order accurate interpolant.

4.2.2 An aggregated example

Next we compare the developed numerical methods for the example

$$u(x,y) = e^x \cos(y), \qquad f(x,y) = e^x, \tag{4.27}$$

which is constructed using $w(z) = e^z$; see Appendix 2.3. Let $\Omega = [0,2] \times [-1/3, 2/3]$ be the computational domain. A straightforward calculation using (4.27) shows

$$\begin{aligned} p(x,y) &= e^x \cos(y), & q(x,y) &= -e^x \sin(y) \\ a(x,y) &= -\frac{\sin(y)+1}{\cos(y)}, & b(x,y) &= \frac{-\sin(y)+1}{\cos(y)}, \end{aligned} \tag{4.28}$$

which we use, along with (4.27), to prescribe u, p, q, a, b on the initial strip $x = 0$, $-\frac{1}{3} \leq y \leq \frac{2}{3}$, accordingly. Equations (4.28) show $a < 0$, $b > 0$ on Ω. Analogously to the default test case, we prescribe a on the lower and b on the upper boundary as dictated by (4.28). We compare the forward Euler, modified Euler and Runge-Kutta methods using second-, third- and fifth-order splines, respectively. The results are shown in Figure 4.11. The left figure shows first-, second- and fourth-order convergence of the global error of u for the forward Euler, modified Euler and Runge-Kutta-method, respectively. These rates of convergence are in agreement with the previous sections. The convergence of ϵ_1, shown in the right figure, also shows first, second and fourth order convergence. Figure 4.12 shows that the error accumulates for increasing x, i.e., the further into the domain, as measured from the initial strip, the higher the error. This is due to the errors propagating along characteristics. Furthermore, at each grid line new numerical errors originate by the chosen explicit one step method and the B-spline interpolation.

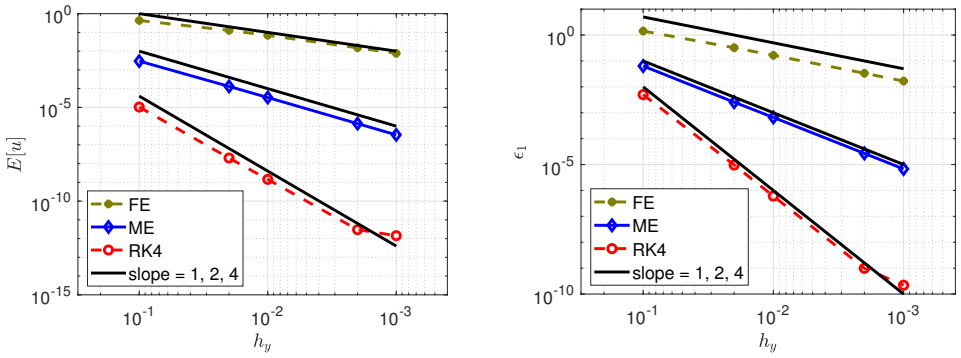


Figure 4.11: Convergence for the global error (left) and the residual (right).

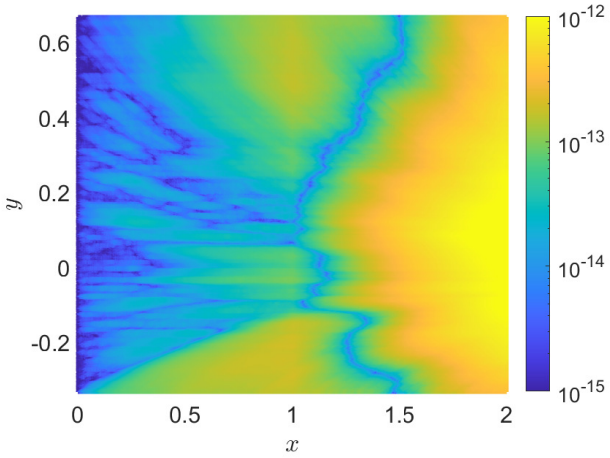


Figure 4.12: The error $|u_{i,j} - u(\mathbf{x}_{i,j})|$ over the domain for the Runge-Kutta method with $h_y = 10^{-3}$.

4.2.3 An initial strip extended over two edges

We will demonstrate an example for the forward Euler based method for which we have two entering characteristics at both the western and northern boundary, and two leaving characteristics on both the southern and eastern boundary. In this case the northern boundary is also an initial strip as discussed in Section 3.4. To this end let

$$u(x,y) = x^3y^2 + 1, \qquad f(x,y) = 2\sqrt{6}x^2y, \tag{4.29}$$

on the domain $\Omega = [1,2]^2$. A straightforward calculation shows that (4.29)

satisfies the Monge-Ampère equation (3.60), and that

$$a(x,y) = \frac{(-3+\sqrt{6})y}{x}, \qquad b(x,y) = -\frac{(3+\sqrt{6})y}{x}, \tag{4.30}$$

which implies $a, b < 0$ on Ω. Let $\hat{\mathbf{n}}_{\mathrm{W}}$ denote the normal at the western boundary segment, and likewise for the other subscripts. It follows that

$$\mathbf{x}_\alpha(1,y)\cdot\hat{\mathbf{n}}_{\mathrm{W}} < 0, \qquad \mathbf{x}_\beta(1,y)\cdot\hat{\mathbf{n}}_{\mathrm{W}} < 0, \tag{4.31a}$$
$$\mathbf{x}_\alpha(x,1)\cdot\hat{\mathbf{n}}_{\mathrm{S}} > 0, \qquad \mathbf{x}_\beta(x,1)\cdot\hat{\mathbf{n}}_{\mathrm{S}} > 0, \tag{4.31b}$$
$$\mathbf{x}_\alpha(2,y)\cdot\hat{\mathbf{n}}_{\mathrm{E}} > 0, \qquad \mathbf{x}_\beta(2,y)\cdot\hat{\mathbf{n}}_{\mathrm{E}} > 0, \tag{4.31c}$$
$$\mathbf{x}_\alpha(x,2)\cdot\hat{\mathbf{n}}_{\mathrm{N}} < 0, \qquad \mathbf{x}_\beta(x,2)\cdot\hat{\mathbf{n}}_{\mathrm{N}} < 0. \tag{4.31d}$$

The classification of the boundary conditions in Section 3.4 implies we should prescribe two boundary conditions at the western and northern boundaries and zero boundary conditions at the eastern and southern segments. By prescribing two boundary conditions at the northern boundary, it is an initial strip. The total set of prescribed boundary conditions for $1 \le x \le 2, 1 \le y \le 2$ thus read

$$u(1,y) = y^2 + 1, \quad p(1,y) = 3y^2, \quad u(x,2) = 4x^3 + 1, \quad q(x,2) = 4x^3. \tag{4.32}$$

From $u(1,y)$ and $p(1,y)$ we obtain q, r, s, t, a, b at the western boundary by (3.73). Analogously, $u(x,2)$ and $q(x,2)$ determine p, a, b, r, s, t at the northern boundary.

Figure 4.13 shows some characteristics for this example, illustrating where characteristics are entering or leaving the domain. Furthermore, the figure shows that Ω is fully covered by the characteristic domain of the two initial strips. The figure on the right also shows the points $(1.7, 1.6)$ and $(1.1, 1.1)$ with their corresponding domains of dependence colored red.

Figure 4.14 shows the convergence for the forward Euler based method for this example, for both the global error and the residual. As expected, both show first-order convergence. Figure 4.15 shows the error in b for this example, calculated using the Runge-Kutta method with fifth-order-splines with $h_y = 1/1000$. The figure shows the accumulation of numerical errors over the domain and shows b is most accurate near the boundaries where both a and b are prescribed.

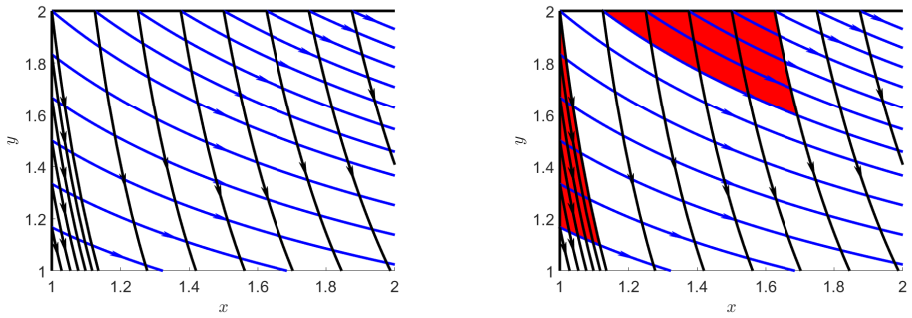


Figure 4.13: Domain Ω and some characteristics, which enter the domain at the western and northern boundaries, with the domain of dependence for the points $(1.7, 1.6)$ and $(1.1, 1.1)$.

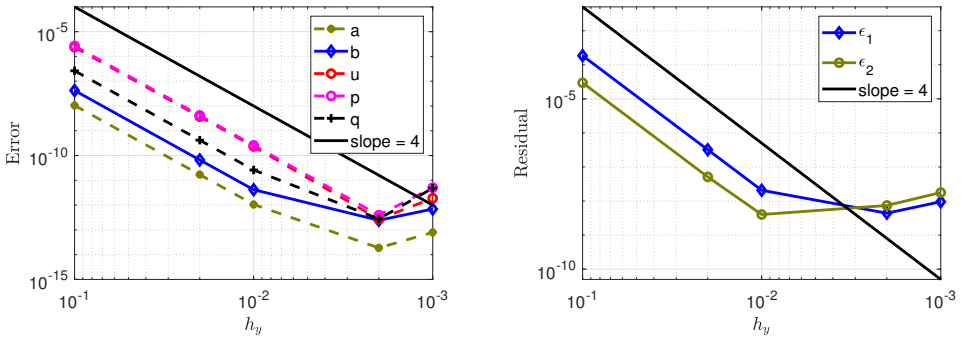


Figure 4.14: Convergence for the global error (left) and the residual (right) for the forward Euler based method with a second-order accurate interpolant.

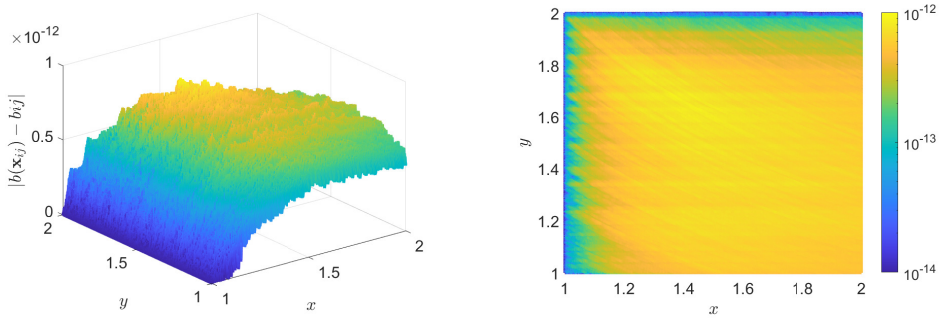


Figure 4.15: The numerical error $|b_{i,j} - b(\mathbf{x}_{i,j})|$.

4.2.4 Varying number of boundary conditions

Next we will show an example for which the number of boundary conditions we prescribe changes along the boundary. To this end let $\Omega = [1, 2.5] \times [-2, -1.5]$ and

$$u(x,y) = 1 + e^{2y/x}, \qquad f(x,y) = \frac{2}{x^2} e^{2y/x}. \tag{4.33}$$

It follows that

$$a(x,y) = 1 + \frac{y}{x}, \qquad b(x,y) = \frac{y}{x}, \tag{4.34}$$

which implies $a \leq 0$ for $x \leq -y$ on Ω and $a > 0$ for $x > -y$ and $b < 0$ on Ω. Let $\hat{\mathbf{n}}$ be the outward unit normal on $\partial\Omega$, then the boundary conditions to be prescribed are

Boundary segment	Classification	Boundary condition(s)
Western,	$(\mathbf{x}_\alpha, \hat{\mathbf{n}}) < 0, \ (\mathbf{x}_\beta, \hat{\mathbf{n}}) < 0,$	a and b,
Southern, $x \leq y$,	$(\mathbf{x}_\alpha, \hat{\mathbf{n}}) > 0, \ (\mathbf{x}_\beta, \hat{\mathbf{n}}) > 0,$	None,
Southern, $x > -y$,	$(\mathbf{x}_\alpha, \hat{\mathbf{n}}) < 0, \ (\mathbf{x}_\beta, \hat{\mathbf{n}}) > 0,$	b,
Northern, $x \leq y$,	$(\mathbf{x}_\alpha, \hat{\mathbf{n}}) < 0, \ (\mathbf{x}_\beta, \hat{\mathbf{n}}) < 0,$	a and b,
Northern, $x > -y$,	$(\mathbf{x}_\alpha, \hat{\mathbf{n}}) > 0, \ (\mathbf{x}_\beta, \hat{\mathbf{n}}) < 0,$	a,
Eastern,	$(\mathbf{x}_\alpha, \hat{\mathbf{n}}) > 0, \ (\mathbf{x}_\beta, \hat{\mathbf{n}}) > 0,$	None,

as illustrated in Figure 4.16.

Figure 4.17 shows the convergence of the global error and the residual for the Runge-Kutta based method with fifth-order splines. The convergence is fourth order as expected, and slowly comes to a halt for a fine grid, as also discussed in the previous section. Figure 4.18 shows the characteristics in the domain (left), and a heat map of the error $|b_{i,j} - b(\mathbf{x}_{i,j})|$ (right). The heat map clearly shows the swirling influence of the α-characteristics, and a lower error near the segments of the boundary where b is prescribed.

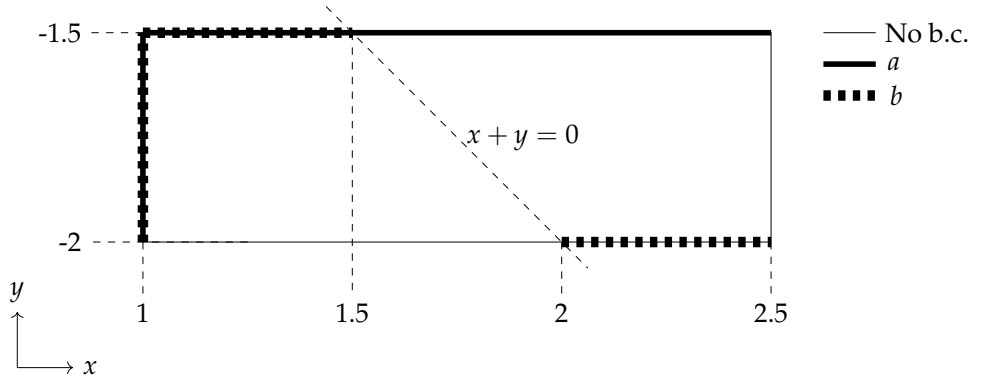


Figure 4.16: Schematic overview of the prescribed boundary conditions and their locations.

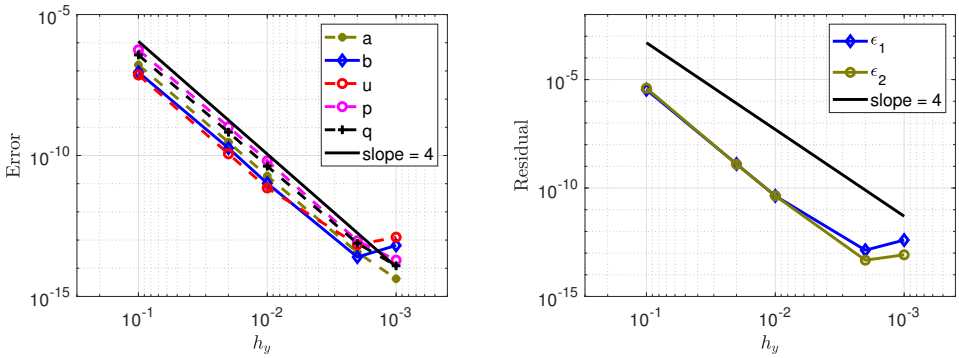


Figure 4.17: Convergence of the global error (left) and the residual (right).

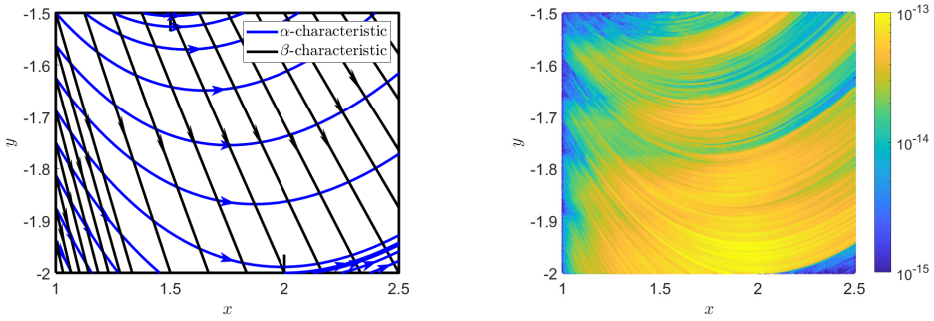


Figure 4.18: Characteristics in the domain (left) and a heat map of the numerical error in b (right).

4.2.5 An example with nonsmooth boundary conditions

The last example is similar to the default test case of Section 4.2.2, but now with nonsmooth boundary conditions. We set $\Omega = [-1, 1] \times [-0.5, 0.5]$. We prescribe the initial conditions

$$u_{\mathrm{W}}(y) = p_{\mathrm{W}}(y) = e^{-1}\cos(y), \tag{4.35}$$

from which the remaining relevant variables follow by applying (3.73). At the lower and upper boundary we prescibe the boundary conditions

$$a_{\mathrm{S}}(x) = \begin{cases} \frac{\sin(0.5)-1}{\cos(0.5)}, & x < 0, \\ -e^{-3x/2}(x^2+1), & x \geq 0, \end{cases} \tag{4.36a}$$

$$b_{\mathrm{N}}(x) = \frac{1-\sin(0.5)}{\cos(0.5)}, \tag{4.36b}$$

respectively. The boundary condition for a is not continuous, i.e., $\lim_{x\uparrow 0} a_{\mathrm{S}}(x) \approx -0.59$ and $\lim_{x\downarrow 0} a_{\mathrm{S}}(x) = -1$.

Generally, the error terms of numerical methods depend on derivatives of the functions to be estimated. As a is nonsmooth, we do not necessarily expect convergence as we did before. Furthermore, no analytical solution is known (for the whole domain) for this particular example, therefore we base convergence on the residual values.

We use the Runge-Kutta method with fifth-order splines for this example. Figure 4.19 shows a heat map of ϵ_1, i.e., the magnitude of the residual on a color scale on the left. The heat map and surface plots in this section are constructed for $N_y = 1001$. The figure shows convergence of the solution, although at a slower rate than for continuous boundary values. The heat map also shows a few characteristics given by the solid green and dashed white curves. Furthermore, it shows that the discontinuity of the derivatives of a in $(0, -0.5)$ yields a rapidly varying residual. This difference in residual is propagated along the characteristic starting in $(0, -0.5)$. This is in agreement with an alternative equivalent definition of a characteristic, from [20, p. 408]: "Discontinuities (of a nature to be specified later) of a solution cannot occur except along characteristics.". We add to this that the discontinuities mentioned only arise in the second-order derivatives, and u, p and q are smooth as seen in Figure 4.20. Furthermore, the region of increased residual increases for increasing x values. This is due to the interpolation along the vertical grid lines spreading the errors vertically. The exact solution at $x \geq 0$, and below the characteristic emanating from $(x, y) = (0, -0.5)$, is not known. We establish

convergence in this region by restricting the residual, viz., by considering

$$\eta_k = \max_{i,j}\{\epsilon_k \mid (x_i, y_j) \in [0.5, 1] \times [-0.5, 0]\}, \text{ for } k = 1, 2, \tag{4.37}$$

as function of h_y, which is shown on the right of Figure 4.19. Figure 4.20 also shows four characteristics, departing from the end points of the initial strip and from $(x, y) = (0, 0.5)$ and $(x, y) = (0, -0.5)$. Lastly, Figure 4.21 shows both r (left) and a (right) to be discontinuous, exactly along the characteristic emanating from $(x, y) = (0, -0.5)$.

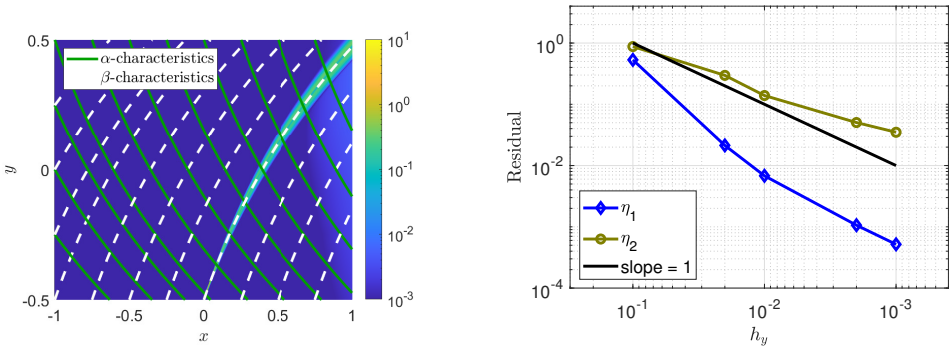


Figure 4.19: A heat map of the residual ϵ_1 (left) and the convergence of the residual η_k (right) for nonsmooth boundary value a.

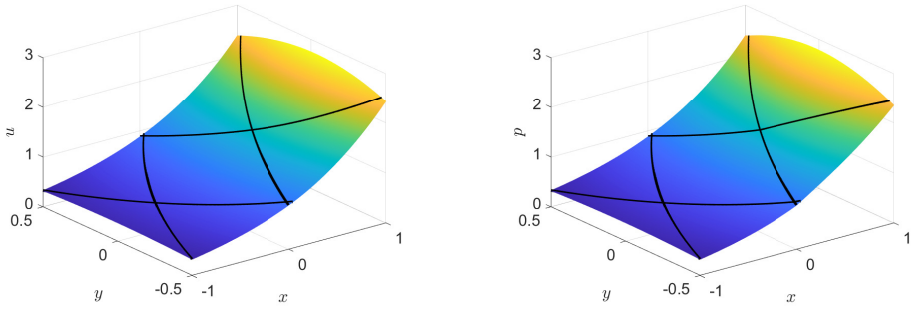


Figure 4.20: Smooth solutions u (left) and p (right) for nonsmooth boundary data.

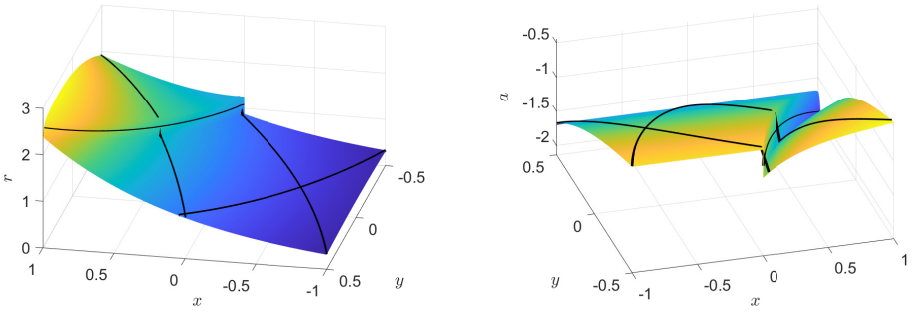


Figure 4.21: Nonsmooth solutions r (left) and a (right) for nonsmooth boundary data.

4.3 Numerical results for the general MA equation

Next we discuss numerical results for the general hyperbolic Monge-Ampère equation (3.42). Recall, that for the general hyperbolic Monge-Ampère equation we have an equivalent systems of ODEs given by (3.41). As is the case for the standard Monge-Ampère equation, we do not require all ODEs and only use the subset

$$
\begin{aligned}
x_\alpha &= \frac{B_2}{2\kappa^\alpha}, & x_\beta &= \frac{B_2}{2\kappa^\beta},\\
y_\alpha &= a\frac{B_2}{2\kappa^\alpha}, & y_\beta &= b\frac{B_2}{2\kappa^\beta},\\
u_\alpha &= (p+aq)\frac{B_2}{2\kappa^\alpha}, & u_\beta &= (p+bq)\frac{B_2}{2\kappa^\beta},\\
p_\alpha &= \left(\frac{a}{2}(A_2-\sqrt{\Delta})-A_3\right)\frac{B_2}{2\kappa^\alpha}, & p_\beta &= \left(\frac{b}{2}(A_2+\sqrt{\Delta})-A_3\right)\frac{B_2}{2\kappa^\beta},\\
q_\alpha &= \left(\frac{1}{2}(A_2+\sqrt{\Delta})-aA_1\right)\frac{B_2}{2\kappa^\alpha}, & q_\beta &= \left(\frac{1}{2}(A_2-\sqrt{\Delta})-bA_1\right)\frac{B_2}{2\kappa^\beta},\\
b_\alpha &= \left(\frac{A_{2,\alpha}-2bA_{1,\alpha}+\frac{2}{\kappa^\alpha}F^y}{2\sqrt{\Delta}}\right. & a_\beta &= \left(\frac{A_{2,\beta}-2aA_{1,\beta}+\frac{2}{\kappa^\beta}F^y}{2\sqrt{\Delta}}\right.\\
&\qquad \left.-\frac{4A_{4,\alpha}+(\Delta_{\tilde{F}})_\alpha}{4\Delta}\right)(a-b), & &\qquad \left.+\frac{4A_{4,\beta}+(\Delta_{\tilde{F}})_\beta}{4\Delta}\right)(a-b),
\end{aligned}
\tag{4.38}
$$

where we used equations (3.44), (3.49), (3.50), (3.51), and (3.55).

All ODEs only depend on x, y, u, p, q, a, b, the functions A_i and their derivatives. Furthermore, we have that Δ is given by (3.43), B_2 by (3.47), $(\Delta_{\tilde{F}})_\beta$ by (3.54). We also have F^y given by (3.58) and $A_{i,\beta}$ by (3.59).

For the numerical methods for the standard Monge-Ampère equation presented in Section 4.1, we assumed $\kappa^\alpha = \kappa^\beta = F_r = \frac{1}{2}B_2$. Here we generalized this approach. If we take $\kappa^\alpha = \kappa^\beta = \frac{1}{2}B_2$, then $x_\alpha = x_\beta = 1$, $y_\alpha = a$ and $y_\beta = b$. This allows the construction of an integration method in the positive x-direction by using x as the independent parameter. Similarly, we can integrate in the negative x-direction by using x as the independent parameter when $\kappa^\alpha = \kappa^\beta = -\frac{1}{2}B_2$. Furthermore, by choosing $(\kappa^\alpha, \kappa^\beta) = \pm\frac{1}{2}B_2(a, b)$ we integrate in the positive or negative y-direction by using y as the independent parameter. In this section we present numerical results for these four choices of κ.

Note that the numerical methods change only slightly in case $(\kappa^\alpha, \kappa^\beta) \neq \frac{1}{2}B_2$. For one, the functions $\mathbf{g}^\alpha$ and $\mathbf{g}^\beta$ are defined analogously to (4.2), containing the right-hand sides of the ODE systems (4.38). Furthermore, if $(\kappa^\alpha, \kappa^\beta) = \pm\frac{1}{2}B_2(a, b)$, we use y as independent variable instead of x and henceforth we formally swap x and y, h_x and h_y. The process is straightforward but tedious, and therefore omitted here. Only the adaptive step size control warrants further consideration. Analogous to Section 4.1.4, we limit $\Delta x(i) = |\tilde{x}^\beta_{j+1}(i) - x_i|$ by imposing restrictions on $(h_y)_j$. We do so by approximating $\Delta x(i)$ by a forward Euler step and subsequently setting $(h_y)_j < h_x$ to find

$$(h_y)_j = \gamma\, h_x \cdot \min_{i\in\{1,\dots,N_x\}} \left\{1, \left|a_{i,j}\right|, \left|b_{i,j}\right|\right\}, \tag{4.39}$$

where we again choose $\gamma = 0.95$.

The general criteria for the boundary conditions as given in Section 3.4, i.e., the cases presented there, do not change. However, the derivation for the function values implied by the boundary condition do change slightly. To see this, let us consider $(\kappa^\alpha, \kappa^\beta) = -\frac{1}{2}B_2$. Because we integrate in the negative x-direction, we do not prescribe the initial strip at the western boundary but at the eastern boundary instead, i.e., instead of prescribing $u_{\mathrm{W}}(y)$ and $p_{\mathrm{W}}(y)$ and calculating the remaining variables via (3.73) for the standard MA equation, we now prescribe $u_{\mathrm{E}}(y)$, $p_{\mathrm{E}}(y)$ and using equations (3.41), (3.44) and (3.46) we calculate

$$q(x_{\max}, y) = u'_{\mathrm{E}}(y), \tag{4.40a}$$
$$t(x_{\max}, y) = u''_{\mathrm{E}}(y), \tag{4.40b}$$
$$s(x_{\max}, y) = p'_{\mathrm{E}}(y), \tag{4.40c}$$
$$A_{i,\mathrm{E}}(y) = A_i(x_{\max}, y, u_{\mathrm{E}}(y), p_{\mathrm{E}}(y), q(x_{\max}, y)), \tag{4.40d}$$
$$\Delta_{\mathrm{E}}(y) = 4A_{4,\mathrm{E}}(y) + A^2_{2,\mathrm{E}}(y) - 4A_{1,\mathrm{E}}(y)A_{3,\mathrm{E}}(y), \tag{4.40e}$$
$$a(x_{\max}, y) = \frac{-2s(x_{\min}, y) + A_{2,\mathrm{E}}(y) + \sqrt{\Delta_{\mathrm{E}}(y)}}{2t(x_{\max}, y) + A_{1,\mathrm{E}}(y)}, \tag{4.40f}$$
$$b(x_{\max}, y) = \frac{-2s(x_{\min}, y) + A_{2,\mathrm{E}}(y) - \sqrt{\Delta_{\mathrm{E}}(y)}}{2t(x_{\max}, y) + A_{1,\mathrm{E}}(y)}, \tag{4.40g}$$
$$r(x_{\max}, y) = \frac{a(x_{\max}, y)b(x_{\max}, y)}{a(x_{\max}, y) - b(x_{\max}, y)}\sqrt{\Delta_{\mathrm{E}}(y)} - A_{3,\mathrm{E}}, \tag{4.40h}$$

with $i = 1, \ldots, 4$. The other cases follow analogously. We omit the details here and instead consider a few examples.

4.3.1 A smooth default test case

In Section 4.2 we discussed the default test case given by (4.22). The example can also be written as a general Monge-Ampère equation with coefficients

$$A_1 = A_2 = A_3 = 0, \qquad A_4 = \tfrac{1}{2}\big(\cos(2y) + \cosh(2x)\big). \tag{4.41}$$

We again consider the domain $\Omega = [0, 1] \times [-0.5, 0.5]$ and solve the general Monge-Ampère equation for the four choices of $(\kappa^\alpha, \kappa^\beta)$ given in Figure 4.22. The figure also shows the required boundary condition per edge of $\partial\Omega$. Numerical solutions have been calculated using the forward Euler, modified Euler and Runge-Kutta methods, using B-splines of order 2, 3 and 5, respectively. For each method we (subsequently) considered each of the four options $(\kappa^\alpha, \kappa^\beta) = \pm\frac{1}{2}B_2(1, 1)$, $(\kappa^\alpha, \kappa^\beta) = \pm\frac{1}{2}B_2(a, b)$. The numerical results converge for each of the algorithms as expected, that is, in line with the results of Section 4.2. We show a subset of the results in Figure 4.23 and 4.24. On the left of Figure 4.23 results for the forard Euler are shown for $(\kappa^\alpha, \kappa^\beta) = -\frac{1}{2}B_2(a, b)$. The figure shows first order convergence for each of the relevant function values, measured by the maximum of the error $|u_{i,j} - u(\mathbf{x}_{ij})|$. Similarily, on the right fourth order convergence for the Runge-Kutta method is observed. In 4.24 the maximum error in b is shown for the modified Euler method and κ^α and κ^β. On the left the errors are a function of h_x or h_y, depending on the choice

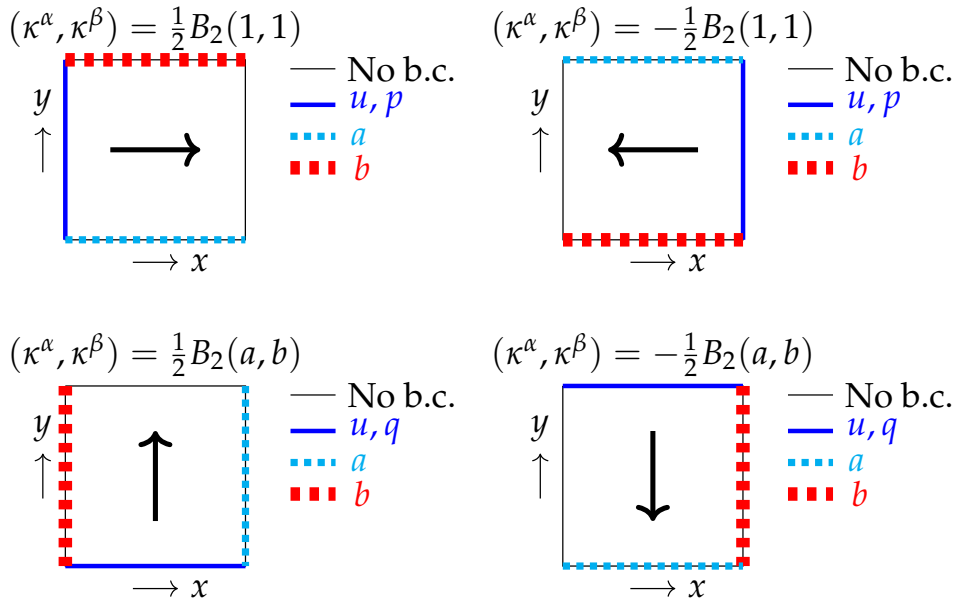


Figure 4.22: Schematic representation of the required boundary conditions for the default test case for various choices of κ^α and κ^β.

for κ^α and κ^β, and on the right as function of the total required number of grid point $N_x \cdot N_y$. Note, due to the adaptive step size control, we prescribe N_x when $(\kappa^\alpha, \kappa^\beta) = \pm\frac{1}{2}B_2(a,b)$ and N_y when $(\kappa^\alpha, \kappa^\beta) = \pm\frac{1}{2}B_2(1,1)$. We calculate N_y, in case of $(\kappa^\alpha, \kappa^\beta) = \pm\frac{1}{2}B_2(a,b)$, and N_x, in case of $(\kappa^\alpha, \kappa^\beta) = \pm\frac{1}{2}B_2(1,1)$, after the numerical solution is obtained, subsequently yielding the total number of grid points used. As function of h_x and h_y, we observe second-order convergence for each choice of κ^α and κ^β. No significant difference in the error of b is observed between the different choices of κ^α and κ^β for a fixed number of total grid points used.

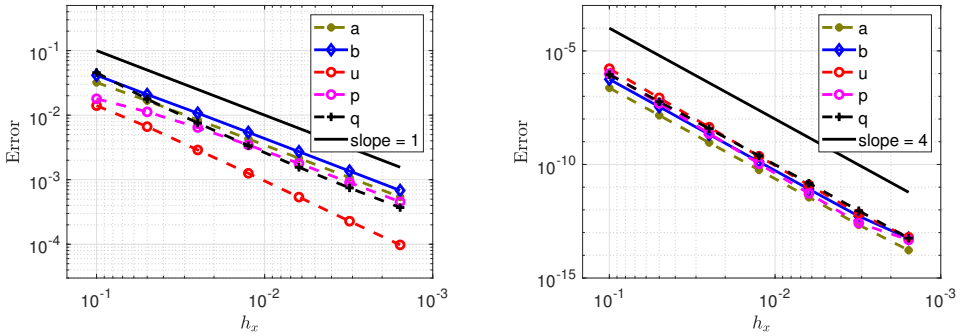


Figure 4.23: Convergence of the errors of the individual function values with $(\kappa^\alpha, \kappa^\beta) = -\frac{1}{2}B_2(a,b)$ for the forward Euler method (left) and the Runge-Kutta method (right)

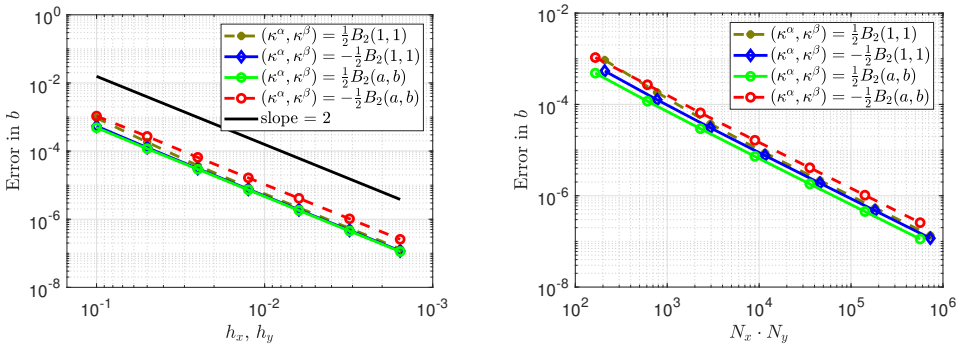


Figure 4.24: Convergence of the maximum error of b for various grids and various κ^α and κ^β for the modified Euler method as function of h_x or h_y (left) and $N_x \cdot N_y$ (right).

4.3.2 Logarithmic test case

We consider the following example:

$$u(x,y) = \tfrac{1}{2}\log(x^2+y^2), \tag{4.42a}$$
$$A_1(x,y,u,p,q) = (p+q)(x-y)\exp(-2u), \tag{4.42b}$$
$$A_2(x,y,u,p,q) = 6pq - \frac{3}{2xy}, \tag{4.42c}$$
$$A_3(x,y,u,p,q) = q^2 - p^2, \tag{4.42d}$$
$$A_4(x,y,u,p,q) = 4xypq\exp(-4u), \tag{4.42e}$$

with domain $\Omega = [1, 1.5] \times [2, 2.5]$. The required boundary conditions are the same as for the default test case and shown in Figure 4.22. All numerical solutions converge as expected, of these, we show two examples here. In Fig-

ure 4.25 the converge is shown for the forward Euler method with $(\kappa^\alpha, \kappa^\beta) = -\frac{1}{2}B_2(1,1)$ on the left and for the modified Euler with $(\kappa^\alpha, \kappa^\beta) = \frac{1}{2}B_2(a,b)$ on the right. Again, the observed convergence is first and second order, respectively. In Figure 4.26 the convergence of the global error in the numerical solution is shown as function of $N_x \cdot N_y$ for each of the four cases for κ^α and κ^β. The figure suggests that the integration direction, and correspondingly the choice for κ^α, κ^β and the placement of the initial base curve, does not influence the convergence of the algorithms.

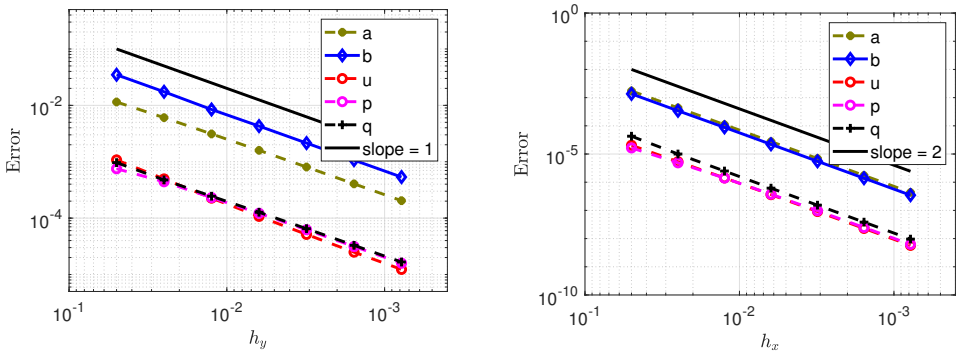


Figure 4.25: Convergence of the errors of the individual function values for the forward Euler method with $(\kappa^\alpha, \kappa^\beta) = -\frac{1}{2}B_2(1,1)$ (left) and the modified Euler for $(\kappa^\alpha, \kappa^\beta) = \frac{1}{2}B_2(a,b)$(right).

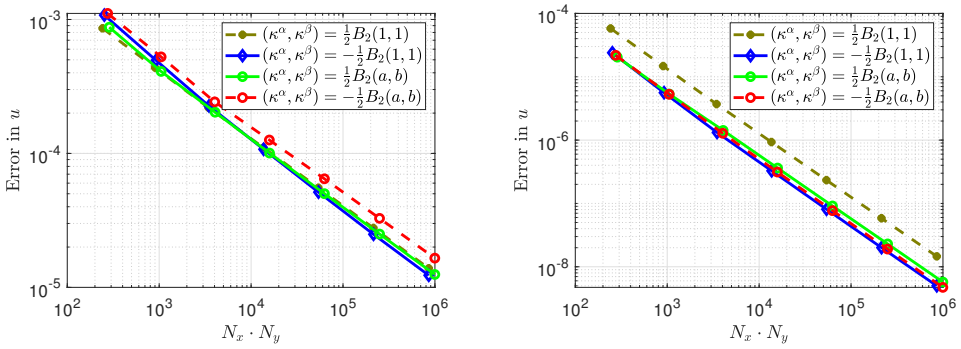


Figure 4.26: The maximum error of u for various grids and κ^α and κ^β for the forward Euler method (left) and the modified Euler method (right).

4.3.3 An initial strip extended over two edges

Next we present an example for which the initial strip extends over two edges. Consider the domain $\Omega = [-1, 0] \times [\frac{3}{4}, \frac{5}{4}]$ and

$$u(x, y) = \exp(x)\cos(y), \tag{4.43a}$$
$$A_1(x, y, u, p, q) = 1 + \frac{up}{\exp(x)\cos(y)}, \tag{4.43b}$$
$$A_2(x, y, u, p, q) = -\frac{1 + u + p}{q} - \exp(x)\sin(y), \tag{4.43c}$$
$$A_3(x, y, u, p, q) = 1 - q\frac{\cos(y)}{\sin(y)}, \tag{4.43d}$$
$$A_4(x, y, u, p, q) = (1 + \exp(x)\cos(y))^2. \tag{4.43e}$$

By (3.44a), i.e., the definition of a and b, we have $a, b > 0$ on Ω. Therefore, according to Section 3.4, we need to prescribe two initial strips. The initial strips for each choice of $(\kappa^\alpha, \kappa^\beta)$ are shown in Figure 4.27. It shows that we prescribe u and p on the edge $\{x_{\min}\} \times [y_{\min}, y_{\max}]$ and u and q on $[x_{\min}, x_{\max}] \times \{y_{\min}\}$ when integrating in either the positive x- or positive y-direction. The figure furthermore shows that we prescribe u and p on the edge $\{x_{\max}\} \times [y_{\min}, y_{\max}]$ and u and q on $[x_{\min}, x_{\max}] \times \{y_{\max}\}$ when integrating in either the negative x- or negative y-direction. Figure 4.28 shows convergence for the modified Euler method when integrating in the positive y-direction (left) and the convergence when integrating in the negative y-direction using the Runge-Kutta method

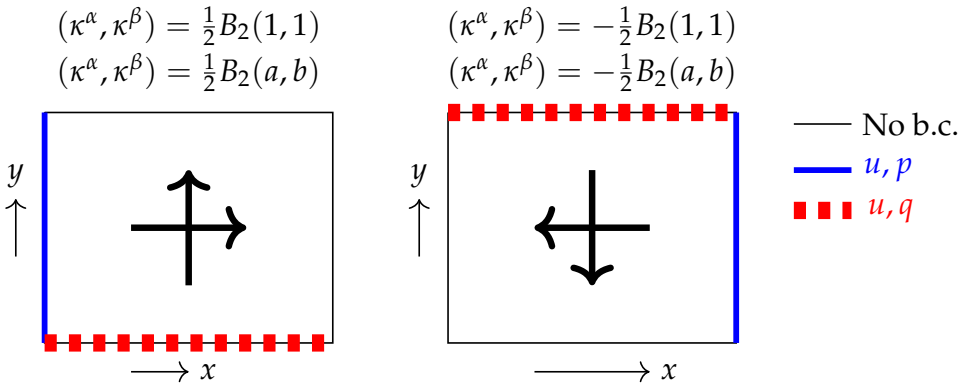


Figure 4.27: Schematic representation of the required boundary conditions for various choices of κ^α and κ^β.

(right). Both figures show the expected orders of convergence. In Figure 4.29 the convergence of the global error in the numerical solution is shown for the Runge-Kutta method for various grids and choices of $\kappa^\alpha, \kappa^\beta$. On the left as function of h_x and h_y, and on the right as function of the total number of grid points. The convergence shown on the left is in accordance with Section 4.2.

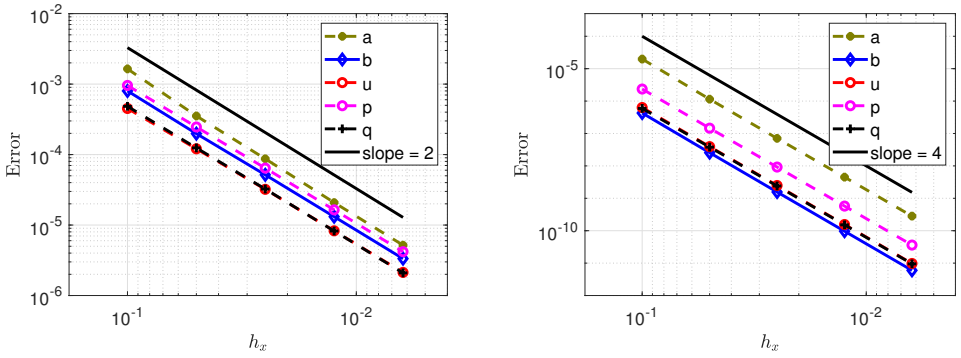


Figure 4.28: Convergence of the errors of the individual function values for the modified Euler method with $(\kappa^\alpha, \kappa^\beta) = \frac{1}{2}B_2(a,b)$ (left) and the Runge-Kutta method for $(\kappa^\alpha, \kappa^\beta) = -\frac{1}{2}B_2(a,b)$(right).

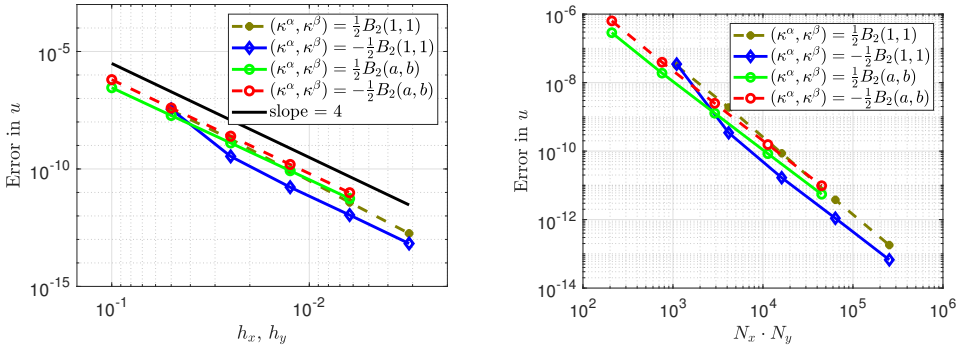


Figure 4.29: The maximum error of u for various grids and κ^α and κ^β for the Runge-Kutta method as function of h_x and h_y (left) and as function of $N_x \cdot N_y$ (right).

4.4 Summary

In this chapter we introduced numerical methods to solve the hyperbolic Monge-Ampère equation using the method of characteristics (MOC). The MOC yields two ODE systems which can be solved using explicit numerical integrators. We presented three such integrators, which are based on one-step methods. The computed characteristics will not pass through grid points. Therefore interpolation is necessary which should be handled carefully as not to spoil numerical convergence. We discussed how the direction of the characteristics at the boundary determines the number of boundary conditions which should be prescribed.

For test cases with known analytical solutions the developed methods are shown to converge to the analytical solution up to computer precision. Furthermore, two measures for the residual are formulated which converge to computer precision. The methods are shown to work for an example where the initial strip is extended over two boundary segments, and for another example for which the number of boundary conditions necessary varies along a boundary segment. A nonsmooth boundary condition was imposed to show the propagation of the nonsmooth boundary data along the characteristics while the algorithm remained stable. Furthermore, we tested the algorithm for the general Monge-Ampère equation where we integrated in the positive and negative x-direction as well as in the positive and negative y-direction.

Chapter 5

A Least-Squares Method for the Monge-Ampère Equation

In this chapter we introduce a least-squares method for the hyperbolic Monge-Ampère equation with transport boundary condition. We are motivated by applications to optical design. In [65] it was found that designing lenses and reflectors for some optical systems is equivalent to solving the standard Monge-Ampère (MA) equation with transport boundary condition. In the next chapter we consider the optical applications in more detail. Here we focus on the hyperbolic Monge-Ampère equation with transport boundary condition, given by

$$\nabla u(\partial\mathcal{X}) = \partial\mathcal{Y}, \tag{5.1}$$

with $u = u(x_1, x_2)$ the solution to the MA equation. So far, to the best of our knowledge, no reliable method to solve the hyperbolic MA equation with transport boundary conditions is known. Ideally we would resort to the method of characteristics introduced in earlier chapters. However, for the MOC we require a parametrization for the characteristics and subsequently Cauchy boundary conditions on an initial curve. The remaining boundary conditions then depend on the location of the characteristics, and consequently, the required boundary conditions are fundamentally different from the transport boundary condition. The transport boundary condition is therefore hard to implement for the method of characteristics.

In this chapter we therefore do not consider the method of characteristics, but instead resort to a least-squares method which has been proven to work for various elliptic problems, among which the standard Monge-Ampère equation [66], the generalized Monge-Ampère equation and the generated Jacobian equation [69]. The least-squares method is an iterative method which does not

directly solve for the solution u, but instead first constructs a mapping $\mathbf{m}$ and subsequently approximates u. The general outline of the least-squares method is as follows: first, we approximate the Jacobi matrix of $\mathbf{m}$ in the interior of the domain by minimizing an error functional. Secondly, $\mathbf{m}$ restricted to the boundary of the domain is approximated. By minimizing another error functional involving the newly found Jacobi matrix and the boundary approximation, we obtain a new approximation for the mapping. We repeat these three steps iteratively until $\mathbf{m}$ no longer changes and subsequently calculate u by minimizing a fourth least-squares functional. One of the benefits of this method is that each of the three stages can be adapted upon individually. For example, the minimization for $\mathbf{m}$ in [66] relies on a finite difference scheme while in [83] it uses a finite volume scheme. This three-stage approach allows us to introduce two new boundary methods, viz. a *segmented projection method* and a *segmented arc length method*, which both lead to better results and higher computational performance than the original *projection method* [66]. Even more importantly, the iterative method using the *segmented projection method* converges in some cases when the original *projection method* does not. And, as we will show, the *segmented arc length method* converges for all examples. Furthermore, we improve upon the first minimization procedure, viz. the procedure for approximating the Jacobi matrix in the interior of the domain, and we establish a method to prevent the crossing of grid lines in target space which is required for proper convergence.

The content of this chapter is as follows. We discuss the theory of the least-squares method for the Monge-Ampère equation in Section 5.1. First, in Section 5.1.1, the least-squares method is introduced. Next, we adapt part of the method, viz. the optimization in the interior domain, in Section 5.1.2. In Section 5.1.3, we introduce various boundary methods to replace the existing projection method and, in Sections 5.1.4 and 5.1.5, we introduce two grid-shock correction methods. In Section 5.2 we compare the boundary methods, show their weaknesses and strengths and elaborate on the convergence of the algorithm for various test cases. Lastly, we end with a discussion of the results followed by a summary in Section 5.3.

5.1 The least-squares formulation

We are interested in the two-dimensional hyperbolic Monge-Ampère equation with transport boundary condition, given by

$$\det\left(\mathrm{D}^2 u(\mathbf{x})\right) + f^2(\mathbf{x}, \nabla u(\mathbf{x})) = 0, \quad \mathbf{x} \in \mathcal{X}, \tag{5.2a}$$

$$\nabla u(\partial\mathcal{X}) = \partial\mathcal{Y}, \tag{5.2b}$$

where $u = u(\mathbf{x})$ is the unknown, $\mathrm{D}^2 u$ is the Hessian matrix of u, $f^2 > 0$ and $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^2$ are connected domains. We call $\mathcal{X}$ the source domain and $\mathcal{Y}$ the target domain. This naming convention will become clear in Chapter 7 where $\mathcal{X}$ represents the light source and $\mathcal{Y}$ represents the illuminated target. We require the boundaries $\partial\mathcal{X}$ and $\partial\mathcal{Y}$ to be orientable. The transport boundary condition (5.2b) can be interpreted as

$$\begin{cases} \forall \mathbf{x} \in \partial\mathcal{X} : \nabla u(\mathbf{x}) \in \partial\mathcal{Y}, \\ \forall \mathbf{y} \in \partial\mathcal{Y} \;\; \exists \mathbf{x} \in \partial\mathcal{X} : \nabla u(\mathbf{x}) = \mathbf{y}, \end{cases} \tag{5.3a}$$

where the latter condition is recognized as surjectivity of ∇u. Bijectivity does not hold in general, not even when restricted to the boundary, as will become apparent by the example discussed in Section 5.2.4. Recall that hyperbolicity of (5.2a) follows from the discriminant of the characteristic condition as stated in Section 3.1.4. The characteristic condition can be obtained by rewriting (5.2a) as

$$F(\mathbf{x}, u, p, q, r, s, t) = rt - s^2 + f^2 = 0, \tag{5.4}$$

where $p = u_{x_1}$, $q = u_{x_2}$, $r = u_{x_1x_1}$, $s = u_{x_1x_2}$ and $t = u_{x_2x_2}$, and is given by

$$F_r \mu^2 - F_s \mu + F_t = 0, \tag{5.5}$$

for the unknown function μ, representing the slope of the characteristics. For the MA equation to be hyperbolic, two real characteristics need to exist for every point in the domain, hence the slopes of the two characteristics, and thus the roots of (5.5), need to be real and distinct. The discriminant reads

$$\Delta = F_s^2 - 4F_rF_t = 4s^2 - 4tr = 4f^2, \tag{5.6}$$

which is, by assumption, strictly positive. Hence, equation (5.2a) is hyperbolic.

5.1.1 Least-squares approach

In [66] a least-squares method was introduced to solve the elliptic Monge-Ampère equation $\det(\mathrm{D}\mathbf{m}) = f^2(\mathbf{x}, \nabla u(\mathbf{x}))$ for $\mathbf{x} \in \mathcal{X}$ and $\mathrm{D}\mathbf{m}$ the Jacobi matrix of $\mathbf{m}$ and $\mathbf{m} = \nabla u$. The main idea of the least-squares method is to reformulate the Monge-Ampère equation in terms of the mapping $\mathbf{m} : \mathcal{X} \to \mathcal{Y}$, representing ∇u, and solve for $\mathbf{m}$. Subsequently, u is reconstructed from $\mathbf{m}$ in a least-squares sense. To solve the hyperbolic problem we replace the right-hand side of the elliptic Monge-Ampère equation by $-f^2(\mathbf{x}, \nabla u(\mathbf{x}))$ yielding

$$\det(\mathrm{D}\mathbf{m}(\mathbf{x})) + f^2(\mathbf{x}, \mathbf{m}(\mathbf{x})) = 0, \quad \mathbf{x} \in \mathcal{X}, \tag{5.7a}$$

$$\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}. \tag{5.7b}$$

Note that by substituting $\mathrm{D}\mathbf{m} = \mathrm{D}^2 u$ we recover (5.2a), the original problem. We formulate a minimization problem for $\mathbf{m}$ which we solve numerically. For this, we introduce the auxiliary functions $\mathbf{P} : \mathcal{X} \to \mathbb{R}^{2\times 2}$ and $\mathbf{b} : \partial\mathcal{X} \to \partial\mathcal{Y}$ which are used to approximate $\mathrm{D}\mathbf{m}$ on the whole domain and $\mathbf{m}$ on the boundary, respectively. This is achieved by the least-squares method, i.e., subsequently minimizing three separate functionals given by

$$J_{\mathrm{I}}(\mathbf{m}, \mathbf{P}) = \tfrac{1}{2} \iint_{\mathcal{X}} \|\mathrm{D}\mathbf{m} - \mathbf{P}\|^2 \,\mathrm{d}\mathbf{x}, \tag{5.8a}$$

$$J_{\mathrm{B}}(\mathbf{m}, \mathbf{b}) = \tfrac{1}{2} \oint_{\partial\mathcal{X}} |\mathbf{m} - \mathbf{b}|^2 \,\mathrm{d}s, \tag{5.8b}$$

$$J(\mathbf{m}, \mathbf{P}, \mathbf{b}) = \alpha J_{\mathrm{I}}(\mathbf{m}, \mathbf{P}) + (1-\alpha) J_{\mathrm{B}}(\mathbf{m}, \mathbf{b}), \tag{5.8c}$$

where $|\cdot|$ is the standard 2-norm, $\|\cdot\|$ is the Frobenius norm defined by $\|\mathbf{A}\|^2 = \mathrm{tr}(\mathbf{A}\mathbf{A}^{\mathrm{T}})$, where $\mathrm{tr}(\mathbf{A})$ is the trace of the matrix $\mathbf{A}$, and $0 < \alpha < 1$ is a constant control parameter to either place weights on the boundary or the interior. Starting with an initial guess $\mathbf{m}^0$, the iterative optimization procedure for $n = 0, 1, 2, \ldots$ reads

$$\mathbf{P}^{n+1} = \operatorname*{argmin}_{\mathbf{P}\in\mathcal{P}(\mathbf{m}^n)} J_{\mathrm{I}}(\mathbf{m}^n, \mathbf{P}), \tag{5.9a}$$

$$\mathbf{b}^{n+1} = \operatorname*{argmin}_{\mathbf{b}\in\mathcal{B}} J_{\mathrm{B}}(\mathbf{m}^n, \mathbf{b}), \tag{5.9b}$$

$$\mathbf{m}^{n+1} = \operatorname*{argmin}_{\mathbf{m}\in\mathcal{M}} J(\mathbf{m}, \mathbf{P}^{n+1}, \mathbf{b}^{n+1}). \tag{5.9c}$$

The spaces $\mathcal{P}(\mathbf{m}^n)$, $\mathcal{B}$ and $\mathcal{M}$ follow from three key observations, viz. first, $\mathbf{m} = \nabla u$ so the Jacobi matrix $\mathrm{D}\mathbf{m} = \mathrm{D}^2 u$ is symmetric and $\det(\mathrm{D}\mathbf{m}) = -f^2(\mathbf{x}, \mathbf{m}(\mathbf{x}))$. Second, by the transport boundary condition, for all $\mathbf{x} \in \partial\mathcal{X}$ we have $\mathbf{m}(\mathbf{x}) \in \partial\mathcal{Y}$. Third, as we require $\mathbf{m}$ to be twice continuously differentiable later on, we impose this requirement. The three sets are then given by

$$\mathcal{P}(\mathbf{m}) = \left\{ \mathbf{P} \in [C^1(\mathcal{X})]^{2\times 2} \mid \det(\mathbf{P}(\mathbf{x})) = -f^2(\mathbf{x}, \mathbf{m}(\mathbf{x})),\ \mathbf{P} = \mathbf{P}^{\mathrm{T}} \right\}, \tag{5.10a}$$

$$\mathcal{B} = \left\{ \mathbf{b} \in [C(\partial\mathcal{X})]^2 \mid \mathbf{b}(\mathbf{x}) \in \partial\mathcal{Y} \right\}, \tag{5.10b}$$

$$\mathcal{M} = [C^2(\mathcal{X})]^2. \tag{5.10c}$$

For the elliptic case, no results of well posedness of (5.9) nor any results of convergence of the total method are known. The method has empirically been shown to work by [65, 68, 82] in a wide variety of cases. On that basis we proceed by adopting the method for the hyperbolic Monge-Ampère equation.

We first outline the minimization of J, as it remains unchanged w.r.t. [66], and in the next sections we elaborate on the minimization of J_I and J_B.

To find the minimizer of J, we calculate the first variation of J w.r.t. $\mathbf{m}$ and equate it to zero. The first variation of J w.r.t. $\mathbf{m}$ for $\boldsymbol{\eta} \in [C^2(\mathcal{X})]^2$ reads

$$\delta J(\mathbf{m}, \mathbf{P}, \mathbf{b})(\boldsymbol{\eta}) = \lim_{\epsilon \to 0} \frac{J(\mathbf{m} + \epsilon\boldsymbol{\eta}, \mathbf{P}, \mathbf{b}) - J(\mathbf{m}, \mathbf{P}, \mathbf{b})}{\epsilon}. \tag{5.11}$$

Assuming that the first variations δJ_I and δJ_B of J_I and J_B w.r.t. $\mathbf{m}$ exist for all $\boldsymbol{\eta} \in [C^2(\mathcal{X})]^2$, we have

$$\delta J(\mathbf{m}, \mathbf{P}, \mathbf{b})(\boldsymbol{\eta}) = \alpha \delta J_\mathrm{I}(\mathbf{m}, \mathbf{P})(\boldsymbol{\eta}) + (1 - \alpha)\delta J_\mathrm{B}(\mathbf{m}, \mathbf{b})(\boldsymbol{\eta}). \tag{5.12}$$

We then obtain δJ_B and δJ_I, viz.

$$\begin{aligned}
\delta J_\mathrm{B}(\mathbf{m}, \mathbf{b})(\boldsymbol{\eta}) &= \lim_{\epsilon \to 0} \frac{J_\mathrm{B}(\mathbf{m} + \epsilon\boldsymbol{\eta}, \mathbf{b}) - J_\mathrm{B}(\mathbf{m}, \mathbf{b})}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{1}{2} \oint_{\partial\mathcal{X}} 2(\mathbf{m} - \mathbf{b}) \cdot \boldsymbol{\eta} + \epsilon|\boldsymbol{\eta}|^2 \,\mathrm{d}\mathbf{s} \\
&= \oint_{\partial\mathcal{X}} (\mathbf{m} - \mathbf{b}) \cdot \boldsymbol{\eta} \,\mathrm{d}\mathbf{s},
\end{aligned} \tag{5.13}$$

$$\begin{aligned}
\delta J_\mathrm{I}(\mathbf{m}, \mathbf{P})(\boldsymbol{\eta}) &= \lim_{\epsilon \to 0} \frac{J_\mathrm{I}(\mathbf{m} + \epsilon\boldsymbol{\eta}, \mathbf{P}) - J_\mathrm{I}(\mathbf{m}, \mathbf{P})}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{1}{2} \iint_{\mathcal{X}} 2(\mathrm{D}\mathbf{m} - \mathbf{P}) : \mathrm{D}\boldsymbol{\eta} + \epsilon\|\mathrm{D}\boldsymbol{\eta}\|^2 \,\mathrm{d}\mathbf{x} \\
&= \iint_{\mathcal{X}} (\mathrm{D}\mathbf{m} - \mathbf{P}) : \mathrm{D}\boldsymbol{\eta} \,\mathrm{d}\mathbf{x},
\end{aligned} \tag{5.14}$$

where we introduced the Frobenius inner product defined by $\mathbf{A} : \mathbf{B} = \mathrm{tr}(\mathbf{A}\mathbf{B}^\mathrm{T})$ for square matrices $\mathbf{A}$ and $\mathbf{B}$. Applying the divergence theorem to δJ_I yields

$$\begin{aligned}
\delta J_\mathrm{I}(\mathbf{m}, \mathbf{P}, \mathbf{b})(\boldsymbol{\eta}) = \oint_{\partial\mathcal{X}} \left[\begin{pmatrix} (\nabla m_1 - \mathbf{p}_1) \cdot \hat{\mathbf{n}} \\ (\nabla m_2 - \mathbf{p}_2) \cdot \hat{\mathbf{n}} \end{pmatrix} \right] \cdot \boldsymbol{\eta} \,\mathrm{d}s \\
- \iint_{\mathcal{X}} \begin{pmatrix} \Delta m_1 - \nabla \cdot \mathbf{p}_1 \\ \Delta m_2 - \nabla \cdot \mathbf{p}_2 \end{pmatrix} \cdot \boldsymbol{\eta} \,\mathrm{d}\mathbf{x},
\end{aligned} \tag{5.15}$$

where we introduced the unit outward normal vector $\hat{\mathbf{n}}$ to $\partial\mathcal{X}$ and the columns of $\mathbf{P}$,

$$\mathbf{p}_1 = \begin{pmatrix} p_{11} \\ p_{21} \end{pmatrix}, \qquad \mathbf{p}_2 = \begin{pmatrix} p_{12} \\ p_{22} \end{pmatrix}, \qquad \mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2]. \tag{5.16}$$

In total, the first variations of J then reads

$$\begin{aligned}\delta J(\mathbf{m},\mathbf{P},\mathbf{b})(\boldsymbol{\eta}) = \oint_{\partial\mathcal{X}} \left[\alpha \begin{pmatrix}(\nabla m_1 - \mathbf{p}_1)\cdot\hat{\mathbf{n}}\\ (\nabla m_2 - \mathbf{p}_2)\cdot\hat{\mathbf{n}}\end{pmatrix} + (1-\alpha)(\mathbf{m}-\mathbf{b})\right]\cdot\boldsymbol{\eta}\,\mathrm{d}s \\ - \iint_{\mathcal{X}} \alpha \begin{pmatrix}\Delta m_1 - \nabla\cdot\mathbf{p}_1\\ \Delta m_2 - \nabla\cdot\mathbf{p}_2\end{pmatrix}\cdot\boldsymbol{\eta}\,\mathrm{d}\mathbf{x}.\end{aligned} \tag{5.17}$$

Choosing $\eta_2 = 0$ and applying the fundamental lemma of calculus of variations [21, p. 185] yields that for the optimal $\mathbf{m}$, its first component satisfies a Poisson equation with Robin boundary condition given by

$$\Delta m_1 = \nabla\cdot\mathbf{p}_1, \qquad \mathbf{x}\in\mathcal{X}, \tag{5.18a}$$
$$(1-\alpha)m_1 + \alpha\nabla m_1\cdot\hat{\mathbf{n}} = (1-\alpha)b_1 + \alpha\mathbf{p}_1\cdot\hat{\mathbf{n}}, \qquad \mathbf{x}\in\partial\mathcal{X}. \tag{5.18b}$$

Similarly, for the second component we find

$$\Delta m_2 = \nabla\cdot\mathbf{p}_2, \qquad \mathbf{x}\in\mathcal{X}, \tag{5.19a}$$
$$(1-\alpha)m_2 + \alpha\nabla m_2\cdot\hat{\mathbf{n}} = (1-\alpha)b_2 + \alpha\mathbf{p}_2\cdot\hat{\mathbf{n}}, \qquad \mathbf{x}\in\partial\mathcal{X}. \tag{5.19b}$$

Upon convergence of (5.9) we reconstruct u from $\mathbf{m}$ by minimizing another least-squares functional, viz.

$$u = \operatorname*{argmin}_{\psi\in C^2(\mathcal{X})} I(\psi), \qquad I(\psi) := \tfrac{1}{2}\iint_{\mathcal{X}} |\nabla\psi - \mathbf{m}|^2\,\mathrm{d}\mathbf{x}. \tag{5.20}$$

To find the minimum we calculate the first variation of I, viz.

$$\begin{aligned}\delta I(u)(\eta) &= \lim_{\epsilon\to 0}\frac{I(u+\epsilon\eta)-I(u)}{\epsilon}\\ &= \lim_{\epsilon\to 0}\frac{1}{2\epsilon}\left[\iint_{\mathcal{X}}|\nabla(u+\epsilon\eta)-\mathbf{m}|^2\,\mathrm{d}\mathbf{x} - \iint_{\mathcal{X}}|\nabla u-\mathbf{m}|^2\,\mathrm{d}\mathbf{x}\right]\\ &= \lim_{\epsilon\to 0}\frac{1}{2}\iint_{\mathcal{X}}\epsilon|\nabla\eta|^2 + 2(\nabla u - \mathbf{m}\cdot\nabla\eta)\,\mathrm{d}\mathbf{x}\\ &= \iint_{\mathcal{X}}(\nabla u-\mathbf{m})\cdot\nabla\eta\,\mathrm{d}\mathbf{x}\\ &= \oint_{\partial\mathcal{X}}(\nabla u-\mathbf{m})\cdot\hat{\mathbf{n}}\,\mathrm{d}s - \iint_{\mathcal{X}}(\Delta u - \nabla\cdot\mathbf{m})\eta\,\mathrm{d}\mathbf{x}, \qquad \forall\eta\in C^2(\mathcal{X}),\end{aligned} \tag{5.21}$$

where in the last step we applied the divergence theorem. For u to minimize $I(u)$, we have $\delta I(u)(\eta) = 0$ for all $\eta\in C^2(\mathcal{X})$. Applying the fundamental lemma of calculus of variations yields

$$\Delta u = \nabla\cdot\mathbf{m}, \qquad \mathbf{x}\in\mathcal{X}, \tag{5.22a}$$
$$\nabla u\cdot\hat{\mathbf{n}} = \mathbf{m}\cdot\hat{\mathbf{n}}, \qquad \mathbf{x}\in\partial\mathcal{X}, \tag{5.22b}$$

which is a Poisson equation with Neumann boundary conditions. For (5.22) to admit a solution, the compatibility condition [1, p. 184]

$$\iint_{\mathcal{X}} \nabla \cdot \mathbf{m} \, \mathrm{d}\mathbf{x} - \oint_{\partial\mathcal{X}} \mathbf{m} \cdot \hat{\mathbf{n}} \, \mathrm{d}s = 0, \tag{5.23}$$

should hold. Note that it is automatically satisfied due to the divergence theorem.

We solve the three Poisson problems (5.18), (5.19) and (5.22) using finite differences (FD), more specifically, standard second-order central differences for both the first- and second-order derivatives. For grid points on the boundary we introduce ghost points, which we eliminate using the normal derivatives in the Robin boundary condition. The system we obtain from discretizing (5.18) and (5.19) needs to be solved in each iteration. In order to increase computational efficiency, we compute the LU-decomposition in the initialization of the algorithm. Note that the solution for u is not unique due to the (transport) boundary condition [31, p. A1438], which is also reflected by (5.22b), so we enforce uniqueness by fixing one function value of u, i.e., let $\mathbf{x} \in \mathcal{X}$ be arbitrary, we then impose the condition $u(\mathbf{x}) = 0$. In practice we assume $\mathcal{X} = [x_\mathrm{m}, x_\mathrm{M}] \times [y_\mathrm{m}, y_\mathrm{M}]$ and we impose $u(x_\mathrm{m}, y_\mathrm{m}) = 0$. Alternatively, one could prescribe the average value of u on the domain [68, p. 177].

5.1.2 Interior error minimization

The matrix $\mathrm{D}\mathbf{m}$ cannot be determined exactly during the iterative process. Because the integrand of J_I, i.e., $\|\mathrm{D}\mathbf{m} - \mathbf{P}\|^2$, does not depend on derivatives of $\mathbf{P}$ we employ a point-wise minimization. To this end we approximate $\mathrm{D}\mathbf{m}$ using standard finite differences. Let $\mathbf{x}_{ij} = ((x_1)_i, (x_2)_j) \in \mathcal{X}$ be the grid points of a Cartesian grid with $i = 1, \ldots, N_{x_1}$ and $j = 1, \ldots, N_{x_2}$ denoting the first and second coordinate, respectively. We write $\mathbf{m}_{ij} \approx \mathbf{m}(\mathbf{x}_{ij})$ and similar for the other variables. We approximate $(\mathrm{D}\mathbf{m})_{ij}$ by $\mathbf{D}_{ij}$ using central and one-sided second-order finite differences in the interior and at the boundary, respectively. This implies that $\mathbf{D}$ is in general not symmetric, while $\mathrm{D}\mathbf{m}$ and $\mathbf{P}$ are. By virtue of the point-wise minimization we proceed to drop the subscripts, e.g., we write $\mathbf{m}$ instead of $\mathbf{m}_{ij}$, for brevity.

Let $F(p_{11}, p_{22}, p_{12}) = \|\mathbf{D} - \mathbf{P}\|^2$; expanding it yields

$$F(p_{11}, p_{22}, p_{12}) = \tfrac{1}{2}\Big((p_{11} - d_{11})^2 + (p_{12} - d_{12})^2 + (p_{12} - d_{21})^2 + (p_{22} - d_{22})^2\Big). \tag{5.24}$$

We replace $\mathbf{D}$ by its symmetric part $\mathbf{D}_\mathrm{s} = \frac{1}{2}(\mathbf{D} + \mathbf{D}^\mathrm{T})$, or written in its compon-

ents, we introduce $d_s = \frac{1}{2}(d_{12} + d_{21})$ and

$$\mathbf{D}_s = \begin{pmatrix} d_{11} & d_s \\ d_s & d_{22} \end{pmatrix}. \tag{5.25}$$

Furthermore, we replace F by $F_s = \frac{1}{2}\|\mathbf{P} - \mathbf{D}_s\|^2$, i.e.,

$$F_s(p_{11}, p_{22}, p_{12}) = \tfrac{1}{2}\Big((p_{11} - d_{11})^2 + 2(p_{12} - d_s)^2 + (p_{22} - d_{22})^2\Big). \tag{5.26}$$

To justify the replacement, note that

$$F(p_{11}, p_{22}, p_{12}) = F_s(p_{11}, p_{22}, p_{12}) + \tfrac{1}{4}(d_{12} - d_{21})^2, \tag{5.27}$$

hence, (p_{11}, p_{22}, p_{12}) minimizes F if and only if it minimizes F_s. To obtain the minimizers, we minimize F_s under the condition $\mathbf{P} \in \mathcal{P}(\mathbf{m})$ using Lagrange multipliers. The Lagrangian is thus given by

$$\Lambda(p_{11}, p_{22}, p_{12}, \lambda) = F_s(p_{11}, p_{22}, p_{12}) + \lambda\left(p_{11}p_{22} - p_{12}^2 + f^2\right). \tag{5.28}$$

By setting the partial derivatives of Λ with respect to p_{11}, p_{22}, p_{12} and λ to zero, we find that the critical points of Λ have to satisfy

$$p_{11} + \lambda p_{22} = d_{11}, \tag{5.29a}$$
$$\lambda p_{11} + p_{22} = d_{22}, \tag{5.29b}$$
$$(1 - \lambda)p_{12} = d_s, \tag{5.29c}$$
$$p_{11}p_{22} - p_{12}^2 = -f^2. \tag{5.29d}$$

This system can be solved analytically and the results are given by Prins et al. [66, p. B942-B947] for the elliptic Monge-Ampère equation, with $-f^2$ replaced by f^2 in (5.29d). Unfortunately, the list of solutions is not complete as for the case $d_{11} = -d_{22}$, two roots of (5.29) are missing in [66]. We propose a different solution strategy here. First, two remarks are in place. While minimizing F_s, the matrix $\mathbf{D}_s$ and the function value of f are given and both $\mathbf{P}$ and λ have to be computed. Hence, we provide a classification in terms of $\mathbf{D}_s$ and the corresponding solutions of (5.29). Furthermore, because the matrix $\mathbf{D}_s$ is an approximation, $\det(\mathbf{D}_s) \neq -f^2$ and in general $\det(\mathbf{D}_s) \geq 0$ may occur. We first write the linear equations of (5.29) as

$$\mathbf{\Lambda p} = \mathbf{d}, \quad \mathbf{\Lambda} = \begin{pmatrix} 1 & \lambda & 0 \\ \lambda & 1 & 0 \\ 0 & 0 & 1 - \lambda \end{pmatrix}, \quad \mathbf{p} = \begin{pmatrix} p_{11} \\ p_{22} \\ p_{12} \end{pmatrix}, \quad \mathbf{d} = \begin{pmatrix} d_{11} \\ d_{22} \\ d_s \end{pmatrix}. \tag{5.30}$$

The vector $\mathbf{p}$ is uniquely determined when $\mathbf{\Lambda}$ is regular, i.e., when $0 \neq \det(\mathbf{\Lambda}) = (1-\lambda)^2(1+\lambda)$. We should therefore distinguish between the cases $\lambda = 1$, $\lambda = -1$ and $\lambda \neq \pm 1$.

Although we should consider the cases $\lambda = 1$, $\lambda = -1$ and $\lambda \neq \pm 1$ separately, $\mathbf{d}$ and f are given and λ and $\mathbf{p}$ are to be calculated. Therefore we consider three cases based on $\mathbf{D}_\mathrm{s}$, viz., ***Case 1:*** $d_{11} = d_{22}$ and $d_\mathrm{s} = 0$, ***Case 2:*** $d_{11} = -d_{22}$ and ***Case 3:*** all other $\mathbf{D}_\mathrm{s}$. We consider $\mathbf{D}_\mathrm{s} = \mathbf{0}$ as a special case of $d_{11} = d_{22}$ and $d_\mathrm{s} = 0$.

We start with some general results, to be used in the subsequent derivations. Using (5.29) we find

$$\delta_\mathrm{s} := \det(\mathbf{D}_\mathrm{s}) = \lambda \mathrm{tr}(\mathbf{P})^2 - (\lambda-1)^2 f^2, \tag{5.31a}$$

$$\mathrm{tr}(\mathbf{D}_\mathrm{s}) = (\lambda+1)\mathrm{tr}(\mathbf{P}). \tag{5.31b}$$

Solving the second equation for $\mathrm{tr}(\mathbf{P})$ and subsequently substituting it in the first equation yields

$$f^2(\lambda^2-1)^2 + \delta_\mathrm{s}(\lambda+1)^2 - \mathrm{tr}(\mathbf{D}_\mathrm{s})^2\lambda = 0. \tag{5.32}$$

Next, we consider the roots of (5.32) and the corresponding solutions $\mathbf{P}$.

Case 1: $d_{11} = d_{22}$ and $d_\mathrm{s} = 0$, which we write as $\mathbf{D}_\mathrm{s} = d\mathbf{I}$ with $d \in \mathbb{R}$. We will show that this condition is equivalent with $\lambda = 1$. So, let $\mathbf{D}_\mathrm{s} = d\mathbf{I}$. We show that $\lambda = 1$ by forcing a contradiction, so, assume $\lambda \neq 1$. Then subtracting (5.29a) from (5.29b) gives $p_{11} = p_{22}$ and by (5.29c) we have $p_{12} = 0$. Substitution of $p_{11} = p_{22}$ and $p_{12} = 0$ in (5.29d) yields $p_{11}^2 = -f^2 < 0$, being a contradiction. Therefore $\lambda = 1$. Conversely, substitution of $\lambda = 1$ in $\mathbf{\Lambda}$ gives

$$\mathbf{\Lambda} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{5.33}$$

In this case the null space of $\mathbf{\Lambda}$ is given by $\mathcal{N}(\mathbf{\Lambda}) = \langle \mathbf{v}_1, \mathbf{v}_2 \rangle$ with $\mathbf{v}_1 = (1,-1,0)^\mathrm{T}$ and $\mathbf{v}_2 = (0,0,1)^\mathrm{T}$. So $\mathbf{\Lambda p} = \mathbf{d}$ only has a solution if $\mathbf{d}$ lies in the column space of $\mathbf{\Lambda}$, i.e., if $d_{11} = d_{22}$ and $d_\mathrm{s} = 0$ or $\mathbf{D}_\mathrm{s} = d\mathbf{I}$ with $d \in \mathbb{R}$. Henceforth we have that $\lambda = 1$ is equivalent with $\mathbf{D}_\mathrm{s} = d\mathbf{I}$ and thus $\lambda = 1$ only occurs in ***Case 1***. The general solution to $\mathbf{\Lambda p} = \mathbf{d}$ is now given by

$$\mathbf{p} = (p,\ d-p,\ 0)^\mathrm{T} + \mu_1\mathbf{v}_1 + \mu_2\mathbf{v}_2, \qquad p, \mu_1, \mu_2 \in \mathbb{R}. \tag{5.34}$$

We aim to minimize F_s. Substitution of (5.34) in F_s gives

$$F_\mathrm{s}(p_{11}, p_{22}, p_{12}) = \tfrac{1}{2}\Big((p+\mu_1-d)^2 + 2\mu_2^2 + (p+\mu_1)^2\Big), \tag{5.35}$$

thus showing $\mu_2 = 0$. By using the substitution $p = \tilde{p} - \mu_1$ we find $\mathbf{p} = (\tilde{p},\ d - \tilde{p},\ 0)^{\mathrm{T}}$ and

$$F_{\mathrm{s}}(p_{11}, p_{22}, p_{12}) = \tfrac{1}{2}\Big((\tilde{p} - d)^2 + \tilde{p}^2\Big). \tag{5.36}$$

The latter equation is only dependent on $\tilde{p}$ and the given value d, effectively reducing the minimization over p and μ_1 to a minimization over $\tilde{p} \in \mathbb{R}$. Subsequent substitution of $\mathbf{p}$ into (5.29d) gives $\tilde{p}(d - \tilde{p}) = -f^2$. This second-order polynomial in $\tilde{p}$ has two real roots, viz.

$$\tilde{p} = \frac{1}{2}\left(d \pm \sqrt{d^2 + 4f^2}\right). \tag{5.37}$$

So in total we find the two solutions

$$p_{11} = \frac{1}{2}\left(d \pm \sqrt{d^2 + 4f^2}\right), \qquad p_{22} = d - p_{11}, \qquad p_{12} = 0. \tag{5.38}$$

In case $d = 0$, i.e., in case $\mathbf{D}_{\mathrm{s}} = \mathbf{0}$, the above derivation still holds so we consider $\mathbf{D}_{\mathrm{s}} = \mathbf{0}$ an instance of ***Case 1***.

Case 2: $d_{11} = -d_{22}$, which we write as $\mathbf{d} = (d,\ -d,\ d_{\mathrm{s}})^{\mathrm{T}}$ with $d, d_{\mathrm{s}} \in \mathbb{R}$. We have that $\mathrm{tr}(\mathbf{D}_{\mathrm{s}}) = 0$ and $\delta_{\mathrm{s}} = -(d^2 + d_{\mathrm{s}}^2)$. For this case the fourth-order polynomial (5.32) can be written as

$$(\lambda + 1)^2\left((\lambda - 1)^2 + \frac{\delta_{\mathrm{s}}}{f^2}\right) = 0. \tag{5.39}$$

It follows that we have three unique roots, $\lambda = -1$ (with multiplicity 2) and $\lambda = 1 \pm \sqrt{|\delta_{\mathrm{s}}|}/f$.

- In case $\lambda = -1$ we have

$$\mathbf{\Lambda} = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}, \tag{5.40}$$

  and the corresponding null space $\mathcal{N}(\mathbf{\Lambda}) = \langle \mathbf{v}_3 \rangle$ with $\mathbf{v}_3 = (1, 1, 0)^{\mathrm{T}}$. For $\mathbf{p}$ to be a solution to $\mathbf{\Lambda}\mathbf{p} = \mathbf{d}$ we require $\mathbf{d}$ to be in the column space of $\mathbf{\Lambda}$. It follows that $\mathbf{d} = (d,\ -d,\ d_{\mathrm{s}})^{\mathrm{T}}$, $d, d_{\mathrm{s}} \in \mathbb{R}$. Henceforth $\lambda = -1$ only occurs for ***Case 2***.

  The general solution to $\mathbf{\Lambda}\mathbf{p} = \mathbf{d}$ is therefore given by

$$\mathbf{p} = (p,\ p - d,\ \tfrac{1}{2}d_{\mathrm{s}})^{\mathrm{T}} + \mu_3\mathbf{v}_3, \qquad p, \mu_3 \in \mathbb{R}. \tag{5.41}$$

Writing $\mathbf{p} = (\tilde{p},\ \tilde{p} - d,\ \frac{1}{2}d_s)^T$ with $\tilde{p} = p + \mu_3$ shows that the actual solution $\mathbf{p}$ does not change by choosing μ_3, so we simply choose $\mu_3 = 0$. By (5.29d) it follows that

$$p(p-d) - \tfrac{1}{4}d_s^2 + f^2 = 0. \tag{5.42}$$

Consequently, solving for p we find that for $|\delta_s| - 4f^2 \geq 0$ we have

$$p_{11} = \tfrac{1}{2}\left(d \pm \sqrt{|\delta_s| - 4f^2}\right), \qquad p_{22} = p_{11} - d, \qquad p_{12} = \tfrac{1}{2}d_s. \tag{5.43}$$

When $|\delta_s| - 4f^2 < 0$ the solution $\mathbf{p}$ is complex. Because we are only interested in real-valued solutions, we do not consider $\lambda = -1$ when $|\delta_s| - 4f^2 < 0$.

- In the case $\lambda = 1 \pm \sqrt{|\delta_s|}/f$, the matrix $\mathbf{\Lambda}^{-1}$ is uniquely defined (see (5.45) for an explicit expression) and by $\mathbf{p} = \mathbf{\Lambda}^{-1}\mathbf{d}$ we obtain

$$p_{11} = \mp\frac{df}{\sqrt{|\delta_s|}}, \qquad p_{22} = -p_{11}, \qquad p_{12} = \mp\frac{d_s f}{\sqrt{|\delta_s|}}. \tag{5.44}$$

The solutions (5.44) are new with respect to those found by Prins et al. [66] and are not specific to the hyperbolic Monge-Ampère equation.

Case 3: All other $\mathbf{D}_s$, i.e., both $\mathbf{D}_s \neq d\mathbf{I}$ and $\mathbf{d}_s \neq (d,\ -d,\ d_s)^T$ for all $d, d_s \in \mathbb{R}$. By ***Case 1*** we have $\lambda \neq 1$ and by ***Case 2*** we have $\lambda \neq -1$. Therefore $\det \mathbf{\Lambda} = (1-\lambda)^2(1+\lambda) \neq 0$. Consequently $\mathbf{\Lambda}$ is invertible and its inverse is given by

$$\mathbf{\Lambda}^{-1} = \frac{1}{1-\lambda^2}\begin{pmatrix} 1 & -\lambda & 0 \\ -\lambda & 1 & 0 \\ 0 & 0 & 1+\lambda \end{pmatrix}. \tag{5.45}$$

The values for λ are obtained by solving (5.32). The roots of this fourth-order polynomial can be determined analytically using Ferrari's method [71, p. 22] and are given in [66, p. B945]. For $\mathbf{p}$ we subsequently find $\mathbf{p} = \mathbf{\Lambda}^{-1}\mathbf{d}$, or more explicitly

$$p_{11} = \frac{\lambda d_{22} - d_{11}}{\lambda^2 - 1}, \qquad p_{22} = \frac{\lambda d_{11} - d_{22}}{\lambda^2 - 1}, \qquad p_{12} = \frac{d_s}{1-\lambda}. \tag{5.46}$$

5.1.3 Boundary method

In [65, p. 131-133] a *projection method* (PM) has been proposed for the minimization of (5.8b). As our numerical results will show, this method proves insufficient for some examples. Therefore we developed two improved methods, viz., the *segmented projection method* (SPM) and the *segmented arc length method* (SALM). Before we introduce the boundary methods, we first introduce some notation.

Let $\mathbf{x}_{ij} = ((x_1)_i, (x_2)_j) \in \mathcal{X}$ be the grid points of a Cartesian grid with $i = 1, \ldots, N_{x_1}$ and $j = 1, \ldots, N_{x_2}$ denoting the first and second coordinate, respectively. Let $\mathbf{x}_l$ be the grid points restricted to $\partial\mathcal{X}$ for $l = 1, \ldots, N$. We index $\mathbf{x}_l = 1, \ldots, N$ in the clockwise direction such that $\mathbf{x}_1 = \mathbf{x}_{1,1}$, i.e. the first point on $\partial\mathcal{X}$ equals the point $\mathbf{x}_{ij}$ with $i = j = 1$. We write $\mathbf{m}_{ij} \approx \mathbf{m}(\mathbf{x}_{ij})$, $\mathbf{m}_l \approx \mathbf{m}(\mathbf{x}_l)$ and similarly for the other variables.

The main idea behind SPM and SALM is to partition the boundaries of the source and target domains in segments. We then uniquely enforce one source segment to be mapped to one target segment. We then distribute $\mathbf{b}_l$, corresponding to $\mathbf{m}_l$ by either a projection (SPM) or by a ratio of arc lengths (SALM).

Let the boundary segments of $\mathcal{X}$ be the curves $\Gamma_k^{\mathcal{X}} \subset \partial\mathcal{X}$ such that for N_Γ boundary segments we have $\cup_{k=1}^{N_\Gamma} \Gamma_k^{\mathcal{X}} = \partial\mathcal{X}$. We denote $\Gamma_{N_\Gamma+1}^{\mathcal{X}} = \Gamma_1^{\mathcal{X}}$ and assume the intersections $\Gamma_{k_1}^{\mathcal{X}} \cap \Gamma_{k_2}^{\mathcal{X}}$ for $k_2 > k_1$ contains precisely one element if $k_2 = k_1 + 1$ or if $k_1 = 1$, $k_2 = N_\Gamma$ and no elements otherwise. Furthermore, we require each $\Gamma_k^{\mathcal{X}}$ to be parametrizable. We assume similar properties for $\Gamma_k^{\mathcal{Y}}$. We aim to map each boundary segment of $\partial\mathcal{X}$ to a boundary segment of $\partial\mathcal{Y}$, hence we enforce $\mathbf{m}(\Gamma_k^{\mathcal{X}}) = \Gamma_k^{\mathcal{Y}}$ for $k = 1, \ldots, N_\Gamma$, from which it follows that

$$\mathbf{m}(\partial\mathcal{X}) = \mathbf{m}(\cup_{k=1}^{N_\Gamma} \Gamma_k^{\mathcal{X}}) = \cup_{k=1}^{N_\Gamma} \mathbf{m}(\Gamma_k^{\mathcal{X}}) = \cup_{k=1}^{N_\Gamma} \Gamma_k^{\mathcal{Y}} = \partial\mathcal{Y}, \tag{5.47}$$

which is the required transport boundary condition.

Figure 5.1 shows a part of $\partial\mathcal{Y}$ and (parts of) three boundary segments. In the following we fix k and for brevity drop the subscript in $\Gamma_k^{\mathcal{X}}$ and $\Gamma_k^{\mathcal{Y}}$.

Let $\mathbf{y}_i$ for $i = 1, \ldots, N_\mathrm{b}$ be a discretization of the boundary segment $\Gamma^{\mathcal{Y}}$ such that for a given counter clockwise parametrization $\mathbf{y}(s) : [0, 1] \to \Gamma^{\mathcal{Y}}$, we have $\mathbf{y}_1 = \mathbf{y}(0)$ and $\mathbf{y}_{N_\mathrm{b}} = \mathbf{y}(1)$. We choose to parametrize $\partial\mathcal{X}$ and $\partial\mathcal{Y}$ in opposite directions as all examples have experimentally shown that $\mathbf{m}$ reverts the direction if it is a solution to the hyperbolic Monge-Ampère equation.

Projection method. We briefly explain PM as introduced in [65, p. 131-133]. We perform the following for each approximation $\mathbf{m}_l$ individually. Let $N_\Gamma = 1$, i.e., we consider the whole boundary as one boundary segment. Furthermore, let

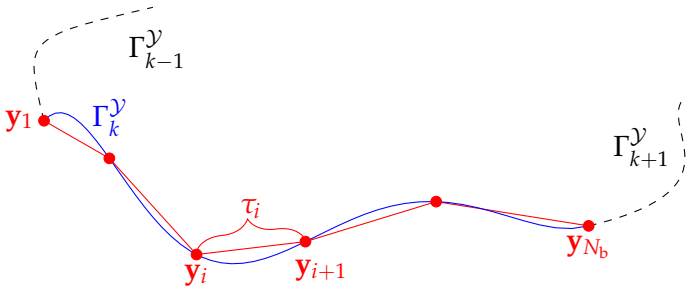


Figure 5.1: Schematic overview of the discretization of $\partial\mathcal{Y}$.

$\mathbf{y}_{N_\mathrm{b}+1} = \mathbf{y}_1$, we connect adjacent points $\mathbf{y}_i$ and $\mathbf{y}_{i+1}$ by straight line segments. The projection of $\mathbf{m}_l$ onto the line connecting $\mathbf{y}_i$ and $\mathbf{y}_{i+1}$ is given by

$$\mathbf{m}_l^{\mathrm{P}}(t_i) = \mathbf{y}_i + t_i(\mathbf{y}_{i+1} - \mathbf{y}_i), \tag{5.48a}$$

$$t_i = \frac{(\mathbf{m}_l - \mathbf{y}_i) \cdot (\mathbf{y}_{i+1} - \mathbf{y}_i)}{|\mathbf{y}_{i+1} - \mathbf{y}_i|^2}. \tag{5.48b}$$

As only $0 \leq t_i \leq 1$ corresponds to a point on the line segment between $\mathbf{y}_i$ and $\mathbf{y}_{i+1}$, we limit t_i according to $\hat{t}_i = \min(1, \max(0, t_i))$. Among all possible line segments, we choose $\mathbf{b}_l$ corresponding to $\mathbf{m}_l$, such that the distance is smallest, i.e.,

$$i_{\min} = \operatorname*{argmin}_i \{|\mathbf{m}_l^{\mathrm{P}}(\hat{t}_i) - \mathbf{m}_l|\}, \tag{5.49a}$$

$$\mathbf{b}_l = \mathbf{m}_l^{\mathrm{P}}(\hat{t}_{i_{\min}}). \tag{5.49b}$$

Segmented projection method. Let $\Gamma_k^{\mathcal{X}} \subseteq \partial\mathcal{X}$ and $\Gamma_k^{\mathcal{Y}} \subseteq \partial\mathcal{Y}$ be boundary segments of the source and target, respectively. For this method, we apply PM to the individual boundary segments instead of the whole boundary at once. Furthermore, we set

$$\mathbf{b}(\Gamma_k^{\mathcal{X}} \cap \Gamma_{k+1}^{\mathcal{X}}) = \Gamma_k^{\mathcal{Y}} \cap \Gamma_{k+1}^{\mathcal{Y}}, \quad 1 \leq k \leq N_\Gamma, \tag{5.50}$$

meaning, we map the end points of the source segments to the end points of the corresponding target segments. In practice, these end points are the corners of the source and target domains.

Segmented arc length method. The core idea of this method is as follows: if $\mathbf{m}(\Gamma^{\mathcal{X}}) = \Gamma^{\mathcal{Y}}$, then the arc length of the curve $\mathbf{m}(\Gamma^{\mathcal{X}})$ should be equal to the

arc length of the curve $\Gamma^{\mathcal{Y}}$. Numerically we approximate this condition by approximating the arc length of both $\Gamma^{\mathcal{Y}}$ and the distance between the points $\{\mathbf{m}(\mathbf{x}_l) \mid \mathbf{x}_l \in \Gamma^{\mathcal{X}}\}$.

We start with the arc length of the curve $\Gamma^{\mathcal{Y}}$. We approximate the arc length between $\mathbf{y}_i$ and $\mathbf{y}_{i+1}$ along $\Gamma^{\mathcal{Y}}$ by the length of the line segment connecting $\mathbf{y}_i$ and $\mathbf{y}_{i+1}$. We denote the approximation by

$$\tau_i = |\mathbf{y}_{i+1} - \mathbf{y}_i|, \quad i = 1, \dots, N_\mathrm{b} - 1. \tag{5.51}$$

The approximate cumulative arc length between $\mathbf{y}_1$ and $\mathbf{y}_i$ in the direction of increasing s is then given by

$$t_i = \sum_{j=1}^{i-1} \tau_j, \quad i = 1, \dots, N_\mathrm{b}. \tag{5.52}$$

The total arc length from $\mathbf{y}_1$ to $\mathbf{y}_{N_\mathrm{b}}$ is then approximated by $L = t_{N_\mathrm{b}}$. We use the cumulative arc lengths to introduce a piece-wise linear interpolation $\mathbf{b}_\mathrm{int}$ approximating $\mathbf{y}(s)$, viz.

$$\mathbf{b}_\mathrm{int}(t) = \mathbf{y}_i + \frac{t - t_i}{t_{i+1} - t_i}(\mathbf{y}_{i+1} - \mathbf{y}_i), \quad t_i \leq t \leq t_{i+1}, \tag{5.53}$$

where the scalar factor is a scaled coordinate between $\mathbf{y}_i$ and $\mathbf{y}_{i+1}$. Note that by construction $\mathbf{b}_\mathrm{int}$ satisfies

$$\mathbf{b}_\mathrm{int}(t_i) = \mathbf{y}_i, \quad i = 1, \dots, N_\mathrm{b}, \tag{5.54}$$

and $\mathbf{b}_\mathrm{int}(t)$ is a linear approximation of $\partial\mathcal{Y}$ for $0 \leq t \leq N_\mathrm{b}$.

Next we consider the points $\mathbf{m}(\mathbf{x}_l)$ with $\mathbf{x}_l \in \Gamma^{\mathcal{X}}$. Let N_m be the number of grid points on $\Gamma^{\mathcal{X}}$ such that $\mathbf{x}_l \in \Gamma^{\mathcal{X}}$ for $l = 1, \dots, N_\mathrm{m}$. Furthermore, let

$$\sigma_l = |\mathbf{m}_{l+1} - \mathbf{m}_l|, \quad l, \dots, N_\mathrm{m} - 1, \tag{5.55}$$

be an approximation of the arc length from $\mathbf{m}(\mathbf{x}_l)$ to $\mathbf{m}(\mathbf{x}_{l+1})$ along $\partial\mathcal{Y}$. This again introduces a cumulative arc length and a total arc length, respectively, given by

$$s_l = \sum_{j=1}^{l-1} \sigma_j, \quad l = 1, \dots, N_\mathrm{m}, \qquad \tilde{L} = s_{N_\mathrm{m}}. \tag{5.56}$$

Because $\mathbf{m}_l$ is an approximation and $\Gamma^{\mathcal{Y}}$ is approximated by straight line segments, $\tilde{L} \neq L$ in general. Hence, $s_{N_\mathrm{m}} \neq L$ may occur such that the end points of $\Gamma^{\mathcal{X}}$ may not be mapped to the end points of $\Gamma^{\mathcal{Y}}$. Furthermore, in

the limits $N_\mathrm{b}, N_x, N_y \to \infty$ with $\mathbf{m}_l$ the exact solution in the point $\mathbf{x}_\ell$, we have $L = \tilde{L}$. By scaling s_l according to

$$\tilde{s}_l = \frac{L}{\tilde{L}} s_l, \tag{5.57}$$

we ensure endpoints of $\Gamma^{\mathcal{X}}$ are mapped to endpoints of $\Gamma^{\mathcal{Y}}$ with proper behaviour in the limit. It follows that $\mathbf{b}_\mathrm{int}(\tilde{s}_l)$ forms a proper approximation for $\mathbf{m}_l$ restricted to $\Gamma^{\mathcal{Y}}$, viz.

$$\mathbf{b}_l = \mathbf{b}_\mathrm{int}(\tilde{s}_l), \quad l = 1, \ldots, N_\mathrm{m}. \tag{5.58}$$

5.1.4 Grid-shock correction

Using the methods outlined above, it is possible that the approximation $\mathbf{m}^n$ of $\mathbf{m}$ contains crossing grid lines, also known as grid shocks [19]. This phenomenon is shown in Figure 5.2 for an example we discuss in Section 5.2.2, with grid parameters $N_{x_1} = N_{x_2} = 321$ after $n = 15,000$ iterations. Though the solution on the left may look visually correct, the grid shock, as seen on the right, prevents proper numerical convergence of our algorithm.

To detect grid shocks, consider a point $\mathbf{x}_{ij} \in \partial\mathcal{X}$ as shown for $j = 1$ in Figure 5.3 on the left, and the corresponding image $\mathbf{m}_{ij}$ shown on the right. If both $\mathbf{m}_{ij}$ and $\mathbf{b}_{ij}$ are exact, then $|\mathbf{m}_{ij} - \mathbf{b}_{ij}| = 0$ and $|\mathbf{m}_{kl} - \mathbf{b}_{ij}| > 0$ for all $(k,l) \neq (i,j)$. Because both $\mathbf{m}$ and $\mathbf{b}$ are approximated, $|\mathbf{m}_{ij} - \mathbf{b}_{ij}| \neq 0$ in general. To detect grid shocks, we compute the distance $|\mathbf{m}_{kl} - \mathbf{b}_{ij}|$ for all (k,l) such that $|(k,l)^\mathrm{T} - (i,j)^\mathrm{T}|_1 \leq 2$. If the minimum distance is found for $(k,l) \neq (i,j)$, then we assume a grid shock occurs and we apply grid-shock correction.

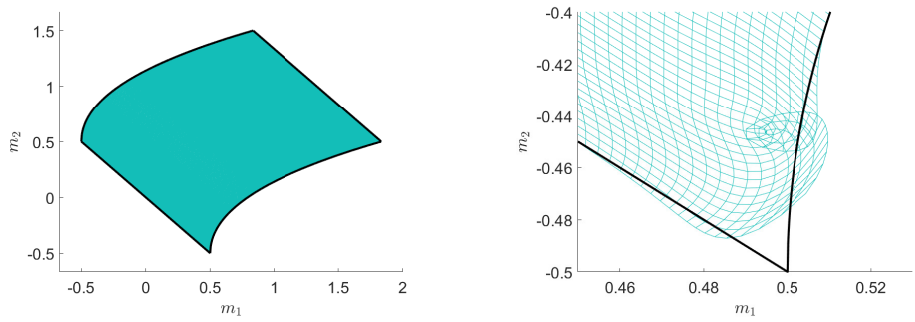


Figure 5.2: Example of grid shock. The global numerical approximation is shown on the left, and a zoomed-in version on the right.

For the gird-shock correction we make α in (5.8c) dependent on the coordinate, i.e., $\alpha = \alpha(\mathbf{x})$ and $\alpha_{ij} = \alpha(\mathbf{x}_{ij})$ and subsequently reduce α_{ij} on the boundary which puts more emphasis on the minimization of $|\mathbf{m}_{ij} - \mathbf{b}_{ij}|$. Introducing the $\mathbf{x}$-dependency, the coefficients α and $1 - \alpha$ in (5.8c) formally have to be moved inside the integrals of J_I and J_B. Doing so yields the functional

$$J(\mathbf{m}, \mathbf{P}, \mathbf{b}, \alpha) = \tfrac{1}{2} \iint_{\mathcal{X}} \alpha \|\mathrm{D}\mathbf{m} - \mathbf{P}\|^2 \,\mathrm{d}\mathbf{x} + \tfrac{1}{2} \oint_{\partial\mathcal{X}} (1 - \alpha)|\mathbf{m} - \mathbf{b}|^2 \,\mathrm{d}s. \tag{5.59}$$

We compute the first variation of $J(\mathbf{m}, \mathbf{P}, \mathbf{b}, \alpha)$ w.r.t. $\mathbf{m}$ to find

$$\delta J(\mathbf{m}, \mathbf{P}, \mathbf{b}, \alpha)(\boldsymbol{\eta}) = \iint_{\mathcal{X}} \alpha(\mathrm{D}\mathbf{m}) - \mathbf{P}) : \mathrm{D}\boldsymbol{\eta} \,\mathrm{d}\mathbf{x} + \oint_{\partial\mathcal{X}} (1 - \alpha)(\mathbf{m} - \mathbf{b}) \cdot \boldsymbol{\eta} \,\mathrm{d}\mathbf{s}. \tag{5.60}$$

We split the area integral as follows:

$$\iint_{\mathcal{X}} \alpha(\nabla(\mathbf{m}^\mathrm{T}) - \mathbf{P}) : \nabla(\boldsymbol{\eta}^\mathrm{T}) \,\mathrm{d}\mathbf{x} = \iint_{\mathcal{X}} \sum_{i=1}^{2} \alpha(\nabla m_i - \mathbf{p}_i) \cdot \nabla \eta_i \,\mathrm{d}\mathbf{x}. \tag{5.61}$$

Next, we have $\alpha(\nabla m_i - \mathbf{p}_i) \cdot \nabla \eta_i = \nabla \cdot (\alpha(\nabla m_i - \mathbf{p}_i)\eta_i) - \eta_i \nabla \cdot (\alpha(\nabla m_i - \mathbf{p}_i))$, such that by applying the divergence theorem to (5.61), we obtain

$$\begin{aligned} &\iint_{\mathcal{X}} \alpha(\mathrm{D}\mathbf{m} - \mathbf{P}) : \mathrm{D}\boldsymbol{\eta} \,\mathrm{d}\mathbf{x} = \\ &\sum_{i=1}^{2} \oint_{\partial\mathcal{X}} \alpha\eta_i(\nabla m_i - \mathbf{p}_i) \cdot \hat{\mathbf{n}} \,\mathrm{d}s - \sum_{i=1}^{2} \iint_{\mathcal{X}} \eta_i \nabla \cdot (\alpha(\nabla \mathbf{m}_i - \mathbf{p}_i)) \,\mathrm{d}\mathbf{x}. \end{aligned} \tag{5.62}$$

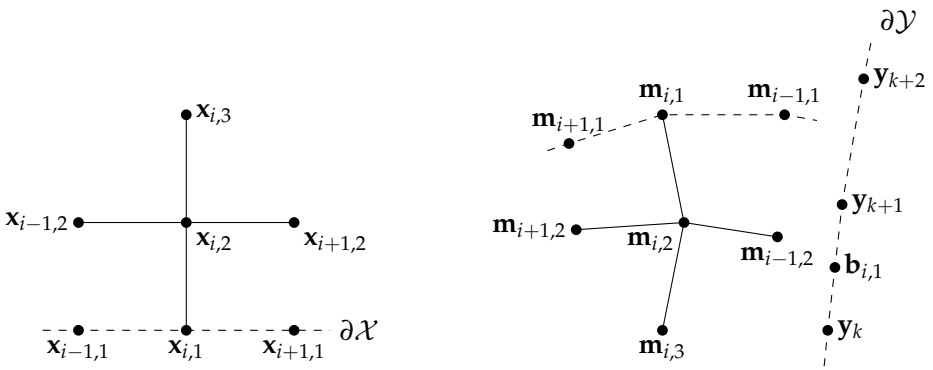


Figure 5.3: Schematic overview of the stencil used for detecting grid shocks. On the left and right the source and the target spaces are shown, respectively. On the right $\mathbf{m}_{i-1,2}$ (the mapping of an interior point) is closer to $\mathbf{b}_{i,1}$ than $\mathbf{m}_{i,1}$ (the mapping of a boundary point) is, so a grid shock occurs.

The minimum of $J(\mathbf{m}, \mathbf{P}, \mathbf{b}, \alpha)$ satisfies $\delta J = 0$. Choosing $\eta_2 = 0$ and applying the fundamental lemma of calculus of variations yields

$$\nabla\alpha \cdot \nabla m_1 + \alpha\Delta m_1 = \nabla\alpha \cdot \mathbf{p}_1 + \alpha\nabla \cdot \mathbf{p}_1, \qquad \mathbf{x} \in \mathcal{X}, \tag{5.63a}$$

$$(1-\alpha)m_1 + \alpha\nabla m_1 \cdot \hat{\mathbf{n}} = (1-\alpha)b_1 + \alpha\mathbf{p}_1 \cdot \hat{\mathbf{n}}, \qquad \mathbf{x} \in \partial\mathcal{X}, \tag{5.63b}$$

for the first component of $\mathbf{m}$. If α is constant in the interior, we have $\nabla\alpha = \mathbf{0}$ in the interior and equations (5.63) reduce to (5.18). By analogy, we have (5.19) for the second component of $\mathbf{m}$. Let $\alpha_1 \in (0,1)$. We set $\alpha(\mathbf{x}_{ij}) = \alpha_1$ for $\mathbf{x}_{ij}$ in the interior of $\mathcal{X}$. For a boundary point $\mathbf{x}_{ij} \in \partial\mathcal{X}$ we instead set

$$\alpha_{ij} = \begin{cases} \alpha_2, & \text{if } \min\limits_{(k,l)\in\mathcal{I}(i,j)} |\mathbf{m}_{kl} - \mathbf{b}_{ij}| < |\mathbf{m}_{ij} - \mathbf{b}_{ij}|, \\ \alpha_1 & \text{otherwise,} \end{cases} \tag{5.64}$$

where $\mathcal{I}(i,j) = \{(k,l) \mid \mathbf{x}_{kl} \in \text{int}(\mathcal{X}), |(k,l)^{\mathrm{T}} - (i,j)^{\mathrm{T}}| \leq 2\}$ is the index set over which we minimize, $\text{int}(\mathcal{X})$ the interior of $\mathcal{X}$ and $\alpha_2 \in (0, \alpha_1)$ a constant. We choose a distance of 2 and the values $\alpha_2 = 0.005$ and $\alpha_1 = 0.2$ since these work well in practice.

If in the n^{th} iteration we obtain $\alpha(\mathbf{x}) \not\equiv \alpha_1$, we solve (5.18) and (5.19) for a second time with the updated α to perform a correction.

Recall, we use a finite difference method for discretizing the Poisson equations (5.18) and (5.19), yielding a system of equations. This system of equations depends on α. Without grid-shock correction, a LU-factorization can be calculated once and used for each subsequent iteration which makes solving the system efficient. In case $\alpha_{ij} \neq \alpha_1$ for any (i,j), the same LU-factorization can no longer be used due to the component $\alpha\nabla m_k \cdot \hat{\mathbf{n}}$ in the Robin boundary conditions and a new LU-factorization has to be calculated for the iteration in which grid-shock correction is applied.

5.1.5 Curl-constrained minimization

By construction we have $\mathbf{m} = \nabla u$, and hence $\mathbf{m}$ is rotation free, i.e., $(\nabla \times \mathbf{m}) \cdot \hat{\mathbf{e}}_z = 0$. Due to the decoupling of $\mathbf{m}$ and u in the iterative algorithm, the rotation-free condition is in general not satisfied. Constraining the optimization of J such that $\mathbf{m}$ is rotation free is therefore a potential alternative to the grid-shock correction as introduced in the previous section, so here we choose $\alpha = \text{const}$. To constrain the optimization of J given by (5.8c), we minimize J over the space

$$\mathcal{M}_{\mathrm{R}} = \{\mathbf{m} \in [C^2(\mathcal{X})]^2 \mid (\nabla \times \mathbf{m}) \cdot \hat{\mathbf{e}}_z = 0\}, \tag{5.65}$$

instead of $\mathcal{M}$. We seek the optimum of J using Lagrange multipliers and calculus of variations. To this end, we define

$$\Lambda_{\mathrm{R}}(\mathbf{m}, \mathbf{P}, \mathbf{b}, \lambda_{\mathrm{R}}) = J(\mathbf{m}, \mathbf{P}, \mathbf{b}) + J_{\mathrm{R}}(\mathbf{m}, \lambda_{\mathrm{R}}), \tag{5.66a}$$

$$J_{\mathrm{R}}(\mathbf{m}, \lambda_{\mathrm{R}}) = \iint_{\mathcal{X}} \lambda_{\mathrm{R}} (\nabla \times \mathbf{m}) \cdot \hat{\mathbf{e}}_z \, \mathrm{d}\mathbf{x}, \tag{5.66b}$$

where $\lambda_{\mathrm{R}} = \lambda_{\mathrm{R}}(\mathbf{x})$ and where we assume $\mathbf{P}$ and $\mathbf{b}$ to be fixed. To ease computations, we introduce the symplectic matrix $\mathbf{J}$ such that

$$\mathbf{J} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad (\nabla \times \mathbf{m}) \cdot \hat{\mathbf{e}}_z = \nabla \cdot (\mathbf{J}\mathbf{m}). \tag{5.67}$$

The first variation of Λ_{R} with respect to $\mathbf{m}$ for $\eta \in [C^2(\mathcal{X})]^2$ is given by

$$\begin{aligned} \delta\Lambda_{\mathrm{R}}(\mathbf{m}, \mathbf{P}, \mathbf{b}, \lambda_{\mathrm{R}})(\boldsymbol{\eta}) &= \lim_{\epsilon \to 0} \frac{\Lambda_{\mathrm{R}}(\mathbf{m} + \epsilon\boldsymbol{\eta}, \mathbf{P}, \mathbf{b}, \lambda_{\mathrm{R}}) - \Lambda_{\mathrm{R}}(\mathbf{m}, \mathbf{P}, \mathbf{b}, \lambda_{\mathrm{R}})}{\epsilon} \\ &= \delta J(\mathbf{m}, \mathbf{P}, \mathbf{b})(\boldsymbol{\eta}) + \delta J_{\mathrm{R}}(\mathbf{m}, \lambda_{\mathrm{R}})(\boldsymbol{\eta}), \end{aligned} \tag{5.68}$$

where δJ and δJ_{R} denote the first variations of J and J_{R}, respectively. The first variation of J is given by (5.17) and reads

$$\begin{aligned} \delta J(\mathbf{m}, \mathbf{P}, \mathbf{b})(\boldsymbol{\eta}) = \oint_{\partial\mathcal{X}} \left[\alpha \begin{pmatrix} (\nabla m_1 - \mathbf{p}_1) \cdot \hat{\mathbf{n}} \\ (\nabla m_2 - \mathbf{p}_2) \cdot \hat{\mathbf{n}} \end{pmatrix} + (1 - \alpha)(\mathbf{m} - \mathbf{b}) \right] \cdot \boldsymbol{\eta} \, \mathrm{d}s \\ - \iint_{\mathcal{X}} \alpha \begin{pmatrix} \Delta m_1 - \nabla \cdot \mathbf{p}_1 \\ \Delta m_2 - \nabla \cdot \mathbf{p}_2 \end{pmatrix} \cdot \boldsymbol{\eta} \, \mathrm{d}\mathbf{x}. \end{aligned} \tag{5.69}$$

We calculate the first variation of J_{R}, viz.

$$\begin{aligned} \delta J_{\mathrm{R}}(\mathbf{m}, \lambda_{\mathrm{R}})(\boldsymbol{\eta}) &= \lim_{\epsilon \to 0} \frac{J_{\mathrm{R}}(\mathbf{m} + \epsilon\boldsymbol{\eta}, \lambda_{\mathrm{R}}) - J_{\mathrm{R}}(\mathbf{m}, \lambda_{\mathrm{R}})}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \iint_{\mathcal{X}} \left[\lambda_{\mathrm{R}} \nabla \cdot \Big(\mathbf{J}(\mathbf{m} + \epsilon\boldsymbol{\eta}) \Big) - \lambda_{\mathrm{R}} \nabla \cdot (\mathbf{J}\mathbf{m}) \right] \mathrm{d}\mathbf{x} \\ &= \iint_{\mathcal{X}} \lambda_{\mathrm{R}} \nabla \cdot (\mathbf{J}\boldsymbol{\eta}) \, \mathrm{d}\mathbf{x} \\ &= \iint_{\mathcal{X}} \Big(\nabla \cdot (\lambda_{\mathrm{R}} \mathbf{J}\boldsymbol{\eta}) - \mathbf{J}\boldsymbol{\eta} \cdot \nabla\lambda_{\mathrm{R}} \Big) \, \mathrm{d}\mathbf{x} \\ &\overset{\star}{=} \oint_{\partial\mathcal{X}} (\lambda_{\mathrm{R}} \mathbf{J}\boldsymbol{\eta}) \cdot \hat{\mathbf{n}} \, \mathrm{d}s - \iint_{\mathcal{X}} \mathbf{J}\boldsymbol{\eta} \cdot \nabla\lambda_{\mathrm{R}} \, \mathrm{d}\mathbf{x} \\ &= \oint_{\partial\mathcal{X}} \lambda_{\mathrm{R}} \mathbf{J}^{\mathrm{T}} \hat{\mathbf{n}} \cdot \boldsymbol{\eta} \, \mathrm{d}s - \iint_{\mathcal{X}} \mathbf{J}^{\mathrm{T}} \nabla\lambda_{\mathrm{R}} \cdot \boldsymbol{\eta} \, \mathrm{d}\mathbf{x}, \end{aligned} \tag{5.70}$$

where we used the divergence theorem at $\star$ and $\mathbf{J}^{\mathrm{T}} = \mathbf{J}$ for the last equality. So in total we have

$$\begin{aligned}\delta\Lambda_{\mathrm{R}}(\mathbf{m},\mathbf{P},\mathbf{b},\lambda_{\mathrm{R}})(\boldsymbol{\eta}) = &-\iint_{\mathcal{X}}\left[\alpha\begin{pmatrix}\Delta m_1 - \nabla\cdot\mathbf{p}_1\\ \Delta m_2 - \nabla\cdot\mathbf{p}_2\end{pmatrix} + \mathbf{J}^{\mathrm{T}}\nabla\lambda_{\mathrm{R}}\right]\cdot\boldsymbol{\eta}\,\mathrm{d}\mathbf{x}\\ &+\oint_{\partial\mathcal{X}}\left[\alpha\begin{pmatrix}(\nabla m_1 - \mathbf{p}_1)\cdot\hat{\mathbf{n}}\\ (\nabla m_2 - \mathbf{p}_2)\cdot\hat{\mathbf{n}}\end{pmatrix} + (1-\alpha)(\mathbf{m}-\mathbf{b}) + \lambda_{\mathrm{R}}\mathbf{J}^{\mathrm{T}}\hat{\mathbf{n}}\right]\cdot\boldsymbol{\eta}\,\mathrm{d}s\end{aligned} \tag{5.71}$$

For $\mathbf{m}$ to be a minimizer, we have $\delta\Lambda_{\mathrm{R}} = 0$ for all $\boldsymbol{\eta} \in [C^2(\mathcal{X})]^2$. Choosing $\eta_2 = 0$ and applying the fundamental lemma of calculus of variations we obtain

$$\alpha\Delta m_1 = \alpha\nabla\cdot\mathbf{p}_1 + \lambda_{\mathrm{R},x_2}, \qquad (x_1,x_2)\in\mathcal{X}, \tag{5.72a}$$

$$(1-\alpha)m_1 + \alpha\nabla m_1\cdot\hat{\mathbf{n}} = (1-\alpha)b_1 + \alpha\mathbf{p}_1\cdot\hat{\mathbf{n}} + \lambda_{\mathrm{R}}n_2, \qquad (x_1,x_2)\in\partial\mathcal{X}, \tag{5.72b}$$

where λ_{R,x_2} is the derivative of λ_{R} w.r.t. x_2. Similarly, choosing $\eta_1 = 0$ and applying the fundamental lemma of calculus of variations, we find

$$\alpha\Delta m_2 = \alpha\nabla\cdot\mathbf{p}_2 - \lambda_{\mathrm{R},x_1}, \qquad (x_1,x_2)\in\mathcal{X}, \tag{5.73a}$$

$$(1-\alpha)m_2 + \alpha\nabla m_2\cdot\hat{\mathbf{n}} = (1-\alpha)b_2 + \alpha\mathbf{p}_2\cdot\hat{\mathbf{n}} - \lambda_{\mathrm{R}}n_1, \qquad (x_1,x_2)\in\partial\mathcal{X}. \tag{5.73b}$$

equations (5.72) and (5.73) are two Poisson equations in the three (independent) variables m_1, m_2 and λ_{R}. To obtain a third equation, we calculate the first variation of Λ_{R} with respect to λ_{R} for $\eta \in [C^2(\mathcal{X})]$, yielding

$$\begin{aligned}\delta J_{\mathrm{R}}(\mathbf{m},\lambda_{\mathrm{R}})(\eta) &= \lim_{\epsilon\to 0}\frac{J_{\mathrm{R}}(\mathbf{m},\lambda_{\mathrm{R}}+\epsilon\eta) - J_{\mathrm{R}}(\mathbf{m},\lambda_{\mathrm{R}})}{\epsilon}\\ &= \iint_{\mathcal{X}}(\lambda_{\mathrm{R}}+\epsilon\eta)\nabla\cdot(\mathbf{J}\mathbf{m}) - \lambda_{\mathrm{R}}\nabla\cdot(\mathbf{J}\mathbf{m})\,\mathrm{d}\mathbf{x}\\ &= \iint_{\mathcal{X}}\eta\nabla\cdot(\mathbf{J}\mathbf{m})\,\mathrm{d}\mathbf{x}.\end{aligned} \tag{5.74}$$

Applying the fundamental lemma of calculus of variations yields the original constraint

$$\nabla\cdot(\mathbf{J}\mathbf{m}) = 0, \qquad (x_1,x_2)\in\mathcal{X}. \tag{5.75}$$

In order to solve for m_1, m_2 and λ_{R}, we formulate an additional equation by using (5.75), (5.72) and (5.73). Conceptually we apply (5.75) to the vector of

equations (5.72), (5.73), or more precisely, we differentiate (5.19) w.r.t. x_1 and subtract (5.18) after differentiation it w.r.t. x_2 to obtain

$$\alpha\Delta(\partial_{x_1} m_2 - \partial_{x_2} m_1) = \alpha\nabla \cdot (\partial_{x_1}\mathbf{p}_2 - \partial_{x_2}\mathbf{p}_1) - \lambda_{\mathrm{R},x_1x_1} - \lambda_{\mathrm{R},x_2x_2}, \quad (x_1, x_2) \in \mathcal{X}. \tag{5.76}$$

Because $\mathbf{P}$ is symmetric by construction, $p_{12} = p_{21}$ and equation (5.76) reads

$$\Delta\lambda_{\mathrm{R}} = \alpha(\partial_{x_1x_1} p_{12} - \partial_{x_2x_2} p_{12} + \partial_{x_1x_2} p_{22} - \partial_{x_1x_2} p_{11}) - \alpha\Delta(\partial_{x_1} m_2 - \partial_{x_2} m_1), \quad (x_1, x_2) \in \mathcal{X}. \tag{5.77a}$$

We obtain a boundary condition for (5.77a) by subtracting n_1 times (5.19) from n_2 times (5.18), viz.

$$\begin{aligned}\lambda_{\mathrm{R}} = {} & (1-\alpha)(n_2(m_1 - b_1) - n_1(m_2 - b_2)) \\ & + \alpha(n_2(\nabla m_1 - \mathbf{p}_1) - n_1(\nabla m_2 - \mathbf{p}_2)) \cdot \hat{\mathbf{n}}, \qquad (x_1, x_2) \in \partial\mathcal{X}.\end{aligned} \tag{5.77b}$$

Note, in case $\mathbf{P}$ and $\mathbf{m}$ are exact we have

$$\begin{aligned}\partial_{x_1x_1} p_{12} &= u_{x_1x_1x_1x_2}, \quad \partial_{x_2x_2} p_{12} = u_{x_1x_2x_2x_2}, \\ \partial_{x_1x_2} p_{22} &= u_{x_1x_2x_2x_2}, \quad \partial_{x_1x_2} p_{11} = u_{x_1x_1x_1x_2},\end{aligned} \tag{5.78}$$

and equation (5.77a) reads

$$\Delta\lambda_{\mathrm{R}} = 0, \qquad (x_1, x_2) \in \mathcal{X}. \tag{5.79}$$

If furthermore $\mathbf{b}(x_1, x_2) = \mathbf{m}(x_1, x_2)$ for $(x_1, x_2) \in \partial\mathcal{X}$, then

$$\lambda_{\mathrm{R}} = 0, \qquad (x_1, x_2) \in \partial\mathcal{X}, \tag{5.80}$$

and consequently $\lambda_{\mathrm{R}} \equiv 0$ on the whole of $\mathcal{X}$ by the maximum principle for the Laplace equation. Under the above assumptions, $\mathbf{m}$ is exact and thus the constraint $\nabla \cdot (\mathbf{J}\mathbf{m}) = 0$ is automatically satisfied and, as expected, equations (5.72) and (5.73) reduce to the PDEs (5.18) and (5.19) because $\lambda_{\mathrm{R}} \equiv 0$.

Equations (5.72), (5.73) and (5.77) form a system of coupled PDEs. To solve these we employ an iterative scheme. First, we assume m_1 and m_2 to be given and solve (5.77) using standard second-order finite differences (FD). The solution λ_{R} is then regarded as a given function and we solve (5.18) and (5.19) sequentially by a similar FD scheme. This process is repeated until m_1 and m_2 no longer change significantly. Solving (5.77) the first time requires an initial guess for m_1 and m_2. For this we use $\mathbf{m}$ from the previous iteration, i.e., $\mathbf{m}^n$ given by (5.9c).

Correcting grid shocks by constraining the curl of $\mathbf{m}$ has experimentally proven to work for some, but not all cases. The grid-shock correction, as presented in Section 5.1.4, has experimentally shown to work better, therefore we do not present numerical results for constraining the curl of $\mathbf{m}$.

5.2 Numerical results

In this section we present numerical results for five examples. For each example we know the exact solution and compare the numerical methods. We choose $\mathcal{X} = [x_1^m, x_1^M] \times [x_2^m, x_2^M]$, a rectangle which may vary per case. For each example we choose $\mathcal{Y}$ such that it has a unique feature to it. We measure the residual

$$\epsilon_r = \left| D_{x_1}[m_1]_{ij} D_{x_2}[m_2]_{ij} - D_{x_1}[m_2]_{ij} D_{x_2}[m_1]_{ij} + f^2(\mathbf{x}_{ij}, \mathbf{m}_{ij}) \right|_\infty, \tag{5.81}$$

with D_{x_1} and D_{x_2} standard second-order (central in interior and one-sided on boundary) finite difference operators for the first-order derivatives with respect to x_1 and x_2, respectively. Furthermore, we measure the global discretization errors ϵ_u, ϵ_{m_1} and ϵ_{m_2} defined by

$$\begin{aligned} \epsilon_u &= \left| \left(u_{ij} - u_{11}\right) - \left(u(\mathbf{x}_{ij}) - u(\mathbf{x}_{11})\right) \right|_\infty, \\ \epsilon_{m_1} &= \left| (m_1)_{ij} - m_1(\mathbf{x}_{ij}) \right|_\infty, \\ \epsilon_{m_2} &= \left| (m_2)_{ij} - m_2(\mathbf{x}_{ij}) \right|_\infty, \end{aligned} \tag{5.82}$$

where the terms u_{11} and $u(\mathbf{x}_{11})$ are introduced due to the nonuniqueness of u given $\mathbf{m}$; see the discussion following equations (5.22). Any fixed grid point could be used, here $\mathbf{x}_{11}$ is chosen. The choice for the ∞-norm is arbitrary in the sense that any standard norm would give similar results. However, the ∞-norm is more sensitive to differences in the errors than, for example, the standard 2-norm. Starting the least-squares algorithm requires an initial guess $\mathbf{m}^0$, so we introduce $\widetilde{\mathcal{Y}} = [y_1^m, y_1^M] \times [y_2^m, y_2^M]$, the smallest bounding box of $\mathcal{Y}$. We then choose $\mathbf{m}^0(\partial\mathcal{X}) = \partial\widetilde{\mathcal{Y}}$, such that $\mathbf{m}_{ij}^0$ is equidistantly distributed, i.e., $\mathbf{m}_{ij}^0$ is the result of a bilinear uniform interpolation of the bounding box of $\mathcal{Y}$ with $\det(\mathrm{D}\mathbf{m}^0) < 0$. The initial guess then reads

$$(m_1^0)_{ij} = \frac{(x_1)_{ij} - x_1^m}{x_1^M - x_1^m} y_1^M + \frac{x_1^M - (x_1)_{ij}}{x_1^M - x_1^m} y_1^m, \tag{5.83a}$$

$$(m_2^0)_{ij} = \frac{x_2^M - (x_2)_{ij}}{x_2^M - x_2^m} y_2^M + \frac{(x_2)_{ij} - x_2^m}{x_2^M - x_2^m} y_2^m. \tag{5.83b}$$

The initial guess is a (discretized) solution of the hyperbolic Monge-Ampère equation with $f^2 = \mathrm{area}(\widetilde{\mathcal{Y}})/\mathrm{area}(\mathcal{X})$, $\mathbf{m} = \nabla u$ and $u = \frac{1}{2}(x^2 - y^2)f$. Three more such initial guesses exist, viz., $u = \frac{1}{2}(y^2 - x^2)f$ and $u = \pm xyf$.

We cut the source boundary in segments, clockwise, according to

$$\begin{aligned} \Gamma_1^{\mathcal{X}} &= \{x_1^m\} \times [x_2^m, x_2^M], & \Gamma_2^{\mathcal{X}} &= [x_1^m, x_1^M] \times \{x_2^M\}, \\ \Gamma_3^{\mathcal{X}} &= \{x_1^M\} \times [x_2^m, x_2^M], & \Gamma_4^{\mathcal{X}} &= [x_1^m, x_1^M] \times \{x_2^m\}, \end{aligned} \tag{5.84}$$

and we write

$$\Gamma_k^{\mathcal{Y}} = \left\{ \mathbf{y}_k(s) \,\middle|\, s \in [0,1] \right\}, \tag{5.85}$$

for $k = 1, \ldots, 4$, i.e., for fixed k the segment $\Gamma_k^{\mathcal{Y}}$ is parametrized counter clockwise by $\mathbf{y}_k(s)$ and formally takes on the role of $\mathbf{y}_k(s)$ as introduced in Section 5.1.3. Furthermore, we apply grid-shock correction only for iteration step $n \geq 100$, as the distance between the boundary of the initial guess and the boundary of the target may be large for small n.

Lastly, we stop the iteration (5.9) based on the update of $\mathbf{m}^n$, i.e., based on

$$\Delta m^n = |\mathbf{m}^n - \mathbf{m}^{n-1}|, \tag{5.86}$$

instead of, the already introduced measures, J_{I} and J_{B}. This is because the values for J_{I} and J_{B} may stagnate over the iterations while Δm^n is still changing. Conversely, if Δm^n has stagnated, then so have the functionals J_{I} and J_{B}. We stop the iterative process when Δm^n reaches floating-point precision.

5.2.1 Example 1: Annulus segment

For the first example we consider $\mathcal{X} = [0,1] \times [-1/2, 1/2]$, $\partial\mathcal{Y} = \cup_{k=1}^4 \Gamma_k^{\mathcal{Y}}$ with

$$\mathbf{y}_1(s) = (\cos(\tfrac{1}{2} - s), \sin(\tfrac{1}{2} - s)), \tag{5.87a}$$
$$\mathbf{y}_2(s) = (e^s \cos(\tfrac{1}{2}), -e^s \sin(\tfrac{1}{2})), \tag{5.87b}$$
$$\mathbf{y}_3(s) = (e \cos(\tfrac{1}{2} - s), e \sin(s - \tfrac{1}{2})), \tag{5.87c}$$
$$\mathbf{y}_4(s) = (e^{1-s} \cos(\tfrac{1}{2}), e^{1-s} \sin(\tfrac{1}{2})), \tag{5.87d}$$

as shown together with the exact mapping on a 21×21 grid in Figure 5.4 on the left. We choose $\Gamma_k^{\mathcal{Y}} = \nabla u(\Gamma_k^{\mathcal{X}})$ for all examples. Furthermore, this choice of $\Gamma_k^{\mathcal{Y}}$ implies that for SALM and SPM the corners of $\partial\mathcal{X}$ are mapped to the corners of $\partial\mathcal{Y}$. Let $f^2(x_1, x_2) = e^{2x_1}$. The solution is then given by

$$u(x_1, x_2) = e^{x_1} \cos(x_2), \tag{5.88}$$

which is symmetric in $x_2 = 0$ as can be seen in Figure 5.4. Unless specified otherwise, we take $N_{\mathrm{b}} = 10^4$ and for each target segment $\Gamma_k^{\mathcal{Y}}$, with $k = 1, \ldots, 4$, we construct $\mathbf{y}_i = \mathbf{y}_k(s_i)$ with $s_i = (i-1)/(N_{\mathrm{b}} - 1)$ and $i = 1, \ldots, N_{\mathrm{b}}$. The results for PM, SPM and SALM are shown in Figure 5.5 for varying grid

configurations with $N_{x_1} = N_{x_2}$. The three figures clearly show second-order convergence of the relevant errors and residual, which is in accordance with the discretization error of the finite difference approximations used. In terms of ϵ_u, PM and SPM ($3 \cdot 10^{-6}$) slightly outperform SALM ($6 \cdot 10^{-6}$) for $N_{x_1} = N_{x_2} = 473$, though the difference is small. Figure 5.6 shows the J-errors over the iterations on a grid of $N_{x_1} = N_{x_2} = 473$. For PM and SPM we obtained $J_{\mathrm{I}} \approx 2 \cdot 10^{-12}$, $J_{\mathrm{B}} \approx 3 \cdot 10^{-13}$ in approximately 60,000 iterations. SALM gave $J_{\mathrm{I}} \approx 4 \cdot 10^{-12}$, $J_{\mathrm{B}} \approx 7 \cdot 10^{-13}$ in 40,000, iterations. SALM consistently requires less iterations as seen on the left in Figure 5.7, where the number of required iterations ($n_{\max}$) for various $N_{x_1} = N_{x_2}$ is shown.

Convergence of u with respect to N_{b} is shown on the right of Figure 5.7 for $N_{x_1} = N_{x_2} = 473$. Two observations are in place. First, for increasing N_{b}, the error ϵ_u reaches an asymptotic value (dashed black line). This phenomenon is to be expected and occurs when the discretization errors in $\mathbf{m}_{ij}$, $\mathbf{P}_{ij}$ and u_{ij}, and the finite differences $\mathbf{D}_{ij}$ become dominant, i.e., when the discretization errors due to the choice of N_{x_1} and N_{x_2} dominate the errors due to discretizing the boundary. Secondly, in the regime prior to the asymptote, the discretization error in u due to the boundary discretization is second-order accurate for all three boundary methods.

Finally, the computational cost per iteration for SALM is lowest, second comes SPM and third PM. The projection methods calculate $C \max(N_{x_1}, N_{x_2}) N_{\mathrm{b}}$ projections and performs $C \max(N_{x_1}, N_{x_2})$ searches over N_{b} points each, with $C \in \mathbb{N}_+$. Similarly, SALM performs a linear interpolation of $C \max(N_{x_1}, N_{x_2})$ points over $N_{\mathrm{b}} - 1$ segments. Therefore, one would expect the average time per iteration to scale linearly in N_x, N_y and N_{b} for PM and SPM when $N_{x_1} = N_{x_2}$, and N_{b} is fixed. For SALM, a linear relation is also expected, with a pos-

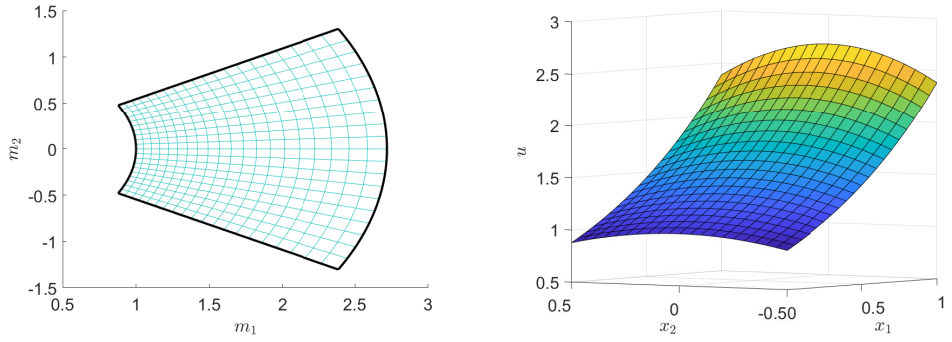


Figure 5.4: The exact mapping $\mathbf{m}$ (left) and solution u (right) on a 21×21 grid for Example 1 (Annulus segment).

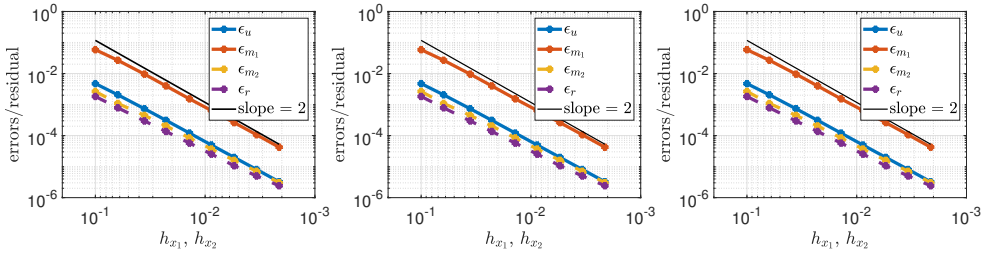


Figure 5.5: Global error and the residual for PM (left), SPM (middle) and SALM (right).

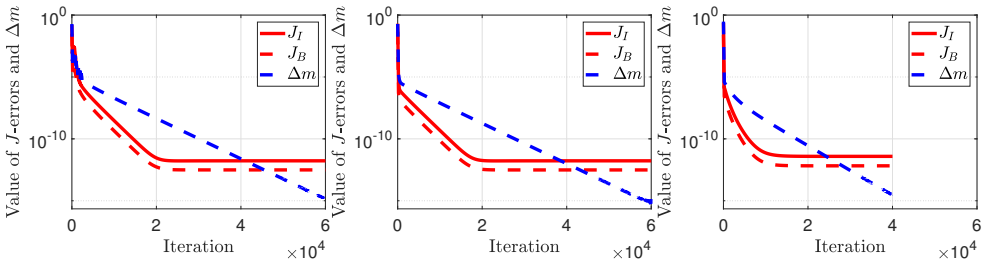


Figure 5.6: J-errors and Δm over the iterations for PM (left), SPM (middle) and SALM (right).

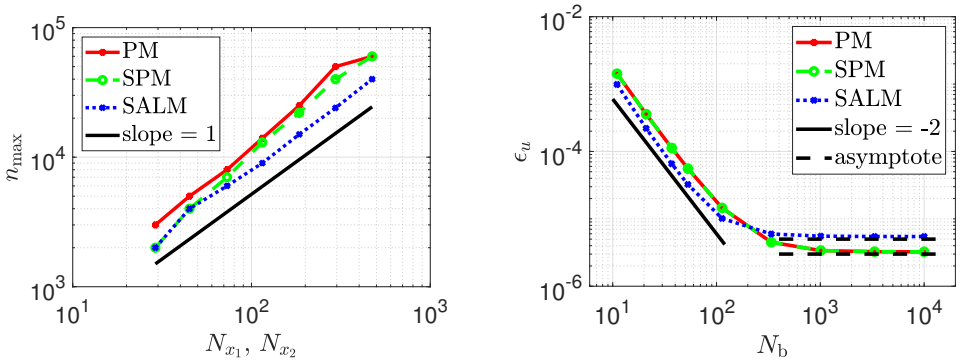


Figure 5.7: The total number of iterations needed for convergence (left) and the influence of the boundary discretization on ϵ_u (right).

sible asymptote when either the computational load due to $\max(N_{x_1}, N_{x_2})$ or N_b dominates. This is also shown in Figure 5.8, where, from left to right, $N_{x_1} = N_{x_2} = 501$ is fixed while N_b varies, $N_b = 37$ is fixed and $N_{x_1} = N_{x_2}$ varies, and lastly, $N_b = 10,007$ is fixed and $N_{x_1} = N_{x_2}$ varies. Additionally, it is observed that SALM, on average, significantly outperforms the projection methods. Lastly, SPM is approximately four times faster than PM because SPM projects one source segment on a target segment (four times) instead of the whole source boundary on the whole target boundary.

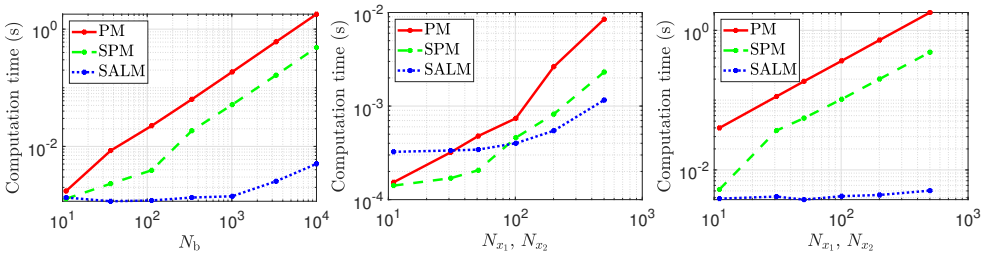


Figure 5.8: Timing results in seconds for the boundary procedures. On the left, $N_{x_1} = N_{x_2} = 501$ is fixed and in the middle and on the right $N_b = 37$ and $N_b = 10,007$, respectively.

5.2.2 Example 2: Deformed square

As a second example we consider a mapping and surface with no symmetries, viz. $\mathcal{X} = [0,1] \times [-1/2, 1/2]$, $\partial\mathcal{Y} = \cup_{k=1}^{4} \Gamma_k^{\mathcal{Y}}$ with

$$\mathbf{y}_1(s) = (s - \tfrac{1}{2},\ -s + \tfrac{1}{2}), \tag{5.89a}$$
$$\mathbf{y}_2(s) = (\tfrac{1}{3}s^3 + s^2 + \tfrac{1}{2},\ s - \tfrac{1}{2}), \tag{5.89b}$$
$$\mathbf{y}_3(s) = (-s + \tfrac{11}{6},\ s + \tfrac{1}{2}), \tag{5.89c}$$
$$\mathbf{y}_4(s) = (-\tfrac{1}{3}s^3 + 2s^2 - 3s + \tfrac{5}{6},\ \tfrac{3}{2} - s), \tag{5.89d}$$

as shown on the left of Figure 5.9. Let $f^2(x_1, x_2) = (x_1 + 1)^2$, then the exact solution is given by

$$u(x_1, x_2) = \tfrac{1}{12}x_1^4 + \tfrac{1}{3}x_1^3 + x_1 x_2 - \tfrac{1}{2}x_2^2, \tag{5.90}$$

which is shown on the right of Figure 5.9. We start with the results for PM. In Figure 5.10 the errors J_I and J_B are shown, both with grid-shock correction (left) and without (right). Clearly the example with grid-shock correction does not converge, the method actually oscillates between intermediate solutions. One may be tempted to think that without grid-shock correction the method does work, as Δm goes to machine precision, but this is not the case as shown in Figure 5.11. The two leftmost figures show the mapping for a 29×29 grid after 50,000 iterations. Clearly, neither of the methods work as intended as there are gaps between the mesh spanned by $\mathbf{m}_l$ and $\partial\mathcal{Y}$, i.e., the transport boundary condition has not been satisfied. The reason why the algorithm with PM does not converge is that the projection of $\mathbf{m}$ onto $\partial\mathcal{Y}$ does not distribute $\mathbf{b}$ well. In particular, no points $\mathbf{b}_l$ near $(0.5, -0.5) \in \partial\mathcal{Y}$ are obtained, as can be seen in Figure 5.11 on the right, where the blue circles represent $\mathbf{m}_l$, the red

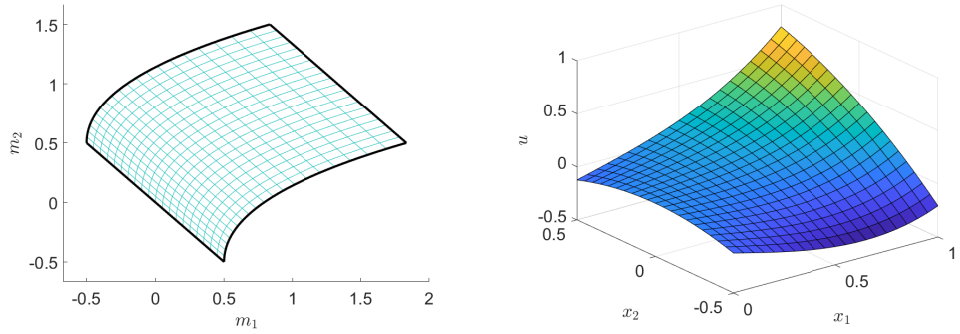


Figure 5.9: The exact mapping **m** (left) and solution u (right) on a 21×21 grid for Example 2 (Deformed square).

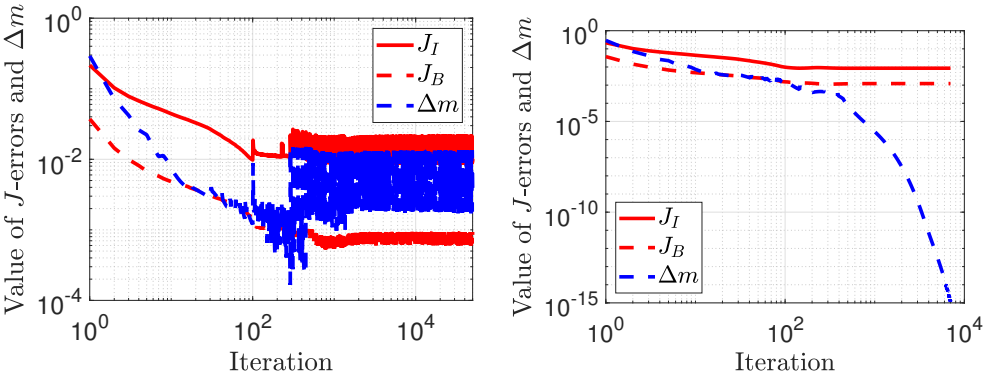


Figure 5.10: The errors J_I, J_B and the update Δm for PM with shock correction (left) and without (right).

squares $\mathbf{b}_l$ and the thin black lines connect $\mathbf{m}_l$ to $\mathbf{b}_l$ for $l = 1, \ldots, N$.

In Figure 5.12 errors and residuals are shown for SPM (left) and SALM (right) for varying grid configurations with $N_{x_1} = N_{x_2}$. Both figures show second-order convergence which is in accordance with the discretization errors of the finite difference approximations used.

Figure 5.13 shows the behavior of J_I, J_B and Δm for SPM and SALM. For SPM, on the left, Δm exhibit oscillations starting at 100 iterations. This is due to the grid-shock correction, enabled in the 100^{th} iteration. As it turns out, this is one example where grid shocks occur using SPM. Without the grid-shock correction, Δm would still go to computer precision but the grid shock (as visualized on the right of Figure 5.2) would remain and subsequently J_I, J_B and the errors ϵ_u, ϵ_{m_1} and ϵ_{m_2} and the residual ϵ_r would be three orders of magnitude higher.

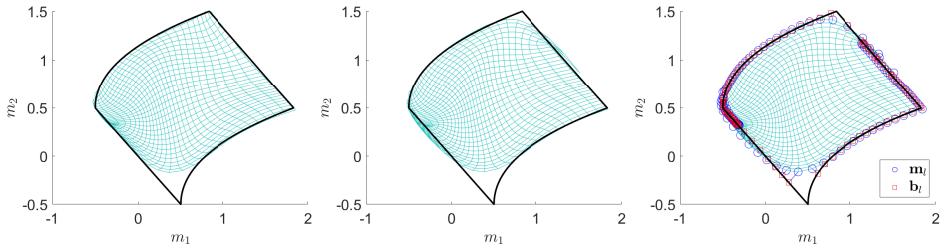


Figure 5.11: The mapping after 50,000 iterations for PM with shock correction (left) and without (middle) and the accompanying projection of $\mathbf{m}_l$ onto $\partial\mathcal{Y}$ for construction of $\mathbf{b}_l$ without shock correction on the right.

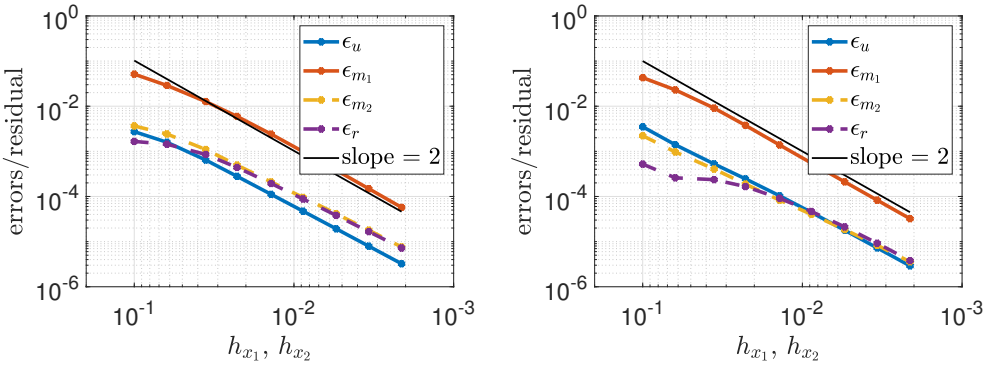


Figure 5.12: Convergence of the global error and the residual for SPM (left) and SALM (right).

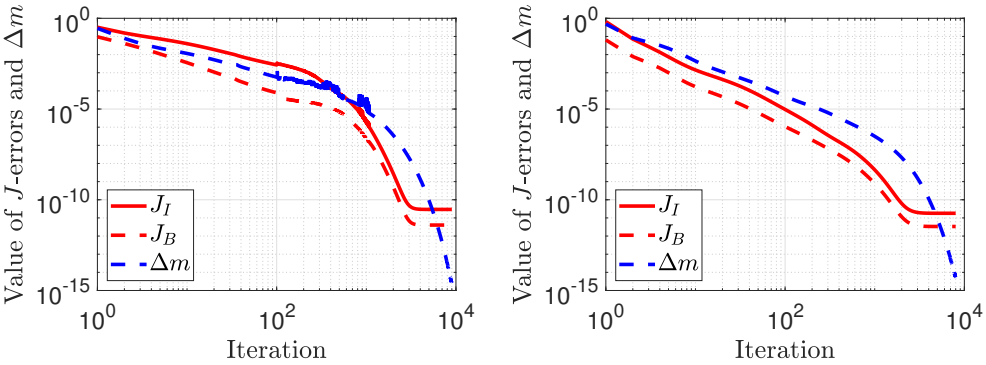


Figure 5.13: Comparison of J_{I}, J_{B} and Δm for SPM (left) and SALM (right) for a grid of 295×295.

5.2.3 Example 3: Inward fold

For this example we consider the target as illustrated in Figure 5.14, for the exact solution on a 61×61 grid. On the right a zoomed-in version of the target is shown. In the figure we have marked two points, one by a solid circle, and one by an asterisk. The former is a point for which the boundary of the target is not differentiable, while for the latter it is. We will come back to this. The example shown in Figure 5.14 corresponds to $\mathcal{X} = [0,1] \times [-1/2, 1/2]$ and $\partial\mathcal{Y} = \cup_{k=1}^{4} \Gamma_k^{\mathcal{Y}}$ with

$$\mathbf{y}_1(s) = (0, -\tfrac{1}{4}s^4 + \tfrac{1}{2}s^3 - \tfrac{3}{8}s^2 - \tfrac{7}{8}s + \tfrac{31}{64}), \tag{5.91a}$$
$$\mathbf{y}_2(s) = (\tfrac{9}{8}s - \tfrac{1}{2}s^3, -\tfrac{1}{4}s^4 + \tfrac{3}{8}s^2 - \tfrac{33}{64}), \tag{5.91b}$$
$$\mathbf{y}_3(s) = (-s^3 + \tfrac{3}{2}s^2 + \tfrac{1}{4}s + \tfrac{5}{8}, -\tfrac{1}{4}s^4 + \tfrac{1}{2}s^3 + \tfrac{9}{8}s^2 - \tfrac{3}{8}s - \tfrac{25}{64}), \tag{5.91c}$$
$$\mathbf{y}_4(s) = (-\tfrac{1}{2}s^3 + \tfrac{3}{2}s^2 - \tfrac{19}{8}s + \tfrac{11}{8}, -\tfrac{1}{4}s^4 + s^3 - \tfrac{9}{8}s^2 + \tfrac{1}{4}s + \tfrac{39}{64}). \tag{5.91d}$$

Furthermore we have

$$f^2(x_1, x_2) = x_1^6 + 3x_1^4x_2^2 + 3x_1^2x_2(x_2^3 - 2) + (1 + x_2^3)^2, \tag{5.92a}$$
$$u(x_1, x_2) = \frac{x_1^2}{2} - \frac{x_1^4x_2}{4} - \frac{x_2^2}{2} + \frac{x_1^2x_2^3}{2} - \frac{x_2^5}{20}. \tag{5.92b}$$

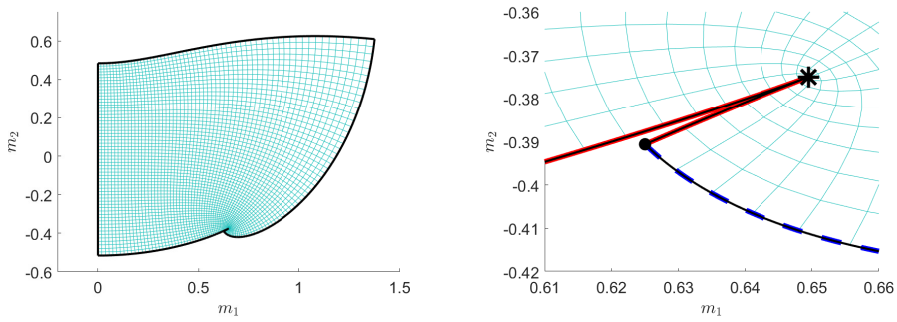


Figure 5.14: The target domain and mapping, with a zoomed-in version for Example 3 (Inward fold).

Taking derivatives of u yields the mapping, i.e.,

$$m_1(x_1, x_2) = u_{x_1}(x_1, x_2) = x_1 - x_1^3 x_2 + x_1 x_2^3, \tag{5.93a}$$

$$m_2(x_1, x_2) = u_{x_2}(x_1, x_2) = -\frac{x_1^4}{4} - x_2 + \frac{3x_1^2 x_2^2}{2} - \frac{x_2^4}{4}. \tag{5.93b}$$

A straightforward calculation shows that $\mathbf{m}(1, 1/2) = (5/8, -25/64)$ which corresponds to the solid circle in Figure 5.14. Henceforth, $\mathbf{m}$ is not differentiable in the point $(1, 1/2)$ as it is the image of a nondifferentiable (corner) point in $\mathcal{X}$ under a continuously differentiable map. The point depicted by the asterisk originates from the source boundary segment $\Gamma_2^{\mathcal{X}} = [0, 1] \times \{\frac{1}{2}\}$. Let m_1 and m_2 along the boundary be parametrized by s. Then in the point indicated by the asterisk, both $\frac{\mathrm{d}m_1(s)}{\mathrm{d}s}$ and $\frac{\mathrm{d}m_2(s)}{\mathrm{d}s}$ change sign. Henceforth, the location of the asterisk can be obtained by solving $\frac{\mathrm{d}m_1(s)}{\mathrm{d}s} = \frac{\mathrm{d}m_2(s)}{\mathrm{d}s} = 0$, which is equivalent to

$$\left.\frac{\partial m_1}{\partial x_1}\right|_{x_2=1/2} = 0, \qquad \left.\frac{\partial m_2}{\partial x_1}\right|_{x_2=\frac{1}{2}} = 0, \qquad 0 \leq x_1 \leq 1. \tag{5.94}$$

Indeed, doing so one obtains the unique solution $x_1 = \sqrt{3}/2$ such that $\mathbf{m}(\sqrt{3}/2, 1/2) = (3\sqrt{3}/8, -3/8)$, which corresponds to the point indicated by an asterisk in Figure 5.14. Furthermore, smoothness of the boundary in said point is implied.

PM and SPM do not yield converging numerical approximations. Figure 5.15 shows a zoomed-in version of two numerical solutions for 161×161 grids. The sharp inward fold seems to be the culprit for the boundary method, as is also seen in Figure 5.16, which shows the projection of $\mathbf{m}_l$ onto $\partial\mathcal{Y}$. The figure clearly shows that the method does not pick $\mathbf{b}_l$ deep within the fold and, consequently, the optimization for $\mathbf{m}$ does not produce a mapping with such a sharp fold.

Because SALM does force points $\mathbf{b}_l$ to be located along the whole boundary, naturally points will end up in the fold. This can be seen in Figure 5.17, where on the left the first iteration of applying SALM to the result of SPM is shown. The 50^{th} iteration of SALM is shown on the right, showing $\mathbf{m}(\partial\mathcal{X})$ being positioned in the fold. Continuation using SALM yields similar results to using SALM starting from the default initial guess. SALM shows approximately second-order convergence, as shown in Figure 5.18 when using (5.83) as initial guess. SALM does not show any visual distortions, in contrast to PM and SPM.

For the remaining results we will not discuss PM, as it performs, at best, as good as SPM while being more computationally expensive.

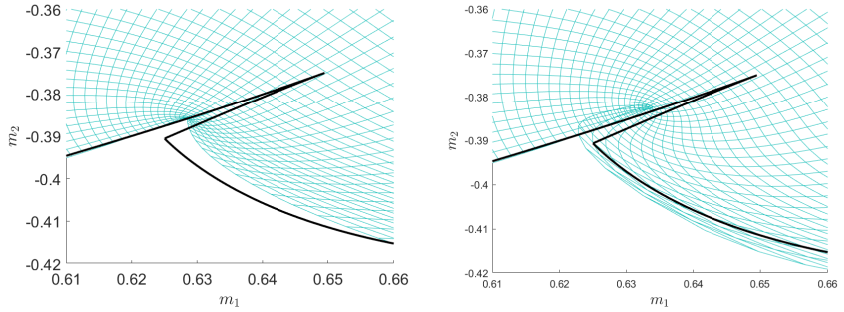


Figure 5.15: Zoomed-in results for PM (left) and SPM (right) on a 161 $\times$ 161 grid.

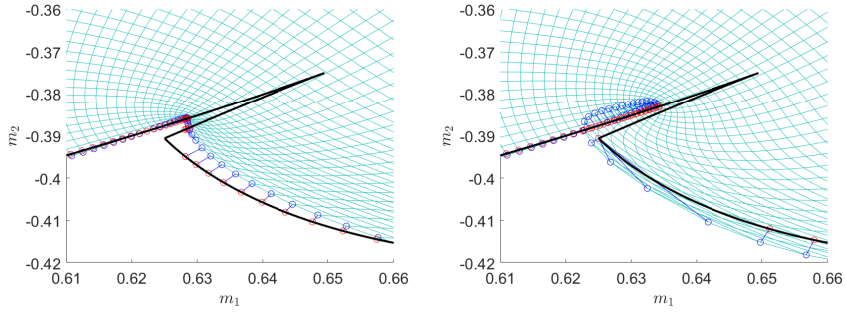


Figure 5.16: Projection step after convergence for PM (left) and for SPM (right) on a 161 $\times$ 161 grid.

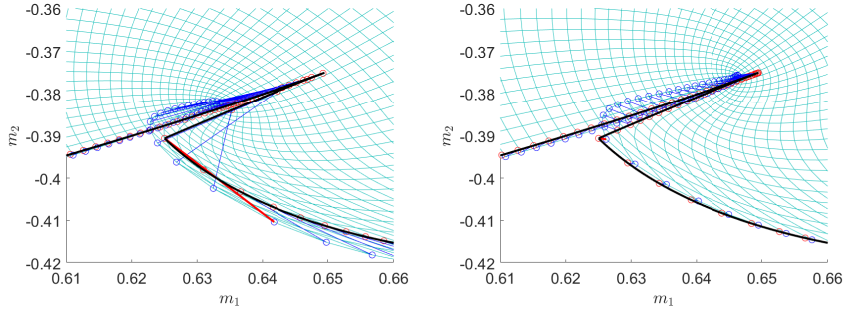


Figure 5.17: The first (left) and 50^{th} (right) iteration of continuation by SALM after SPM has converged on a 161 $\times$ 161 grid.

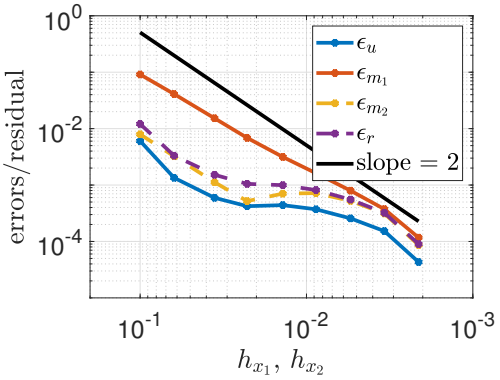


Figure 5.18: Convergence of the errors and residual for SALM starting from a uniform initial guess.

5.2.4 Example 4: Annulus

For this example we consider the target given in Figure 5.19, where the central part near $\mathbf{m} = (0,0)$ is not part of $\mathcal{Y}$. Let $\mathcal{X} = [0, 2\pi] \times [-1/2, 1/2]$, $\partial\mathcal{Y} = \cup_{k=1}^{4} \Gamma_k^{\mathcal{Y}}$ with

$$\mathbf{y}_1(s) = (0,\, e^{s-\frac{1}{2}}), \tag{5.95a}$$
$$\mathbf{y}_2(s) = \sqrt{e}(-\sin(2\pi s),\, \cos(2\pi s)), \tag{5.95b}$$
$$\mathbf{y}_3(s) = (0,\, e^{\frac{1}{2}-s}), \tag{5.95c}$$
$$\mathbf{y}_4(s) = \frac{1}{\sqrt{e}}(\sin(2\pi s),\, \cos(2\pi s)), \tag{5.95d}$$

and $f^2(x_1, x_2) = e^{2x_2}$, such that the exact solution is given by

$$u(x_1, x_2) = e^{x_2}\cos(x_1), \tag{5.96}$$

as shown on the right of Figure 5.19. Observe that $\mathbf{m}|_{\partial\mathcal{X}}$ is not bijective, as $\Gamma_1^{\mathcal{Y}} = \Gamma_3^{\mathcal{Y}}$. Nevertheless, we introduce both $\Gamma_1^{\mathcal{Y}}$ and $\Gamma_3^{\mathcal{Y}}$ as the orientation, i.e., the parametrization of the segments matters for SALM.

Figure 5.20 shows results for SPM. On the left the mapping after the algorithm has converged for a grid with $N_{x_1} = 115$ and $N_{x_2} = 19$. The grid parameters are chosen such that $h_{x_1} \approx h_{x_2}$ as $N_{x_1}/N_{x_2} \approx (x_1^M - x_1^m)/(x_2^M - x_2^m) = 2\pi$. Although the figure on the right clearly shows J_{I} and J_{B} have converged, and that Δm reached computer precision, the algorithm does not

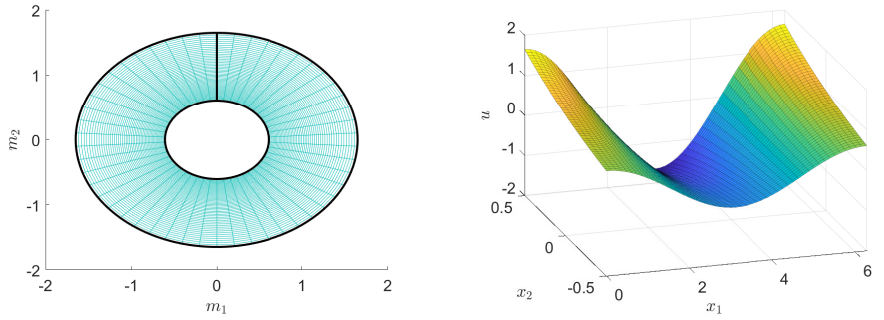


Figure 5.19: The target and the exact mapping on the left, and the solution surface on the right, both shown on a 51×51 grid for Example 4 (Annulus).

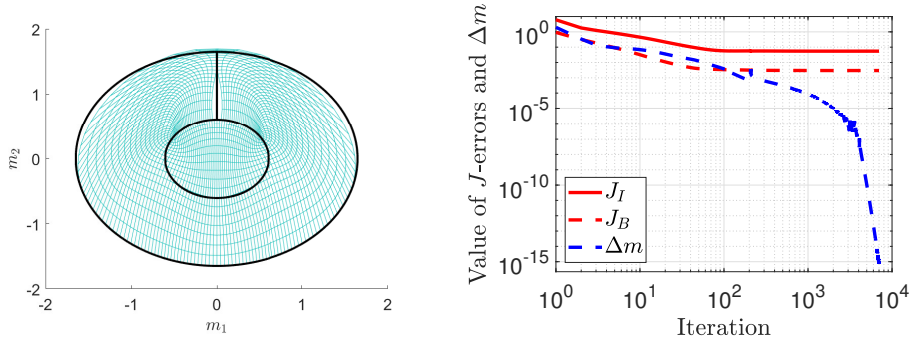


Figure 5.20: The numerical mapping $\mathbf{m}$ after convergence (left) and the history of J_I, J_B and Δm for SPM.

yield a correct solution. This is evident by the fact that the solution does not satisfy the transport boundary condition because there are points $\mathbf{m}_{ij}$ which lie outside $\mathcal{Y}$, nor does the solution solve the hyperbolic Monge-Ampère equation as is shown by the residual ϵ_r in Figure 5.21 on the left. The reason why SPM does not produce accurate solutions is easiest demonstrated by visualizing a few iterations. To this end, consider the boundary routine for the first, third and tenth iteration as shown in Figure 5.22. We focus on one segment of the mapping of the boundary, i.e, $\mathbf{m}_{ij}$ with $\mathbf{x}_{ij} \in \Gamma_4^{\mathcal{X}} = [0, 2\pi] \times \{-\frac{1}{2}\}$ for the initial guess shown in Figure 5.22. For the exact solution, $\Gamma_4^{\mathcal{X}}$ needs to be mapped to the entire inner circle of the target, i.e., $\Gamma_4^{\mathcal{Y}}$. As shown for the first iteration, $\Gamma_4^{\mathcal{X}}$ is mapped to only part of $\Gamma_4^{\mathcal{Y}}$, viz., the accompanying $\mathbf{b}_l$'s lie on the northern part of $\Gamma_4^{\mathcal{Y}}$ (the inner circle). In subsequent iterations, shown in the middle and on the right in Figure 5.22, $\Gamma_4^{\mathcal{X}}$ will again not be mapped to the whole of $\Gamma_4^{\mathcal{Y}}$; it is only mapped to the top part of the inner circle. This process continues indefinitely.

For SALM such an accumulation of $\mathbf{b}_l$'s does not occur, as by construction, the $\mathbf{b}_l$'s are distributed over the boundary segments. The results of the first, second and third iteration of the $\mathbf{b}$-minimization are shown in Figure 5.23. Clearly, SALM does not suffer from the same flaws as SPM. As such, the convergence is expected to behave as for the other examples, which is confirmed by the results shown in Figure 5.21 on the right.

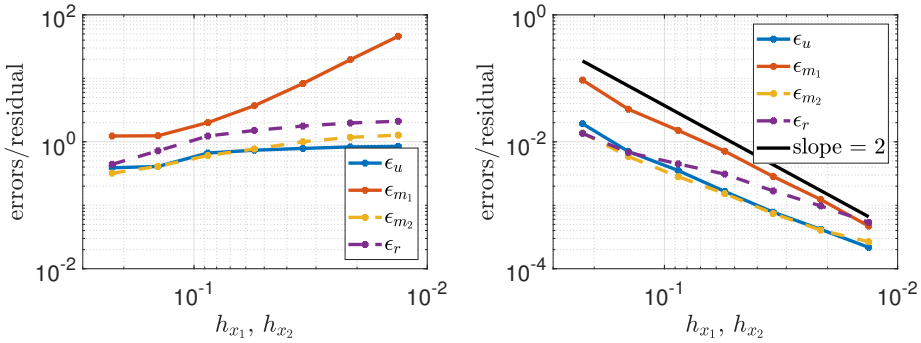


Figure 5.21: Residual for SPM (left) and SALM (right). No convergence for SPM, and second-order convergence for SALM is observed.

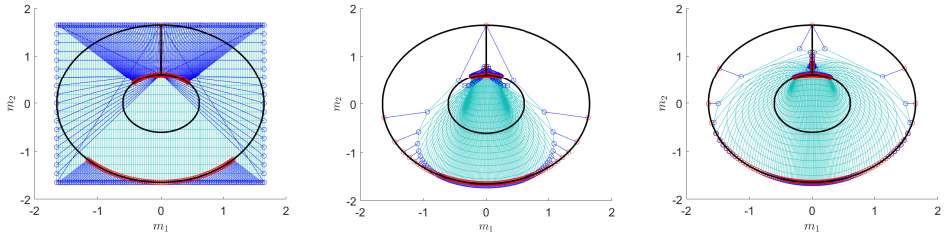


Figure 5.22: From left to right, the first, third and tenth iteration for SPM.

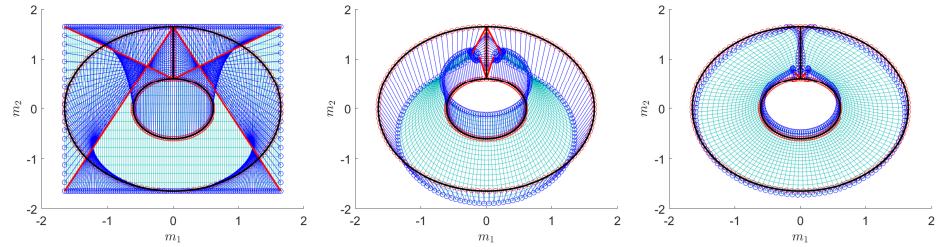


Figure 5.23: From left to right, the first, second and third iteration for SALM.

5.2.5 Example 5: Gradient dependent problem

Lastly we consider an example with f dependent on the gradient of the solution, i.e., $f = f(x_1, x_2, \nabla u)$, viz.

$$f^2(\mathbf{x}, \mathbf{m}) = 3x_2^2 - m_1 \sin(x_1) - \frac{1}{4} m_2^2. \tag{5.97}$$

We consider the domain $\mathcal{X} = [-1, 1] \times [1, 3/2]$ and $\partial\mathcal{Y} = \cup_{k=1}^{4} \Gamma_k^{\mathcal{Y}}$ with

$$\mathbf{y}_1(s) = (\tfrac{\sin(1)}{4} s^2 + \sin(1)s + \sin(1),\ \cos(1)s + 2\cos(1)), \tag{5.98a}$$
$$\mathbf{y}_2(s) = (-\tfrac{9}{4} \sin(2s - 1),\ 3\cos(2s - 1)), \tag{5.98b}$$
$$\mathbf{y}_3(s) = (-\tfrac{\sin(1)}{4} s^2 + \tfrac{3}{2}s - \tfrac{9\sin(1)}{4},\ -\cos(1)s + 3\cos(1)), \tag{5.98c}$$
$$\mathbf{y}_4(s) = (\sin(2s - 1),\ 2\cos(2s - 1)). \tag{5.98d}$$

The exact solution is given by

$$u(x_1, x_2) = x_2^2 \cos(x_1), \tag{5.99}$$

and is, together with the mapping and target domain, shown in Figure 5.24.

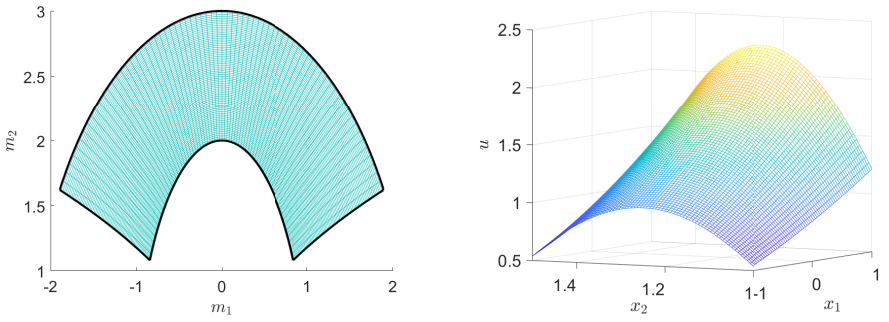


Figure 5.24: The target and the exact mapping on the left, and the solution surface on the right, both shown on a 51×51 grid for Example 5 (Gradient dependent problem).

By construction of the algorithm, little effort is required for f to be dependent on the mapping $\mathbf{m}$. The difference being that during the n^{th} iteration, $f^2(\mathbf{x}_{ij}, \mathbf{m}_{ij}^n)$ has to be evaluated instead of $f^2(\mathbf{x}_{ij})$ in the optimization of $\mathbf{P}$. The results for SPM and SALM with $N_{x_1} = N_{x_2}$ are given in Figure 5.25, showing second-order convergence for both methods. In this case grid-shock correction is needed for SPM to ensure proper convergence.

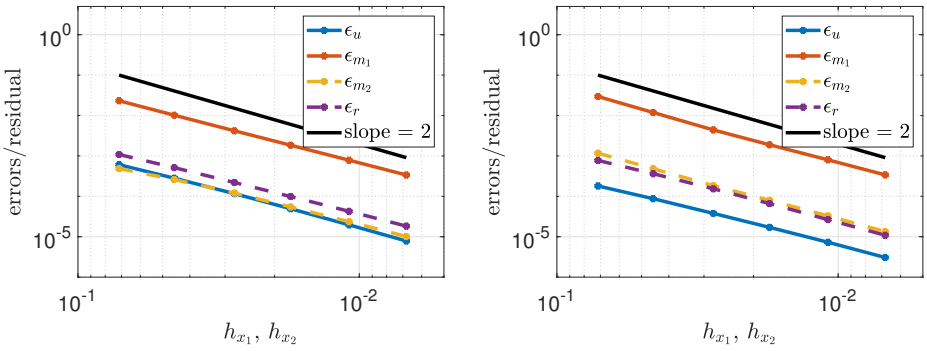


Figure 5.25: Convergence of SPM (left) and SALM (right).

5.3 Summary

We have introduced a least-squares solver for the hyperbolic Monge-Ampère equation with transport boundary conditions. The algorithm, originally introduced by Prins et al. [66] for the elliptic Monge-Ampère equation, has been improved to encompass a more complete description of the roots for the P-optimization. Furthermore, we introduced two new boundary methods, the *segmented projection mathod* and the *segmented arc length method*. All three boundary methods, if convergent, show second-order convergence of the residual and the global discretization errors as function of the mesh size, and also second-order convergence as function of the number of boundary points. Of the three boundary methods, the *segmented arc length method* is both the only method to converge for all tried examples and is computationally most efficient, both in terms of computation time per iteration, as in total number of iterations required.

Chapter 6

Freeform Illumination Optics

The aim of this chapter is to derive a framework for designing freeform optical systems consisting of saddle-shaped optical surfaces, i.e., surfaces without any symmetries with both positive and negative curvature. The branch of optics which concerns itself with the design of optical systems is called illumination optics.

We start this chapter with an introduction to optics. First, we introduce Maxwell equations and the underlying reason to consider those. Via these equations, we derive the eikonal equation which lays the foundation for geometrical optics, the branch of optics which considers light as rays. Next we introduce Fermat's principle and the Hamiltonian formalism.

With the use of Hamilton's characteristics, we subsequently formulate the design process of four optical systems as solutions to the Monge-Ampère equation. This is achieved by combining conservation of energy with a cost balance, which is to be derived for each optical system individually. Each optical system consists of a parallel light source and either a far-field or parallel target. Furthermore, the optical systems constructed consist of the minimum required number of freeform lenses or reflectors. The optical systems presented here are a subset of those known in the literature [3].

6.1 A primer on optics

The mathematical theory of optics is vast and deep. Because the content of this thesis is rather mathematical, we do not assume any prior knowledge on optics from the reader and instead choose the fundamentals as starting point. Let us consider the outline of this chapter. Our aim is to design optical systems for which a light source and target distribution are given. To do so, we first

need to introduce light and its propagation. This is described by the Maxwell equations. When the wavelength of light is small compared to the optical elements in a system, i.e., compared to the size of reflectors and lenses, light can be considered to move along light rays, which follows from the eikonal equation. The path a light ray takes is described by the ray equation and, as it turns out, is a stationary point of the optical path length. This is known as the principle of Fermat. It follows that each ray is a solution to one of four Hamiltonian systems.

We structure this section similar to those found in [82] and [68]. For a more rigorous treatment of the subject we suggest the book by Born and Wolf [11] or, for a more applied physics oriented approach the book by Hecht [39].

6.1.1 Maxwell's equations

In 1820, the first connection between electric and magnetic effects was discovered by Hans Christian Ørsted when he found that electric currents produce magnetic forces [25], implying electric and magnetic fields should be considered jointly which gave rise to the term electromagnetism. In 1845, Michael Faraday observed that light is influenced by magnetic fields [45], meaning light had to be electromagnetic.

Let us consider an electric field $\mathbf{E} = \mathbf{E}(\mathbf{x}, t)$ and magnetic field $\mathbf{H} = \mathbf{H}(\mathbf{x}, t)$, both in $\mathbb{R}^3$ with spatial coordinates $\mathbf{x}$ and time t. The equations describing the relation between the electric and magnetic fields are called the Maxwell equations. In the presence of a dielectric, i.e., nonconducting, medium, the Maxwell equations in differential form read

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon}, \tag{6.1a}$$

$$\nabla \cdot \mathbf{H} = 0, \tag{6.1b}$$

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t}, \tag{6.1c}$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \epsilon \frac{\partial \mathbf{E}}{\partial t}, \tag{6.1d}$$

where ρ is the free electric charge density, $\mathbf{J}$ the electric current density, and ϵ and μ the permittivity and permeability of the dielectric, respectively. For the purpose of illumination optics, we have that $\rho = 0$, $\mathbf{J} = \mathbf{0}$ and both ϵ and μ are

constants, simplifying the Maxwell equations to

$$\nabla \cdot \mathbf{E} = 0, \tag{6.2a}$$
$$\nabla \cdot \mathbf{H} = 0, \tag{6.2b}$$
$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t}, \tag{6.2c}$$
$$\nabla \times \mathbf{H} = \epsilon \frac{\partial \mathbf{E}}{\partial t}. \tag{6.2d}$$

The electric and magnetic fields $\mathbf{E}$ and $\mathbf{H}$ propagate as waves. To see this, we take the curl of (6.2c) and (6.2d) to obtain

$$\nabla \times (\nabla \times \mathbf{E}) = -\mu \nabla \times \frac{\partial \mathbf{H}}{\partial t} = -\mu \frac{\partial}{\partial t} (\nabla \times \mathbf{H}) = -\mu\epsilon \frac{\partial^2 \mathbf{E}}{\partial t^2}, \tag{6.3a}$$
$$\nabla \times (\nabla \times \mathbf{H}) = \quad \epsilon \nabla \times \frac{\partial \mathbf{E}}{\partial t} \quad = \quad \epsilon \frac{\partial}{\partial t} (\nabla \times \mathbf{E}) \quad = -\mu\epsilon \frac{\partial^2 \mathbf{H}}{\partial t^2}. \tag{6.3b}$$

Applying the vector identity $\nabla \times (\nabla \times \mathbf{v}) = \nabla(\nabla \cdot \mathbf{v}) - \nabla^2 \mathbf{v}$ to (6.3) and subsequently using (6.2a) and (6.2b) yields the wave equations for $\mathbf{E}$ and $\mathbf{H}$, viz.

$$\frac{\partial^2 \mathbf{E}}{\partial t^2} = \frac{1}{\mu\epsilon} \nabla^2 \mathbf{E}, \tag{6.4a}$$
$$\frac{\partial^2 \mathbf{H}}{\partial t^2} = \frac{1}{\mu\epsilon} \nabla^2 \mathbf{H}. \tag{6.4b}$$

The factor $1/(\mu\epsilon)$ is the square of the speed of propagation of the electric and magnetic waves, i.e., $\mathbf{E}$ and $\mathbf{H}$ propagate at a speed of $v = 1/\sqrt{\mu\epsilon}$. Because light consists of both an $\mathbf{E}$ and $\mathbf{H}$ field, we call v the speed of light. In a vacuum the permittivity and permeability are given by $\epsilon_0 \approx 8.85 \cdot 10^{-12}$ $\mathrm{A}^2 \cdot \mathrm{s}^4 \cdot \mathrm{kg}^{-1} \cdot \mathrm{m}^{-3}$ and $\mu_0 \approx 1.26 \cdot 10^{-6}$ $\mathrm{kg} \cdot \mathrm{m} \cdot \mathrm{s}^{-2} \cdot \mathrm{A}^{-2}$ such that the speed of light in vacuum is $c = 1/\sqrt{\mu_0\epsilon_0} \approx 3.00 \cdot 10^8$ $\mathrm{m} \cdot \mathrm{s}^{-1}$. Note that c is constant and v depends on the medium, and may therefore depend on $\mathbf{x}$. The ratio between the speed of light in vacuum and in a medium is called the refractive index and is given by

$$n(\mathbf{x}) = \frac{c}{v(\mathbf{x})}, \tag{6.5}$$

with $n \in \mathbb{R}$ and $n > 1$. In this thesis we only consider media with a constant refractive index, i.e., $n(\mathbf{x})$ is a piecewise constant function.

The energy density of an electromagnetic wave is defined by

$$\mathcal{U} = \frac{1}{2}(\epsilon |\mathbf{E}|^2 + \mu |\mathbf{H}|^2). \tag{6.6}$$

This energy is carried by the electromagnetic wave in the direction

$$\mathbf{S} = \mathbf{E} \times \mathbf{H}, \tag{6.7}$$

called the Poynting vector, named after its discoverer John Poynting. To see why the energy flows in the direction $\mathbf{S}$, we derive the conservation law for $\mathcal{U}$. To this end, we take the inner product of (6.2c) with $\mathbf{H}$ and subtract this from the inner product of (6.2d) with $\mathbf{E}$ and find

$$(\nabla \times \mathbf{H}) \cdot \mathbf{E} - (\nabla \times \mathbf{E}) \cdot \mathbf{H} - \left(\epsilon \mathbf{E} \cdot \frac{\partial \mathbf{E}}{\partial t} + \mu \mathbf{H} \cdot \frac{\partial \mathbf{H}}{\partial t} \right) = 0. \tag{6.8}$$

Subsequently using the vector identity $\nabla \cdot (\mathbf{u} \times \mathbf{v}) = (\nabla \times \mathbf{u}) \cdot \mathbf{v} - \mathbf{u} \cdot (\nabla \times \mathbf{v})$ yields

$$\nabla \cdot (\mathbf{E} \times \mathbf{H}) + \frac{1}{2} \frac{\partial}{\partial t} (\epsilon |\mathbf{E}|^2 + \mu |\mathbf{H}|^2) = 0. \tag{6.9}$$

Thus, energy conservation is given by

$$\frac{\partial \mathcal{U}}{\partial t} + \nabla \cdot \mathbf{S} = 0, \tag{6.10}$$

and indeed the electromagnetic energy $\mathcal{U}$ is transported in the direction of the Poynting vector $\mathbf{S}$.

6.1.2 Geometrical optics

For common optical devices the dimension of optical elements, lenses and reflectors, is typically in the order of millimeters to centimeters, while the wavelength of light is in the order of nanometers. This is the starting point for geometrical optics, also named the short wavelength approximation. Let us consider time harmonic fields as solutions to the Maxwell equations (6.2), viz.

$$\mathbf{E}(\mathbf{x}, t) = \mathbf{e}(\mathbf{x}) e^{i\kappa(\varphi(\mathbf{x}) - ct)}, \tag{6.11a}$$

$$\mathbf{H}(\mathbf{x}, t) = \mathbf{h}(\mathbf{x}) e^{i\kappa(\varphi(\mathbf{x}) - ct)}, \tag{6.11b}$$

where $\kappa = \omega / c$ the free-space wave number with ω is the angular frequency, φ the function incorporating the spatial dependence of the phase and $\mathbf{e}$ and $\mathbf{h}$ the amplitudes of their respective fields. The wavelength and the free-space wave number are related by $\lambda \kappa = 2\pi$. Substitution of the harmonic fields (6.11) in

the Maxwell equations (6.2) yields

$$\frac{1}{i\kappa}\nabla\cdot\mathbf{e}+\nabla\varphi\cdot\mathbf{e}=0, \tag{6.12a}$$

$$\frac{1}{i\kappa}\nabla\cdot\mathbf{h}+\nabla\varphi\cdot\mathbf{h}=0, \tag{6.12b}$$

$$\frac{1}{i\kappa}\nabla\times\mathbf{e}+\nabla\varphi\times\mathbf{e}=\mu c\mathbf{h}, \tag{6.12c}$$

$$\frac{1}{i\kappa}\nabla\times\mathbf{h}+\nabla\varphi\times\mathbf{h}=-\epsilon c\mathbf{e}. \tag{6.12d}$$

The short wavelength approximation, $\lambda\to 0$, implies $\kappa\to\infty$ and consequently (6.12) reduces to

$$\nabla\varphi\cdot\mathbf{e}=0, \tag{6.13a}$$

$$\nabla\varphi\cdot\mathbf{h}=0, \tag{6.13b}$$

$$\nabla\varphi\times\mathbf{e}=c\mu\mathbf{h}, \tag{6.13c}$$

$$\nabla\varphi\times\mathbf{h}=-c\epsilon\mathbf{e}. \tag{6.13d}$$

Equations (6.13a) and (6.13b) imply that $\nabla\varphi$ is perpendicular to both $\mathbf{e}$ and $\mathbf{h}$, i.e., $\mathbf{E}$ and $\mathbf{H}$ are transverse waves. Equations (6.13c) and (6.13d) both imply that $\mathbf{e}$ and $\mathbf{h}$ are perpendicular to each other.

For (6.13) to admit a nontrivial solution for $\mathbf{e}$ and $\mathbf{h}$, a consistency requirement on φ needs to be satisfied [11, §. 3.1]. This consistency requirement can be obtained by eliminating $\mathbf{e}$ and $\mathbf{h}$ from (6.13). Note, (6.13a) and (6.13b) are implied by separately taking the inner product of $\nabla\varphi$ with (6.13c) and (6.13d). Therefore, we should utilize (6.13c) and (6.13d), viz., we solve $\mathbf{h}$ from (6.13c) and subsequently substitute it into (6.13d) to obtain

$$\nabla\varphi\times(\nabla\varphi\times\mathbf{e})+\epsilon\mu c^2\mathbf{e}=\mathbf{0}. \tag{6.14}$$

Applying the vector identity $\mathbf{u}\times(\mathbf{v}\times\mathbf{w})=(\mathbf{u}\cdot\mathbf{w})\mathbf{v}-(\mathbf{u}\cdot\mathbf{v})\mathbf{w}$ yields

$$(\nabla\varphi\cdot\mathbf{e})\nabla\varphi-(\nabla\varphi\cdot\nabla\varphi)\mathbf{e}+\epsilon\mu c^2\mathbf{e}=\mathbf{0}. \tag{6.15}$$

By (6.13a), we obtain

$$(|\nabla\varphi|^2-\epsilon\mu c^2)\mathbf{e}=\mathbf{0}. \tag{6.16}$$

Analogously, by solving $\mathbf{e}$ from (6.13d) and substituting it into (6.13c) we obtain

$$(|\nabla\varphi|^2-\epsilon\mu c^2)\mathbf{h}=\mathbf{0}. \tag{6.17}$$

For non-trivial solutions $\mathbf{e}$ and $\mathbf{h}$ cannot simultaneously be $\mathbf{0}$, hence

$$|\nabla\varphi| = \sqrt{\epsilon\mu}c = n, \tag{6.18}$$

where we used (6.5) and $v = 1/\sqrt{\mu\epsilon}$. The nonlinear PDE (6.18) is known as the eikonal equation and the function φ is called the eikonal. The surfaces $\varphi(\mathbf{x}) = \text{const}$ are called geometric wave surfaces or geometric wave fronts [11, p.119]. The eikonal equation is fundamental to geometrical optics in the sense that most properties of geometrical optics are either related to, or derived from (6.18).

To conclude this section, we show that the Poynting vector is parallel to $\nabla\varphi$ and hence perpendicular to a wave fronts. By the definition of the Poynting vector and the harmonic fields (6.11) it follows that

$$\begin{aligned}\mathbf{S}(\mathbf{x},t) &= \mathbf{E}(\mathbf{x},t) \times \mathbf{H}(\mathbf{x},t) \\ &= (\mathbf{e}(\mathbf{x},t) \times \mathbf{h}(\mathbf{x},t))e^{2i\kappa(\varphi(\mathbf{x})-ct)}.\end{aligned} \tag{6.19}$$

We define $\mathbf{s} = \mathbf{e} \times \mathbf{h}$ such that $\mathbf{S}(\mathbf{x},t) = \mathbf{s}(\mathbf{x},t)e^{2i\kappa(\varphi(\mathbf{x})-ct)}$. By substituting $\mathbf{e}$ and $\mathbf{h}$, given by equations (6.13c) and (6.13c), into $\mathbf{s}$ and then applying the vector identities $\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) = (\mathbf{u} \cdot \mathbf{w})\mathbf{v} - (\mathbf{u} \cdot \mathbf{v})\mathbf{w}$, then $\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) = \mathbf{v} \cdot (\mathbf{w} \times \mathbf{u})$ and $\mathbf{u} \cdot (\mathbf{u} \times \mathbf{v}) = \mathbf{0}$ we find

$$\begin{aligned}\mathbf{s} &= \mathbf{e} \times \mathbf{h} \\ &= -\frac{1}{c^2\mu\epsilon}(\nabla\varphi \times \mathbf{h}) \times (\nabla\varphi \times \mathbf{e}) \\ &= -\frac{1}{n^2}[((\nabla\varphi \times \mathbf{h}) \cdot \mathbf{e}))\nabla\varphi - (\nabla\varphi \cdot (\nabla\varphi \times \mathbf{h})\mathbf{e})] \\ &= -\frac{1}{n^2}(\nabla\varphi \cdot (\mathbf{h} \times \mathbf{e}))\nabla\varphi \\ &= \frac{1}{n^2}(\nabla\varphi \cdot \mathbf{s})\nabla\varphi.\end{aligned} \tag{6.20}$$

It follows that $\mathbf{s}$ is parallel to $\nabla\varphi$ and henceforth $\mathbf{S}$ is parallel to $\nabla\varphi$. Because the electromagnetic energy $\mathcal{U}$ is carried in the direction of the Poynting vector $\mathbf{S}$, the energy is propagated in the direction $\nabla\varphi$, i.e., in the direction perpendicular to the wave front, justifying the notion of light rays.

6.1.3 The ray equation and Fermat's principle

The eikonal equation $|\nabla\varphi| = n$ is one example of a hyperbolic first-order nonlinear PDE in three variables. We aim to apply the method of characteristics to

this equation. To this end, consider a general first-order nonlinear PDE given by

$$F(\mathbf{x}, u, \boldsymbol{p}^*) = 0, \tag{6.21}$$

for the unknown function $u = u(\mathbf{x})$ and $\boldsymbol{p}^* = \nabla u$. We parametrize the characteristics of (6.21) by s and write $\mathbf{x} = \mathbf{x}(s)$, $u = u(\mathbf{x}(s))$ and $\boldsymbol{p}^* = \boldsymbol{p}^*(\mathbf{x}(s))$. The evolution of the solution u along a characteristic is then given by the ODE system [24]

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}s} = \frac{\partial F}{\partial \boldsymbol{p}^*}, \tag{6.22a}$$

$$\frac{\mathrm{d}u}{\mathrm{d}s} = \boldsymbol{p}^* \cdot \frac{\partial F}{\partial \boldsymbol{p}^*}, \tag{6.22b}$$

$$\frac{\mathrm{d}\boldsymbol{p}^*}{\mathrm{d}s} = -\boldsymbol{p}^* \frac{\partial F}{\partial u} - \frac{\partial F}{\partial \mathbf{x}}. \tag{6.22c}$$

Applying the method of characteristics to the eikonal equation $F(\mathbf{x}, \varphi, \nabla\varphi) = |\nabla\varphi| - n = 0$ with $n = n(\mathbf{x})$, we obtain the ODE system

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}s} = \frac{1}{n}\nabla\varphi, \tag{6.23a}$$

$$\frac{\mathrm{d}\varphi}{\mathrm{d}s} = n, \tag{6.23b}$$

$$\frac{\mathrm{d}(\nabla\varphi)}{\mathrm{d}s} = \nabla n. \tag{6.23c}$$

Equation (6.23a) shows that $\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}s}$ is parallel to $\nabla\varphi$, i.e., light rays are characteristics curves of the eikonal equation. By the eikonal equation and (6.23a), we find $\left|\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}s}\right| = 1$ and hence, s is the arc length of a light ray. Multiplying (6.23a) by n and taking the derivative w.r.t. the arc length shows

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(n\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}s}\right) = \frac{\mathrm{d}(\nabla\varphi)}{\mathrm{d}s} = \nabla n, \tag{6.24}$$

where we used (6.23c) in the last equality. Equation (6.24) is known as the ray equation for the unknown $\mathbf{x}$. It shows that the path a light ray takes only depends on the refractive index $n = n(\mathbf{x})$ and, among others, implies that light rays follow straight paths when n is constant. When n is piece-wise constant, e.g., in the case of an isotropic lens surrounded by air, the path consists of straight line segments and only changes direction at the interface. Solving the ray equation will, in general, be a tedious exercise due to, for example, n

being discontinuous at an optical interface. We instead consider the optical path length

$$V[\mathbf{x}] := \int_{\mathcal{C}} n(\mathbf{x}(s))\,\mathrm{d}s, \tag{6.25}$$

over a continuous curve $\mathcal{C}$ between the points P_1 and P_2. The ray equation (6.24) is then recovered as the Euler-Lagrange equation [24] of the optical path length V if $\mathbf{x}$ is stationary w.r.t. V, which is asserted by Fermat's principle. We do not give a derivation here, but instead refer the reader to [11, § 3.3.2]. Fermat's principle, also know as the principle of shortest optical path, states that the optical path length

$$\int_{P_1}^{P_2} n\,\mathrm{d}s, \tag{6.26}$$

of a light ray between two fixed points P_1 and P_2 is stationary with respect to its path.

6.1.4 Hamiltonian optics

In geometrical optics, Fermat's principle is the notion of light rays being stationary with respect to their path. This principle is mathematically similar to Lagrangian mechanics, a branch of classical mechanics, which is founded on the stationary-action principle (also known as the principle of least action). Sir William Rowan Hamilton introduced Hamiltonian mechanics as a reformulation of Lagrangian mechanics and subsequently applied his ideas to geometrical optics, founding the branch of Hamiltonian optics in his series of papers [36–38]. Here we discuss some of the ideas of Hamiltonian optics.

6.1.4.1 From Fermat to Hamilton

We consider the path of one light ray between two reference planes in $\mathbb{R}^3$. Without loss of generality, we choose the planes perpendicular to the z-axis, i.e., parallel to the xy-plane. We parametrize the light ray as $\mathbf{x} = \mathbf{x}(z) = (q_1(z), q_2(z), z)$ where $\mathbf{q} = (q_1, q_2)$ are generalized coordinates. The optical path length (6.25) from the plane $z = z_\mathrm{s}$ to the plane $z = z_\mathrm{t} > z_\mathrm{s}$ is, by a coordinate transformation, found to be

$$V[\mathbf{q}] = \int_{z_\mathrm{s}}^{z_\mathrm{t}} n(\mathbf{q}, z)\sqrt{1 + |\mathbf{q}'|^2}\,\mathrm{d}z, \tag{6.27}$$

where a prime denotes the derivative with respect to z. The Euler-Lagrange equations of (6.27) are given by

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{n\mathbf{q}'}{\sqrt{1+|\mathbf{q}'|^2}}\right) - \sqrt{1+|\mathbf{q}'|^2}\frac{\partial n}{\partial \mathbf{q}} = \mathbf{0}. \tag{6.28}$$

Note that according to Fermat's principle, a light ray with path $(q_1(z), q_2(z), z)$ satisfies equation (6.28) for $z_\mathrm{s} \leq z \leq z_\mathrm{t}$. We introduce the momentum vector $\mathbf{p} = (p_1, p_2)^\mathrm{T}$, given by

$$\mathbf{p} = \frac{n\mathbf{q}'}{\sqrt{1+|\mathbf{q}'|^2}}. \tag{6.29}$$

Note that $|\mathbf{p}| \leq n$. Elaborating on $n^2 - |\mathbf{p}|^2$ using (6.29) straightforwardly yields

$$\sqrt{1+|\mathbf{q}'|^2} = \frac{n}{\sqrt{n^2-|\mathbf{p}|^2}}. \tag{6.30}$$

Combining equations (6.29) and (6.30) gives

$$\mathbf{q}' = \frac{\mathbf{p}}{\sqrt{n^2-|\mathbf{p}|^2}}. \tag{6.31a}$$

Substitution of (6.29) and (6.30) into (6.28) yields

$$\mathbf{p}' = \frac{n}{\sqrt{n^2-|\mathbf{p}|^2}}\frac{\partial n}{\partial \mathbf{q}}. \tag{6.31b}$$

Equations (6.31) form a coupled ODE system of four equations. Using the Hamiltonian

$$H = H(z, \mathbf{q}, \mathbf{p}) = -\sqrt{n^2-|\mathbf{p}|^2}, \tag{6.32}$$

the ODE system can be written as a so-called Hamiltonian system, viz.

$$\mathbf{q}' = \frac{\partial H}{\partial \mathbf{p}} = \frac{\mathbf{p}}{\sqrt{n^2-|\mathbf{p}|^2}}, \qquad \mathbf{p}' = -\frac{\partial H}{\partial \mathbf{q}} = \frac{n}{\sqrt{n^2-|\mathbf{p}|^2}}\frac{\partial n}{\partial \mathbf{q}}. \tag{6.33}$$

Note that by (6.30)

$$n\sqrt{1+|\mathbf{q}'|^2} = \frac{n^2}{\sqrt{n^2-|\mathbf{p}|^2}} = \mathbf{q}'\cdot\mathbf{p} - H(z, \mathbf{q}, \mathbf{p}) =: L(z, \mathbf{q}, \mathbf{q}'), \tag{6.34}$$

where we introduce the Lagrangian $L(z, \mathbf{q}, \mathbf{q}')$ conform (6.27). The optical path length can thus be written as

$$V[\mathbf{q}] = \int_{z_s}^{z_t} L(z, \mathbf{q}, \mathbf{q}')\, \mathrm{d}z. \tag{6.35}$$

The Hamiltonian and Lagrangian approaches are two different formalisms for the same phenomena and one can freely transform between the two. We do not delve into details here but instead refer to [56] for an excellent overview on Hamiltonian optics and to [52] for the dualism of the Lagrangian and the Hamiltonian approaches and their applications to classical mechanics.

6.1.4.2 Hamilton's characteristic functions

The Hamiltonian system (6.33) is an ODE system, naturally requiring boundary conditions. The main result of Hamilton's theory is that chosen either the coordinate or momentum at the plane $z = z_s$ and either the coordinate or momentum at the plane $z = z_t$ as boundary conditions for (6.33), the light ray from the plane $z = z_s$ to $z = z_t$ can be determined completely if a solution exists.

As an example, consider the Hamiltonian boundary value problem

$$\begin{aligned} \mathbf{q}' &= \frac{\partial H}{\partial \mathbf{p}}, \qquad \mathbf{p}' = -\frac{\partial H}{\partial \mathbf{q}}, \qquad z_s < z < z_t, \\ \mathbf{q}(z_s) &= \mathbf{q}_s, \qquad \mathbf{q}(z_t) = \mathbf{q}_t. \end{aligned} \tag{6.36}$$

The system represents a light ray passing through the coordinates $\mathbf{q}_s$ and $\mathbf{q}_t$ at $z = z_s$ and $z = z_t$, respectively. We write the general solution to (6.36) as $\mathbf{q} = \mathbf{q}(z; z_s, z_t, \mathbf{q}_s, \mathbf{q}_t)$ and $\mathbf{p} = \mathbf{p}(z; z_s, z_t, \mathbf{q}_s, \mathbf{q}_t)$ and find that the momenta $\mathbf{p}_s$ and $\mathbf{p}_t$ can subsequently be calculated. We show this in Section 6.1.4.3. The optical path length of this specific solution is the so-called point characteristic. In total four such characteristics exist, given by (6.37), one for each combination of $\{\mathbf{q}_s, \mathbf{q}_t, \mathbf{p}_s, \mathbf{p}_t\}$ with exactly one source and one target vector. These characteristics describe the optical path length of a ray through the optical system. Furthermore, each characteristic function satisfies an eikonal equation.

Point characteristic:

$$V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t) = \int_{z_s}^{z_t} n(\mathbf{q}, z)\sqrt{1 + |\mathbf{q}'|^2}\,\mathrm{d}z, \tag{6.37a}$$

$$\mathbf{p}_s = -\frac{\partial V}{\partial \mathbf{q}_s}, \qquad \mathbf{p}_t = \frac{\partial V}{\partial \mathbf{q}_t}. \tag{6.37b}$$

Mixed characteristic of the first kind:

$$W(z_s, z_t, \mathbf{q}_s, \mathbf{p}_t) = V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t) - \mathbf{q}_t \cdot \mathbf{p}_t, \tag{6.37c}$$

$$\mathbf{p}_s = -\frac{\partial W}{\partial \mathbf{q}_s}, \qquad \mathbf{q}_t = -\frac{\partial W}{\partial \mathbf{p}_t}. \tag{6.37d}$$

Mixed characteristic of the second kind:

$$W^*(z_s, z_t, \mathbf{p}_s, \mathbf{q}_t) = V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t) + \mathbf{q}_s \cdot \mathbf{p}_s, \tag{6.37e}$$

$$\mathbf{q}_s = \frac{\partial W^*}{\partial \mathbf{p}_s}, \qquad \mathbf{p}_t = \frac{\partial W^*}{\partial \mathbf{q}_t}. \tag{6.37f}$$

Angular characteristic:

$$T(z_s, z_t, \mathbf{p}_s, \mathbf{p}_t) = V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t) + \mathbf{q}_s \cdot \mathbf{p}_s - \mathbf{q}_t \cdot \mathbf{p}_t, \tag{6.37g}$$

$$\mathbf{q}_s = \frac{\partial T}{\partial \mathbf{p}_s}, \qquad \mathbf{q}_t = -\frac{\partial T}{\partial \mathbf{p}_t}. \tag{6.37h}$$

In the next section we discuss the point characteristic and present the corresponding Hamiltonian system. The derivations of the remaining characteristic functions follows a similar structure and we refer the interested reader to [68, § 2.7].

6.1.4.3 The point characteristic

We consider the optical path length (6.35), given by

$$V[\mathbf{q}] = \int_{z_s}^{z_t} L(z, \mathbf{q}, \mathbf{q}')\,\mathrm{d}z, \tag{6.38}$$

with the Lagrangian $L(z, \mathbf{q}, \mathbf{q}') = \mathbf{q}' \cdot \mathbf{p} - H(z, \mathbf{q}, \mathbf{p})$. Let $\mathbf{q} = \mathbf{q}(z; z_s, z_t, \mathbf{q}_s, \mathbf{q}_t)$ and $\mathbf{p} = \mathbf{p}(z; z_s, z_t, \mathbf{q}_s, \mathbf{q}_t)$ denote the solution to the Hamiltonian boundary

value problem (6.36) for fixed $\mathbf{q}_\mathrm{s}$ and $\mathbf{q}_\mathrm{t}$. Given this solution, the optical path length can be expressed as $V = V(z_\mathrm{s}, z_\mathrm{s}, \mathbf{q}_\mathrm{s}, \mathbf{q}_\mathrm{t})$, which is known as the point characteristic.

Next, we derive explicit expressions for $\mathbf{p}_\mathrm{s}$ and $\mathbf{p}_\mathrm{t}$. To this end, let $c \in \{q_{\mathrm{s},1}, q_{\mathrm{s},2}, q_{\mathrm{t},1}, q_{\mathrm{t},2}\}$, where $q_{\mathrm{s},1}$ denotes the first component of $\mathbf{q}_\mathrm{s}$ and similarly for the rest. The derivative of $L(z, \mathbf{q}, \mathbf{q}')$ with respect to c then reads

$$\frac{\partial L}{\partial c} = \frac{\partial \mathbf{q}'}{\partial c} \cdot \mathbf{p} + \mathbf{q}' \cdot \frac{\partial \mathbf{p}}{\partial c} - \frac{\partial H}{\partial \mathbf{q}} \cdot \frac{\partial \mathbf{q}}{\partial c} - \frac{\partial H}{\partial \mathbf{p}} \cdot \frac{\partial \mathbf{p}}{\partial c}. \tag{6.39}$$

By the Hamiltonian system (6.33) the derivative reads

$$\frac{\partial L}{\partial c} = \frac{\partial \mathbf{q}'}{\partial c} \cdot \mathbf{p} + \mathbf{p}' \cdot \frac{\partial \mathbf{q}}{\partial c} = \left(\frac{\partial \mathbf{q}}{\partial c} \cdot \mathbf{p} \right)'. \tag{6.40}$$

By the Leibniz integral rule and the (second) fundamental theorem of calculus the derivative of V with respect to c reads

$$\frac{\partial V}{\partial c} = \frac{\partial}{\partial c} \int_{z_\mathrm{s}}^{z_\mathrm{t}} L \, \mathrm{d}z = \int_{z_\mathrm{s}}^{z_\mathrm{t}} \frac{\partial L}{\partial c} \, \mathrm{d}z = \left[\frac{\partial \mathbf{q}}{\partial c} \cdot \mathbf{p} \right]_{z_\mathrm{s}}^{z_\mathrm{t}} = \frac{\partial \mathbf{q}_\mathrm{t}}{\partial c} \cdot \mathbf{p}_\mathrm{t} - \frac{\partial \mathbf{q}_\mathrm{s}}{\partial c} \cdot \mathbf{p}_\mathrm{s}. \tag{6.41}$$

By the chain rule for differentiation we find the derivative of $V(z_\mathrm{s}, z_\mathrm{s}, \mathbf{q}_\mathrm{s}, \mathbf{q}_\mathrm{t})$ with respect to c to be

$$\frac{\partial V}{\partial c} = \frac{\partial V}{\partial \mathbf{q}_\mathrm{s}} \cdot \frac{\partial \mathbf{q}_\mathrm{s}}{\partial c} + \frac{\partial V}{\partial \mathbf{q}_\mathrm{t}} \cdot \frac{\partial \mathbf{q}_\mathrm{t}}{\partial c}. \tag{6.42}$$

Comparing equations (6.41) and (6.42) we find

$$\mathbf{p}_\mathrm{s} = -\frac{\partial V}{\partial \mathbf{q}_\mathrm{s}}, \qquad \mathbf{p}_\mathrm{t} = \frac{\partial V}{\partial \mathbf{q}_\mathrm{t}}. \tag{6.43}$$

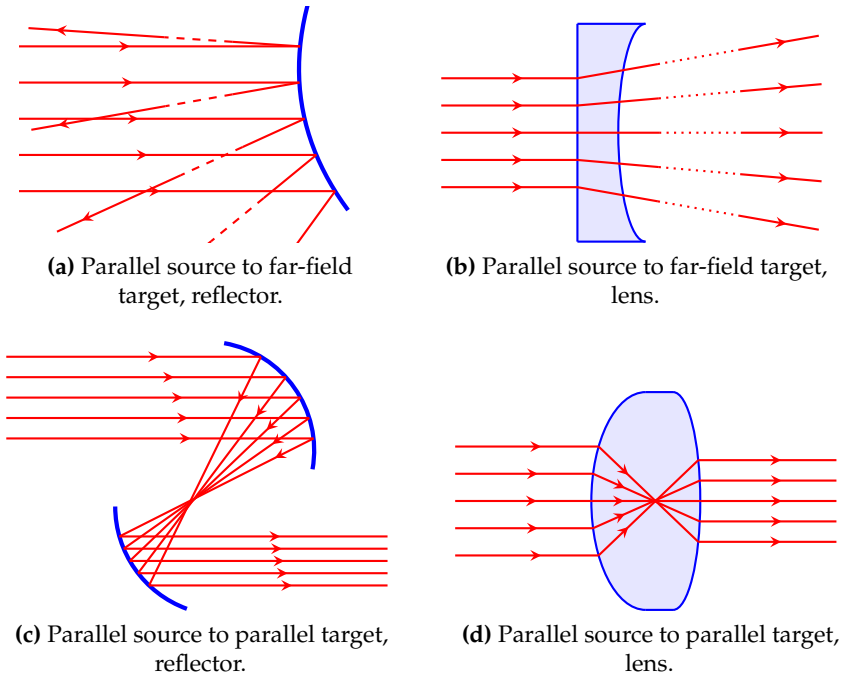
6.2 Freeform optical systems

We consider the design of freeform optical systems where the source and target distributions are specified, and the optical surfaces are to be calculated. The systems to be considered are shown in Figure 6.1. We consider a source with emittance $f(\mathbf{x})$, with $\mathbf{x} \in \mathcal{X}$. Here $\mathcal{X}$ is the source domain and $\mathbf{x}$ are Cartesian coordinates. Furthermore, we consider either a target intensity or illuminance $g(\mathbf{y})$ with $\mathbf{y} \in \mathcal{Y}$ for the target domain $\mathcal{Y}$. The optical map $\mathbf{m} : \mathcal{X} \to \mathcal{Y}$ maps a point on the source domain to a point on the target domain via $\mathbf{y} = \mathbf{m}(\mathbf{x})$. We show that each of the optical systems satisfies a Monge-Ampère equation.

To do so, we first derive a general energy balance which holds for each of the optical systems. We posit, for now, that each optical system individually satisfies the cost balance $u_1(\mathbf{x}) + u_2(\mathbf{y}) = c(\mathbf{x}, \mathbf{y})$, where the function c is the so-called cost function and $u_1(\mathbf{x})$ and $u_2(\mathbf{y})$ are functions related to the optical surface(s). Combining the energy balance with the cost balance results in a Monge-Ampère equation. Lastly we derive the cost balance for each system.

6.2.1 Energy balance

We consider four freeform optical systems as depicted in Figure 6.1. Let $\hat{\mathbf{s}} = \hat{\mathbf{e}}_z$ be the direction of an emitted light ray and $\hat{\mathbf{t}}$ the direction of a light ray incident on the target. Let $A \subseteq \mathcal{X}$. We use Cartesian coordinate $\mathbf{x} \in \mathcal{X}$ for the source domain and either Cartesian or stereographic coordinates, to be introduced in Section 6.2.3, for $\mathbf{y} \in \mathcal{Y}$ on the target domain. Let $J_{\mathcal{Y}}(\mathbf{y}) = \left| \frac{\partial \hat{\mathbf{t}}}{\partial y_1} \times \frac{\partial \hat{\mathbf{t}}}{\partial y_2} \right|$ the Jacobian of the coordinate system for $\hat{\mathbf{t}}$ with respect to $\mathbf{y}$. By conservation of luminous flux, the emitted flux $f(\mathbf{x})$, ends up at the target at $\mathbf{m}(\mathbf{x})$ and should



(a) Parallel source to far-field target, reflector.

(b) Parallel source to far-field target, lens.

(c) Parallel source to parallel target, reflector.

(d) Parallel source to parallel target, lens.

Figure 6.1: Overview of optical systems.

be conserved. Henceforth we have conservation of energy, viz.

$$\begin{aligned}\iint_A f(\mathbf{x})\,\mathrm{d}A(\mathbf{x}) &= \iint_{\mathbf{m}(A)} g(\mathbf{y})J_{\mathcal{Y}}(\mathbf{y})\,\mathrm{d}A(\mathbf{y}) \\ &= \iint_A g(\mathbf{m}(\mathbf{x}))J_{\mathcal{Y}}(\mathbf{m}(\mathbf{x}))\,|\det(\mathrm{D}\mathbf{m}(\mathbf{x}))|\,\mathrm{d}A(\mathbf{x}).\end{aligned} \tag{6.44}$$

Because equation (6.44) should hold for all $A \subseteq \mathcal{X}$, we find the law of conservation of energy in differential form

$$\det(\mathrm{D}\mathbf{m}(\mathbf{x})) = \pm\frac{f(\mathbf{x})}{g(\mathbf{m}(\mathbf{x}))J_{\mathcal{Y}}(\mathbf{m}(\mathbf{x}))}, \tag{6.45}$$

where the absolute value in (6.44) has been replaced by $\pm$. The choice for either the plus or the minus sign determines the shape of the freeform surface(s) to be either convex, concave or saddle shaped. This will become apparent in the next sections.

6.2.2 Monge-Ampère equation

For each of the optical systems we are able to formulate a geometrical description in term of a cost function $c(\mathbf{x}, \mathbf{y})$. The so-called cost balance in terms of $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ reads

$$u_1(\mathbf{x}) + u_2(\mathbf{y}) = c(\mathbf{x}, \mathbf{y}), \tag{6.46}$$

where the functions u_1 and u_2 are related to the optical surface(s) and are, together with the cost function c, derived in Sections 6.2.3, 6.2.4, 6.2.5 and 6.2.6. Hamiltonian optics tells us that there exists a relation, the so-called optical mapping, between a point on the source and the resulting point on the target. We find this relation by first differentiating (6.46) w.r.t. $\mathbf{x}$ to obtain

$$\nabla_{\mathbf{x}}\Big(u_1(\mathbf{x}) - c(\mathbf{x}, \mathbf{y})\Big) = \mathbf{0}. \tag{6.47}$$

For certain cases one can explicitly determine the optical mapping $\mathbf{y} = \mathbf{m}(\mathbf{x})$ from (6.47). For all optical systems we consider here, the existence of $\mathbf{y} = \mathbf{m}(\mathbf{x})$ is established by applying the implicit function theory to (6.47). In this context, the condition under which we can apply the implicit function theory is called the twist condition [77, p. 216], and is satisfied for all optical systems we consider. Henceforth, we assume existence of a unique map $\mathbf{m}$ such that $\mathbf{y} = \mathbf{m}(\mathbf{x})$.

We proceed by finding an expression for $\det(\mathrm{D}\mathbf{m})$ using (6.47), which can be combined with the energy balance (6.45). This is achieved by substituting $\mathbf{y} = \mathbf{m}(\mathbf{x})$ in (6.47) and subsequently differentiating w.r.t. $\mathbf{x}$, yielding

$$\mathrm{D}^2 u_1(\mathbf{x}) - \mathrm{D}_{\mathbf{xx}} c(\mathbf{x}, \mathbf{m}(\mathbf{x})) - \mathrm{D}_{\mathbf{xy}} c(\mathbf{x}, \mathbf{m}(\mathbf{x}))\mathrm{D}\mathbf{m}(\mathbf{x}) = \mathbf{0}. \tag{6.48}$$

The matrices $\mathrm{D}^2 u_1(\mathbf{x})$ and $\mathrm{D}_{\mathbf{xx}} c(\mathbf{x}, \mathbf{m}(\mathbf{x}))$ are Hessian matrices (with respect to $\mathbf{x}$), $\mathrm{D}_{\mathbf{xy}} c(\mathbf{x}, \mathbf{m}(\mathbf{x})) = (c_{x_i, y_j})$ is the matrix of mixed derivatives and $\mathrm{D}\mathbf{m}$ the Jacobi matrix of $\mathbf{m}$. For convenience we introduce matrices $\mathbf{P}$ and $\mathbf{C}$ such that

$$\mathbf{C}\mathrm{D}\mathbf{m} = \mathbf{P}, \quad \mathbf{C}(\mathbf{x}) = \mathrm{D}_{\mathbf{xy}} c(\mathbf{x}, \mathbf{m}(\mathbf{x})), \quad \mathbf{P}(\mathbf{x}) = \mathrm{D}^2 u_1(\mathbf{x}) - \mathrm{D}_{\mathbf{xx}} c(\mathbf{x}, \mathbf{m}(\mathbf{x})). \tag{6.49}$$

Consequently, $\det(\mathbf{P}) = \det(\mathbf{C}) \det(\mathrm{D}\mathbf{m})$ and substitution of $\det(\mathrm{D}\mathbf{m})$ into the energy balance (6.45) gives

$$\det(\mathbf{P}) = \det(\mathrm{D}^2 u_1(\mathbf{x}) - \mathrm{D}_{\mathbf{xx}} c(\mathbf{x}, \mathbf{m}(\mathbf{x}))) = \pm \frac{\det(\mathbf{C}) f(\mathbf{x})}{g(\mathbf{m}(\mathbf{x})) J_{\mathcal{Y}}(\mathbf{m}(\mathbf{x}))}, \tag{6.50}$$

which is a (general) Monge-Ampère equation. The PDE (6.50) with unknown $u_1(\mathbf{x})$ is either elliptic for the plus sign (+) or hyperbolic for the minus sign (-). To see this, note that (6.50) can be written as

$$\begin{aligned} \det(\mathrm{D}^2 u_1(\mathbf{x})) + \det(\mathrm{D}_{\mathbf{xx}} c(\mathbf{x}, \mathbf{m}(\mathbf{x}))) - c_{22} u_{1,x_1 x_1} \\ + 2c_{12} u_{1,x_1 x_2} - c_{11} u_{1,x_2 x_2} \mp \frac{\det(\mathbf{C}) f(\mathbf{x})}{g(\mathbf{m}(\mathbf{x})) J_{\mathcal{Y}}(\mathbf{m}(\mathbf{x}))} = 0, \end{aligned} \tag{6.51}$$

where we used the notation $\mathrm{D}_{\mathbf{xx}} c(\mathbf{x}, \mathbf{m}(\mathbf{x})) = (c_{ij})$ and u_{1,x_1} for the derivative of u_1 w.r.t. x_1 and similar for the others. According to the classification of PDEs established in Section 2.1, by equations (2.1) and (2.9) we have

$$\begin{aligned} \Delta &= 4A_4 + A_2^2 - 4A_1 A_3 \\ &= 4\Big(\det(\mathrm{D}_{\mathbf{xx}} c(\mathbf{x}, \mathbf{m}(\mathbf{x}))) \mp \frac{\det(\mathbf{C}) f(\mathbf{x})}{g(\mathbf{m}(\mathbf{x})) J_{\mathcal{Y}}(\mathbf{m}(\mathbf{x}))} \Big) + 4c_{12}^2 - 4c_{11} c_{22} \\ &= \mp \frac{\det(\mathbf{C}) f(\mathbf{x})}{g(\mathbf{m}(\mathbf{x})) J_{\mathcal{Y}}(\mathbf{m}(\mathbf{x}))}. \end{aligned} \tag{6.52}$$

For the cost functions and $J_{\mathcal{Y}}$, as given in the next sections, we have $\det(\mathbf{C}) > 0$, $J_{\mathcal{Y}} > 0$. Furthermore, because $f, g > 0$, if the + sign is chosen in (6.50) then $\Delta < 0$ and (6.50) is elliptic. Similarly, if the - sign is chosen then $\Delta > 0$ and (6.50) is hyperbolic.

For both the elliptic and hyperbolic variant, (6.50) describes solutions for u_1, which we solve using (an adjusted version of) the least-squares solver

as shown in Section 7.1. Note, whether a solution exists (6.50) exists is not straightforward. Proving existence for the elliptic case is done using results from Optimal Transport Theory, and can for example be found in Romijn's thesis [68, Ch. 4]. As far as the author is aware, currently no existence results exist for the hyperbolic case.

6.2.3 Parallel-to-far-field reflector

The first optical system we consider consists of one reflector, a parallel source and a far-field target, as shown in Figure 6.2. We assume the light source to be located at $(\mathbf{q}_s, z_s)$ with $z_s = 0$ and the target at $(\mathbf{q}_t, z_t)$ with $z_t = -L$ for $L > 0$. Let $z = u(\mathbf{x})$ be the surface of the reflector. We set the refractive index to $n = 1$. Because the source coordinates $\mathbf{x}$ are Cartesian and the target is in the far-field, i.e., directional, we use the mixed characteristics of the first kind, i.e., $W = W(z_s, z_t, \mathbf{q}_s, \mathbf{p}_t)$. The source ray $\hat{\mathbf{s}}$ originates at $O_s = (\mathbf{q}_s, 0)$, we write $\mathbf{q}_s = \mathbf{x}$ and we have $\mathbf{p}_s = \mathbf{0}$. The reflected ray $\hat{\mathbf{t}}$ intersects the target at $O_t = (\mathbf{q}_t, -L)$ with position and direction coordinates $\mathbf{q}_t$ and $\mathbf{p}_t$. The source ray intersects the reflector at $P = (\mathbf{x}, u(\mathbf{x}))$ and the target at $(\mathbf{q}_t, z_t)$ as shown in Figure 6.2.

By the definition of the mixed characteristic W, given by (6.37c), we have

$$W(z_s, z_t, \mathbf{q}_s, \mathbf{p}_t) = V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t) - \mathbf{q}_t \cdot \mathbf{p}_t. \tag{6.53}$$

Because $n = 1$, the optical path length between O_s and O_t equals the Euclidean

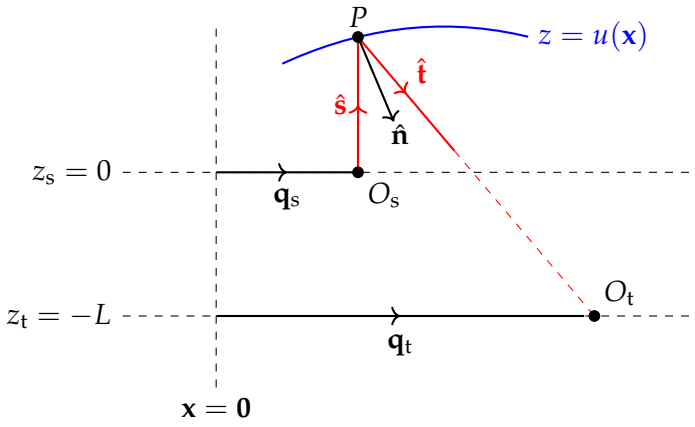


Figure 6.2: Schematic representation of the reflector setup for a parallel source and far-field target.

distance along the ray, i.e.,

$$V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t) = u(\mathbf{x}) + \sqrt{|\mathbf{q}_t - \mathbf{x}|^2 + (u(\mathbf{x}) + L)^2}. \tag{6.54}$$

The mixed characteristic W is thus given by

$$W(\mathbf{q}_s, \mathbf{p}_t) = u(\mathbf{x}) + \sqrt{|\mathbf{q}_t - \mathbf{x}|^2 + (u(\mathbf{x}) + L)^2} - \mathbf{q}_t \cdot \mathbf{p}_t. \tag{6.55}$$

By connecting the points P and O_t, we find that

$$\mathbf{r}_{O_t} = \mathbf{r}_P + d(P, O_t)\hat{\mathbf{t}}, \tag{6.56}$$

with $\mathbf{r}_{O_t}$ the position vector of O_t and $d(P, O_t)$ the distance between P and O_t. Using $O_t = (\mathbf{q}_t, -L)$ and $\hat{\mathbf{t}} = (\mathbf{p}_t, t_3)$ we deduce that

$$\mathbf{p}_t = \frac{\mathbf{q}_t - \mathbf{x}}{\sqrt{|\mathbf{q}_t - \mathbf{x}|^2 + (u(\mathbf{x}) + L)^2}}, \quad t_3 = -\frac{L + u(\mathbf{x})}{\sqrt{|\mathbf{q}_t - \mathbf{x}|^2 + (u(\mathbf{x}) + L)^2}}. \tag{6.57}$$

Furthermore, W is independent of $\mathbf{q}_s$ by $\mathbf{p}_s = 0$ and equation (6.37d). Therefore, W reads

$$\begin{aligned} W(\mathbf{p}_t) &= u(\mathbf{x}) + \frac{|\mathbf{q}_t - \mathbf{x}|^2 - \mathbf{q}_t \cdot (\mathbf{q}_t - \mathbf{x}) + (u(\mathbf{x}) + L)^2}{\sqrt{|\mathbf{q}_t - \mathbf{x}|^2 + (u(\mathbf{x}) + L)^2}} \\ &= u(\mathbf{x}) - \mathbf{x} \cdot \mathbf{p}_t - (L + u(\mathbf{x}))t_3 \\ &= u(\mathbf{x})(1 - t_3) - x_1 t_1 - x_2 t_2 - L t_3, \end{aligned} \tag{6.58}$$

where we used (6.57) for the second equal sign. This expression can be rewritten as

$$u(\mathbf{x}) - \frac{W(\mathbf{p}_t) + L t_3}{1 - t_3} = \frac{x_1 t_1 + x_2 t_2}{1 - t_3}. \tag{6.59}$$

We aim to prescribe the target in stereographic coordinates $\mathbf{y}$, thus requiring a transformation from $\hat{\mathbf{t}}$ to $\mathbf{y}$. For this we use a stereographic projection from the north pole. The projection and its inverse read [68, p. 60]

$$\mathbf{y} = \frac{1}{1 - t_3}\begin{pmatrix} t_1 \\ t_2 \end{pmatrix}, \qquad \hat{\mathbf{t}} = \frac{1}{|\mathbf{y}|^2 + 1}\begin{pmatrix} 2y_1 \\ 2y_2 \\ |\mathbf{y}|^2 - 1 \end{pmatrix}. \tag{6.60}$$

Applying this coordinate transformation, equation (6.59) becomes

$$u(\mathbf{x}) - w(\mathbf{y}) = \mathbf{x} \cdot \mathbf{y}, \qquad w(\mathbf{y}) := \frac{W(\mathbf{p}_t) + L t_3}{1 - t_3}. \tag{6.61}$$

Note, we write $w(\mathbf{y})$ though its right-hand side is not in terms of $\mathbf{y}$, though it can solely be written as such due to $\hat{\mathbf{t}} = (\mathbf{p}_t, t_3)^T$ and the projection (6.60). We omit the details here as these are not needed for the remainder of this section. By writing $u_1(\mathbf{x}) = u(\mathbf{x})$ and $u_2(\mathbf{y}) = -w(\mathbf{y})$ we obtain the cost balance

$$u_1(\mathbf{x}) + u_2(\mathbf{y}) = \mathbf{x} \cdot \mathbf{y} =: c(\mathbf{x}, \mathbf{y}). \tag{6.62}$$

In the literature on optimal transport [77], the cost function $c(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$ is known as a quadratic cost function. It follows that

$$\mathbf{C} = \mathrm{D}_{\mathbf{xy}} c(\mathbf{x}, \mathbf{y}) = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -\mathbf{I}, \tag{6.63}$$

and $\det(\mathbf{C}) = 1$. Lastly we need to calculate the Jacobian $J_y(\mathbf{y})$. We find, using (6.60), that

$$J_y(\mathbf{y}) = \left| \frac{\partial \hat{\mathbf{t}}}{\partial y_1} \times \frac{\partial \hat{\mathbf{t}}}{\partial y_2} \right| = \frac{4}{(1 + |\mathbf{y}|^2)^2}. \tag{6.64}$$

6.2.4 Parallel-to-far-field lens

The second optical system we consider consists of one lens with one refracting surface, a parallel source and a far-field target, as shown in Figure 6.3. We assume the light source to be located at $(\mathbf{q}_s, z_s)$ with $z_s = 0$ and the target at $(\mathbf{q}_t, z_t)$ with $z_t = L$ for $L > 0$. Let $z = u(\mathbf{x})$ be the refracting surface of the lens and we fix the refractive index $n > 1$.

Again, because the source coordinates $\mathbf{x}$ are Cartesian and the target is far-field, we use the mixed characteristics of the first kind, i.e., $W = W(z_s, z_t, \mathbf{q}_s, \mathbf{p}_t)$. The source ray $\hat{\mathbf{s}}$ originates at O_s with position and

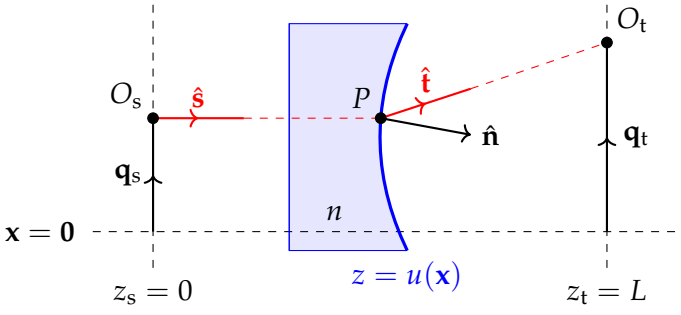


Figure 6.3: Schematic representation of the lens setup for a parallel source and far-field target.

direction coordinates $\mathbf{q}_s$ and $\mathbf{p}_s$. We write $\mathbf{q}_s = \mathbf{x}$ and we have $\mathbf{p}_s = 0$. The source ray refracts in direction $\hat{\mathbf{t}}$ at the point $P = (\mathbf{x}, u(\mathbf{x}))$ and intersects the target at the point O_t with position $(\mathbf{q}_t, z_t)$ and direction $\mathbf{p}_t$ as shown in Figure 6.3.

By the definition of the mixed characteristic W, given by (6.37c), we have

$$W(z_s, z_t, \mathbf{q}_s, \mathbf{p}_t) = V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t) - \mathbf{q}_t \cdot \mathbf{p}_t. \tag{6.65}$$

Because the first surface of the lens does not refract the rays, we assume the lens to extend till the source. The optical path length between O_s and O_t then becomes

$$V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t) = n \cdot u(\mathbf{x}) + \sqrt{|\mathbf{q}_t - \mathbf{x}|^2 + (L - u(\mathbf{x}))^2}. \tag{6.66}$$

By connecting the points P and O_t, we find that

$$\mathbf{p}_t = \frac{\mathbf{q}_t - \mathbf{x}}{\sqrt{|\mathbf{q}_t - \mathbf{x}|^2 + (L - u(\mathbf{x}))^2}}, \quad t_3 = \frac{L - u(\mathbf{x})}{\sqrt{|\mathbf{q}_t - \mathbf{x}|^2 + (L - u(\mathbf{x}))^2}}, \tag{6.67}$$

and the mixed characteristic W is thus given by

$$\begin{aligned} W(\mathbf{p}_t) &= n \cdot u(\mathbf{x}) + \sqrt{|\mathbf{q}_t - \mathbf{x}|^2 + (L - u(\mathbf{x}))^2} - \mathbf{q}_t \cdot \mathbf{p}_t \\ &= n \cdot u(\mathbf{x}) + \frac{|\mathbf{q}_t - \mathbf{x}|^2 - \mathbf{q}_t \cdot (\mathbf{q}_t - \mathbf{x}) + (L - u(\mathbf{x}))^2}{\sqrt{|\mathbf{q}_t - \mathbf{x}|^2 + (L - u(\mathbf{x}))^2}} \\ &= n \cdot u(\mathbf{x}) - \mathbf{x} \cdot \mathbf{p}_t + (L - u(\mathbf{x}))t_3 \\ &= u(\mathbf{x})(n - t_3) - x_1 t_1 - x_2 t_2 + L t_3, \end{aligned} \tag{6.68}$$

where we used that W is independent of $\mathbf{q}_s$ by $\mathbf{p}_s = 0$ and equation (6.37d). We rewrite (6.68) as

$$u(\mathbf{x}) - \frac{W(\mathbf{p}_t) - L t_3}{n - t_3} = \frac{x_1 t_1 + x_2 t_2}{n - t_3}. \tag{6.69}$$

As for the reflector case, we require a transformation from $\hat{\mathbf{t}}$ to $\mathbf{y}$. For this case we again use a stereographic projection but now we do not project from the north pole $(0, 0, 1)$, but from $(0, 0, n)$ instead. The orthogonal projection and its inverse are given by [82, p. 39-40]

$$\mathbf{y} = \frac{n}{n - t_3} \begin{pmatrix} t_1 \\ t_2 \end{pmatrix}, \quad \hat{\mathbf{t}} = \frac{1}{|\mathbf{y}|^2 + n^2} \begin{pmatrix} n^2 + y_1 \sqrt{n^2 + (1 - n^2)|\mathbf{y}|^2} \\ n^2 + y_2 \sqrt{n^2 + (1 - n^2)|\mathbf{y}|^2} \\ n|\mathbf{y}|^2 - n\sqrt{n^2 + (1 - n^2)|\mathbf{y}|^2} \end{pmatrix}. \tag{6.70}$$

Applying this stereographic projection to (6.69) we obtain

$$u(\mathbf{x}) - w(\mathbf{y}) = \frac{1}{n}\mathbf{x}\cdot\mathbf{y}, \qquad w(\mathbf{y}) := \frac{W(\mathbf{p}_t) - Lt_3}{n - t_3}. \tag{6.71}$$

By splitting the scalar product $\mathbf{x}\cdot\mathbf{y}$ and subsequently introducing the auxiliary functions

$$u_1(\mathbf{x}) = \tfrac{1}{2}|\mathbf{x}|^2 - nu(\mathbf{x}), \qquad u_2(\mathbf{y}) = \tfrac{1}{2}|\mathbf{y}|^2 + nw(\mathbf{y}), \tag{6.72}$$

we obtain the cost balance

$$u_1(\mathbf{x}) + u_2(\mathbf{y}) = \tfrac{1}{2}|\mathbf{x} - \mathbf{y}|^2 =: c(\mathbf{x}, \mathbf{y}). \tag{6.73}$$

Analogously to the parallel to far-field reflector system, we find $\mathbf{C} = -\mathbf{I}$. and $J_{\mathcal{Y}}(\mathbf{y}) = 4(1 + |\mathbf{y}|^2)^{-2}$.

6.2.5 Parallel-to-parallel reflectors

For the next example we consider an optical system consisting of a parallel source, two freeform reflectors and a parallel target, as shown in Figure 6.4. Let the source be positioned in the plane $z = 0$ with Cartesian coordinates $\mathbf{x} \in \mathcal{X}$. Let the target be positioned in the plane $z = L$ with $L > 0$ and Cartesian coordinates $\mathbf{y} \in \mathcal{Y}$. We assume both a parallel source and target, i.e., the direction of both a source ray and a target ray equal $\hat{\mathbf{e}}_z$. Let the first reflector surface be given by $z = u_1(\mathbf{x})$ and the second by $z = L - u_2(\mathbf{y})$.

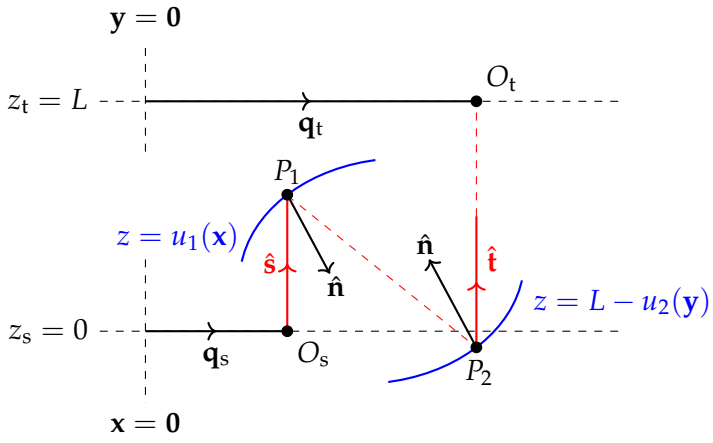


Figure 6.4: Schematic representation of the reflector setup for a parallel source and parallel target.

Such a parametrization is possible because both source and target rays are perpendicular to the planes $z = \text{const}$. We assume $n = 1$. Because $\mathbf{p}_s = \mathbf{p}_t = \mathbf{0}$ we use the point characteristic, i.e., the optical path length

$$V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t) = V(z_s, z_t, \mathbf{x}, \mathbf{y}) = u_1(\mathbf{x}) + d(P_1, P_2) + u_2(\mathbf{y}), \tag{6.74}$$

where we parametrize $\mathbf{q}_s$ by $\mathbf{x}$, $\mathbf{q}_t$ by $\mathbf{y}$ and where $d(P_1, P_2)$ is the Euclidean distance between $P_1 = (\mathbf{x}, u_1(\mathbf{x}))$ and $P_2 = (\mathbf{y}, L - u_2(\mathbf{y}))$. Because both the source and target are considered parallel, we have $\mathbf{p}_s = \mathbf{p}_t = \mathbf{0}$. By (6.37a) the optical path length $V(z_s, z_t, \mathbf{x}, \mathbf{y})$ is independent of $\mathbf{q}_s$ and $\mathbf{q}_t$, i.e., it is independent of the position vectors $\mathbf{x}$ and $\mathbf{y}$ and therefore $V(z_s, z_t, \mathbf{x}, \mathbf{y}) = V > 0$. The optical distance between the two reflectors, which is equal to the Euclidean distance between P_1 and P_2 due to $n = 1$, can be written as

$$d^2(\mathbf{x}, \mathbf{y}) = (u_1(\mathbf{x}) + u_2(\mathbf{y}) - L)^2 + |\mathbf{x} - \mathbf{y}|^2. \tag{6.75}$$

Substitution of (6.75) into (6.74), and subsequently reordering, yields

$$u_1(\mathbf{x}) + u_2(\mathbf{y}) = \frac{\beta^2 + 2\beta L - |\mathbf{x} - \mathbf{y}|^2}{2\beta} =: c(\mathbf{x}, \mathbf{y}), \tag{6.76}$$

which is the sought cost balance and cost function c, with $\beta = V - L$ the so-called reduced optical path length [72]. It follows that

$$\mathbf{C} = \mathrm{D}_{\mathbf{xy}} c(\mathbf{x}, \mathbf{y}) = \frac{1}{\beta}\mathbf{I}, \tag{6.77}$$

and $\det(\mathbf{C}) = \frac{1}{\beta^2}$. Lastly we need to calculate the Jacobian $J_{\mathcal{Y}}(\mathbf{y})$. Because $\mathbf{y}$ are Cartesian coordinates, we have $J_{\mathcal{Y}}(\mathbf{y}) = 1$.

6.2.6 Parallel-to-parallel lens

As a last optical system we consider a parallel source, a parallel target and a lens with two freeform surfaces, as shown in Figure 6.5. Let the source be positioned in the plane $z = 0$ with Cartesian coordinates $\mathbf{x} \in \mathcal{X}$. Let the target be positioned in the plane $z = L$ with $L > 0$ and Cartesian coordinates $\mathbf{y} \in \mathcal{Y}$. We assume both a parallel source and target with the direction of both source and target rays equal to $\hat{\mathbf{e}}_z$. Let n_o be the refractive index outside of the lens, and n_i the refractive index inside the lens. Let the first refracting surface be given by $z = u_1(\mathbf{x})$ and the second by $z = L - u_2(\mathbf{y})$. Such a parametrization is possible beacause both the source and target rays are perpendicular to the

planes $z = \text{const}$. Because $\mathbf{p}_s = \mathbf{p}_t = \mathbf{0}$, we use the point characteristic, i.e., the optical path length

$$V(z_s, z_t, \mathbf{q}_s, \mathbf{q}_t) = V(z_s, z_t, \mathbf{x}, \mathbf{y}) = n_o u_1(\mathbf{x}) + n_i d(\mathbf{x}, \mathbf{y}) + n_o u_2(\mathbf{y}), \tag{6.78}$$

where we parametrize $\mathbf{q}_s$ by $\mathbf{x}$, $\mathbf{q}_t$ by $\mathbf{y}$ and where $d(P_1, P_2)$ is the Euclidean distance between $P_1 = (\mathbf{x}, u_1(\mathbf{x}))$ and $P_2 = (\mathbf{y}, L - u_2(\mathbf{y}))$. By (6.37a) the optical path length $V(z_s, z_t, \mathbf{x}, \mathbf{y})$ is independent of $\mathbf{q}_s$ and $\mathbf{q}_t$, i.e., it is independent of the position vectors $\mathbf{x}$ and $\mathbf{y}$ and therefore $V(z_s, z_t, \mathbf{x}, \mathbf{y}) = V > 0$. The Euclidean distance between P_1 and P_2 can be written as

$$d^2(\mathbf{x}, \mathbf{y}) = (u_1(\mathbf{x}) + u_2(\mathbf{y}) - L)^2 + |\mathbf{x} - \mathbf{y}|^2. \tag{6.79}$$

Combining equations (6.78) and (6.79) yields

$$n_i^2 (u_1(\mathbf{x}) + u_2(\mathbf{y}) - L)^2 + n_i^2 |\mathbf{x} - \mathbf{y}|^2 = (V - n_o u_1(\mathbf{x}) - n_o u_2(\mathbf{y}))^2. \tag{6.80}$$

Subsequently solving for $u_1(\mathbf{x}) + u_2(\mathbf{y})$ we find

$$u_1(\mathbf{x}) + u_2(\mathbf{y}) = c_\pm(\mathbf{x}, \mathbf{y}), \tag{6.81}$$

where, using $\beta = V - n_o L$, the cost function is given by

$$c_\pm(\mathbf{x}, \mathbf{y}) = L + \frac{n_o \beta}{n_o^2 - n_i^2} \pm \frac{n_i}{n_o^2 - n_i^2} \sqrt{\beta^2 + (n_o^2 - n_i^2)|\mathbf{x} - \mathbf{y}|^2}, \tag{6.82}$$

with the sign in front of the square root yet undetermined and the implied condition $\beta^2 + (n_o^2 - n_i^2)|\mathbf{x} - \mathbf{y}|^2 > 0$. To choose the sign, we consider a practical example with a source in air, i.e., $n_o = 1$ and a lens with, for example, plastic or glass material with, approximately, $1.3 < n_i < 1.7$ [57]. We therefore

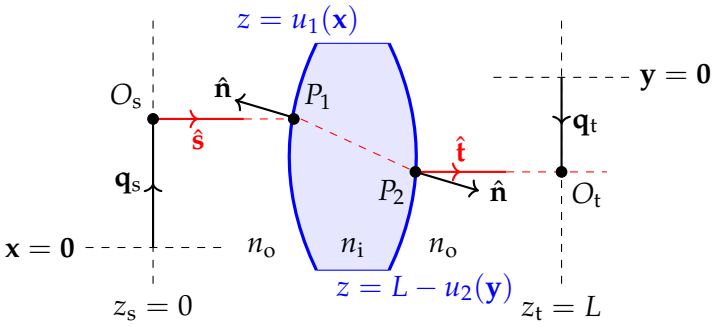


Figure 6.5: Schematic representation of the reflector setup for a parallel source and parallel target.

have $n_o < n_i$. Furthermore, because we assume u_1 to be the first surface, and u_2 to be the second, we have $u_1(\mathbf{x}) + u_2(\mathbf{y}) < L$. Combining this inequality with equations (6.81), (6.82) and $n_o < n_i$, we obtain

$$n_o\beta \pm n_i\sqrt{\beta^2 + (n_o^2 - n_i^2)|\mathbf{x}-\mathbf{y}|^2} > 0. \tag{6.83}$$

By combining $V(z_s, z_t, \mathbf{x}, \mathbf{y}) = V > 0$, equations (6.78) and (6.79), we find

$$\beta^2 + (n_o^2 - n_i^2)|\mathbf{x}-\mathbf{y}|^2 = \frac{1}{n_o^2}(n_i\beta + d(n_o^2 - n_i^2))^2 \geq 0. \tag{6.84}$$

Because $\beta = V - n_o L > 0$, by choosing the plus sign in (6.82), condition (6.83) is automatically satisfied. We therefore choose the plus sign and henceforth have

$$u_1(\mathbf{x}) + u_2(\mathbf{y}) = L + \frac{n_o\beta}{n_o^2 - n_i^2} + \frac{n_i}{n_o^2 - n_i^2}\sqrt{\beta^2 + (n_o^2 - n_i^2)|\mathbf{x}-\mathbf{y}|^2} =: c(\mathbf{x}, \mathbf{y}). \tag{6.85}$$

Calculating the derivatives of c straightforwardly shows that

$$\mathbf{C} = \frac{n_i}{r^3}\left[(n_o^2 - n_i^2)\mathbf{J}(\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^T\mathbf{J} - \beta^2\mathbf{I}\right], \tag{6.86a}$$

$$\mathbf{J} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{6.86b}$$

$$r = \sqrt{\beta^2 + (n_o^2 - n_i^2)|\mathbf{x}-\mathbf{y}|^2}, \tag{6.86c}$$

and subsequently we obtain

$$\det(\mathbf{C}) = \frac{\beta^2 n_i^2}{(\beta^2 - (n_i^2 - n_o^2)|\mathbf{x}-\mathbf{y}|^2)^2}. \tag{6.87}$$

Lastly, the Jacobian reads $J_{\mathcal{Y}}(\mathbf{y}) = 1$.

6.3 Summary

In this chapter we provided a brief introduction to optics. We introduced the fundamentals of geometrical optics and showed how these can be derived from the Maxwell equations. By means of a short wave length approximation, we derived the eikonal equations laying the basis for the ray equation, Fermat's principle and Hamiltonian optics. By use of Hamilton's characteristics we derived cost functions and corresponding cost balances for the optical elements of four freeform optical systems.

Chapter 7

Least-Squares Solutions to four Optical Systems

In this chapter we present numerical results for four distinct optical systems.

To calculate the optical surfaces, we propose a least-squares solver. This solver is an adaption of the least-squares method introduced in the Chapter 5, such that it includes the cost balances introduced in Chapter 6.

For each of the optical systems, we show that there exist multiple distinct solutions and present four. Two such solutions are so-called c-convex and c-concave solutions and are solutions to an elliptic Monge-Ampère equation [68]. The other two solutions we name c-saddle solutions and are solutions to a hyperbolic Monge-Ampère equation. These solutions differ in the way light rays are mapped from the source to the target. This mapping is called the optical map. Furthermore, we present the results of ray-tracing the optical systems and compare the various solutions.

7.1 Adjusted Least-squares method

In this section we present a least-squares method for the design problem of the optical systems presented in Section 6.2. The aim is to adjust the least-squares method introduced in Chapter 5 to include the cost balances derived in Chapter 6. The Monge-Ampère equation introduced in Section 5.1.1, given by

$$\det\left(\mathrm{D}\mathbf{m}(\mathbf{x})\right) + f^2(\mathbf{x}, \mathbf{m}(\mathbf{x})) = 0, \quad \mathbf{x} \in \mathcal{X}, \tag{7.1}$$

and the Monge-Ampère equation of Section 6.2.2, i.e.,

$$\det(\mathbf{P}) = \det(\mathrm{D}^2 u_1(\mathbf{x}) - \mathrm{D}_{\mathbf{xx}} c(\mathbf{x}, \mathbf{m}(\mathbf{x}))) = \pm \frac{\det(\mathbf{C}) f(\mathbf{x})}{g(\mathbf{m}(\mathbf{x})) J_{\mathcal{Y}}(\mathbf{m}(\mathbf{x}))}, \tag{7.2}$$

differ slightly. Certain parts of the least-squares method introduced in Section 5.1.1 remain the same, so we only highlight the differences. Recall equations (6.49), viz.

$$\mathbf{C}\mathrm{D}\mathbf{m} = \mathbf{P}, \quad \mathbf{C}(\mathbf{x}) = \mathrm{D}_{\mathbf{xy}} c(\mathbf{x}, \mathbf{m}(\mathbf{x})), \quad \mathbf{P}(\mathbf{x}) = \mathrm{D}^2 u_1(\mathbf{x}) - \mathrm{D}_{\mathbf{xx}} c(\mathbf{x}, \mathbf{m}(\mathbf{x})). \tag{7.3}$$

Instead of approximating $\mathrm{D}\mathbf{m}$ by $\mathbf{P}$, as we did on basis of equation (7.1), we now approximate $\mathbf{C}\mathrm{D}\mathbf{m}$ by $\mathbf{P}$. We do so by minimizing the functionals

$$J_{\mathrm{I}}(\mathbf{m}, \mathbf{P}) = \tfrac{1}{2} \iint_{\mathcal{X}} \|\mathbf{C}\mathrm{D}\mathbf{m} - \mathbf{P}\|^2 \, \mathrm{d}\mathbf{x}, \tag{7.4a}$$

$$J_{\mathrm{B}}(\mathbf{m}, \mathbf{b}) = \tfrac{1}{2} \oint_{\partial\mathcal{X}} |\mathbf{m} - \mathbf{b}|^2 \, \mathrm{d}s, \tag{7.4b}$$

$$J(\mathbf{m}, \mathbf{P}, \mathbf{b}) = \alpha J_{\mathrm{I}}(\mathbf{m}, \mathbf{P}) + (1 - \alpha) J_{\mathrm{B}}(\mathbf{m}, \mathbf{b}), \tag{7.4c}$$

iteratively, according to

$$\mathbf{P}^{n+1} = \operatorname*{argmin}_{\mathbf{P} \in \mathcal{P}(\mathbf{m}^n)} J_{\mathrm{I}}(\mathbf{m}^n, \mathbf{P}), \tag{7.5a}$$

$$\mathbf{b}^{n+1} = \operatorname*{argmin}_{\mathbf{b} \in \mathcal{B}} J_{\mathrm{B}}(\mathbf{m}^n, \mathbf{b}), \tag{7.5b}$$

$$\mathbf{m}^{n+1} = \operatorname*{argmin}_{\mathbf{m} \in \mathcal{M}} J(\mathbf{m}, \mathbf{P}^{n+1}, \mathbf{b}^{n+1}), \tag{7.5c}$$

where the spaces $\mathcal{B}$ and $\mathcal{M}$ are given by (5.10) and

$$\mathcal{P}(\mathbf{m}) = \left\{ \mathbf{P} \in [C^1(\mathcal{X})]^{2\times 2} \mid \det(\mathbf{P}(\mathbf{x})) = -\frac{\det(\mathbf{C}) f(\mathbf{x})}{g(\mathbf{m}(\mathbf{x})) J_{\mathcal{Y}}(\mathbf{m}(\mathbf{x}))}, \, \mathbf{P} = \mathbf{P}^{\mathrm{T}} \right\}. \tag{7.6}$$

Here, we have chosen the minus sign in (7.2), which corresponds to a hyperbolic solution. For a treatment of the elliptic variant we refer to [68, Ch. 6]. Upon convergence of (7.5) we reconstruct u_1 from $\mathbf{m}$ by minimizing a least-squares functional, based on (6.47), viz.

$$u = \operatorname*{argmin}_{\psi \in C^2(\mathcal{X})} I(\psi), \qquad I(\psi) := \tfrac{1}{2} \iint_{\mathcal{X}} |\nabla \psi - \nabla_{\mathbf{x}} c(\mathbf{x}, \mathbf{m}(\mathbf{x}))|^2 \, \mathrm{d}\mathbf{x}. \tag{7.7}$$

Let us consider the minimization procedures for (7.4)-(7.5) and (7.7). The minimization of (7.4a), as given in Section 5.1.2, changes only slightly. The minimization remains point-wise, but instead of minimizing $\|\mathrm{D}\mathbf{m} - \mathbf{P}\|^2$ under the constraint $\det \mathbf{P} = -f^2(\mathbf{x}, \mathbf{m})$ with $\mathbf{P} = \mathbf{P}^{\mathrm{T}}$, we minimize $\|\mathbf{C}\mathrm{D}\mathbf{m} - \mathbf{P}\|^2$ under the constraint $\det \mathbf{P} = -\det(\mathbf{C}) f(\mathbf{x}) / (g(\mathbf{m}(\mathbf{x})) J_{\mathcal{Y}}(\mathbf{m}(\mathbf{x})))$ with $\mathbf{P} = \mathbf{P}^{\mathrm{T}}$. The method outlined in Section 5.1.2 can still be used for the current case by formally replacing $f^2(\mathbf{x}, \mathbf{m})$ by $\det(\mathbf{C}) f(\mathbf{x}) / g(\mathbf{m}(\mathbf{x})) J_{\mathcal{Y}}(\mathbf{m}(\mathbf{x}))$. In Section 5.1.2 we approximated the Jacobian matrix $\mathrm{D}\mathbf{m}$ by $\mathbf{D}$ using finite difference. Subsequently, we used the symmetric part of $\mathbf{D}$, denoted by $\mathbf{D}_{\mathrm{s}}$, for the remainder of the section. Now we define $\mathbf{D}_{\mathrm{s}}$ as the symmetric part of $\mathbf{CD}$ such that the derivations in Section 5.1.2 remain valid, i.e., we let $\mathbf{D}_{\mathrm{s}} = \frac{1}{2}\big(\mathbf{CD} + (\mathbf{CD})^{\mathrm{T}}\big)$.

The minimization of (7.4b) remains unchanged and is given in Section 5.1.3.

The minimization of (7.4c) changes significantly, so let us consider the current case. The first variation of J_{I} with respect to $\mathbf{m}$ is given by

$$\begin{aligned}
\delta J_{\mathrm{I}}(\mathbf{m}, \mathbf{P})(\boldsymbol{\eta}) &= \lim_{\epsilon \to 0} \frac{J_{\mathrm{I}}(\mathbf{m} + \epsilon\boldsymbol{\eta}, \mathbf{P}) - J_{\mathrm{I}}(\mathbf{m}, \mathbf{P})}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{1}{2} \iint_{\mathcal{X}} \Big(2(\mathbf{C}\mathrm{D}\mathbf{m} - \mathbf{P}) : \mathbf{C}\mathrm{D}\boldsymbol{\eta} + \epsilon \|\mathbf{C}\mathrm{D}\boldsymbol{\eta}\|^2 \Big) \, \mathrm{d}\mathbf{x} \qquad (7.8) \\
&= \iint_{\mathcal{X}} (\mathbf{C}\mathrm{D}\mathbf{m} - \mathbf{P}) : \mathbf{C}\mathrm{D}\boldsymbol{\eta} \, \mathrm{d}\mathbf{x},
\end{aligned}$$

with $:$ the Frobenius inner product. Let $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2] = \mathbf{C}^{\mathrm{T}}(\mathbf{C}\mathrm{D}\mathbf{m} - \mathbf{P})$ and $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2] = \mathbf{V}^{\mathrm{T}}$, then

$$\begin{aligned}
(\mathbf{C}\mathrm{D}\mathbf{m} - \mathbf{P}) : \mathbf{C}\mathrm{D}\boldsymbol{\eta} &= \mathbf{C}^{\mathrm{T}}(\mathbf{C}\mathrm{D}\mathbf{m} - \mathbf{P}) : \mathrm{D}\boldsymbol{\eta} \\
&= \mathbf{V} : \mathrm{D}\boldsymbol{\eta} = \mathbf{V}^{\mathrm{T}} : (\mathrm{D}\boldsymbol{\eta})^{\mathrm{T}} \\
&= \mathbf{w}_1 \cdot \nabla \eta_1 + \mathbf{w}_2 \cdot \nabla \eta_2 \\
&= \nabla \cdot (\eta_1 \mathbf{w}_1 + \eta_2 \mathbf{w}_2) - (\eta_1 \nabla \cdot \mathbf{w}_1 + \eta_2 \nabla \cdot \mathbf{w}_2) \qquad (7.9) \\
&= \nabla \cdot (\mathbf{W}\boldsymbol{\eta}) - \boldsymbol{\eta} \cdot \begin{pmatrix} \nabla \cdot \mathbf{w}_1 \\ \nabla \cdot \mathbf{w}_2 \end{pmatrix} \\
&= \nabla \cdot (\mathbf{V}^{\mathrm{T}}\boldsymbol{\eta}) - \boldsymbol{\eta} \cdot (\nabla \cdot \mathbf{V}),
\end{aligned}$$

where we used the definitions of $\mathbf{V}$ and $\mathbf{W}$ in the first and second step, and the divergence of $\mathbf{V}$ in the last step, which is defined by

$$\nabla \cdot \mathbf{V} = \begin{pmatrix} \frac{\partial v_{11}}{\partial x_1} + \frac{\partial v_{12}}{\partial x_2} \\ \frac{\partial v_{21}}{\partial x_1} + \frac{\partial v_{22}}{\partial x_2} \end{pmatrix} = \begin{pmatrix} \nabla \cdot \mathbf{w}_1 \\ \nabla \cdot \mathbf{w}_2 \end{pmatrix}. \qquad (7.10)$$

Using the divergence theorem we can write the first variation of J_{I} as

$$\begin{aligned}\delta J_{\mathrm{I}}(\mathbf{m},\mathbf{P})(\boldsymbol{\eta}) &= \iint_{\mathcal{X}} \Big(\nabla\cdot(\mathbf{V}^{\mathrm{T}}\boldsymbol{\eta}) - \boldsymbol{\eta}\cdot(\nabla\cdot\mathbf{V})\Big)\,\mathrm{d}\mathbf{x} \\ &= \oint_{\partial\mathcal{X}} (\mathbf{V}^{\mathrm{T}}\boldsymbol{\eta})\cdot\hat{\mathbf{n}}\,\mathrm{d}s - \iint_{\mathcal{X}} \boldsymbol{\eta}\cdot(\nabla\cdot\mathbf{V})\,\mathrm{d}\mathbf{x},\end{aligned} \tag{7.11}$$

where $\hat{\mathbf{n}}$ is the outward unit normal of $\mathcal{X}$. The minimizer for J is given by $\delta J(\mathbf{m},\mathbf{P},\mathbf{b})(\boldsymbol{\eta}) = \alpha\delta J_{\mathrm{I}}(\mathbf{m},\mathbf{P})(\boldsymbol{\eta}) + (1-\alpha)\delta J_{\mathrm{B}}(\mathbf{m},\mathbf{b})(\boldsymbol{\eta}) = 0$ for all $\boldsymbol{\eta}\in[C^2(\mathcal{X})]^2$, with δJ_{B} given by (5.13). It follows that for $\mathbf{m}$ to be a minimum of J, we have the necessary condition

$$\oint_{\partial\mathcal{X}} (\alpha\mathbf{V}\hat{\mathbf{n}} + (1-\alpha)(\mathbf{m}-\mathbf{b}))\cdot\boldsymbol{\eta}\,\mathrm{d}s - \alpha\iint_{\mathcal{X}} \boldsymbol{\eta}\cdot(\nabla\cdot\mathbf{V})\,\mathrm{d}\mathbf{x} = 0. \tag{7.12}$$

Applying the fundamental lemma of calculus of variations [21, p. 185] twice, once for $\eta_1 = 0$ and once for $\eta_2 = 0$, we obtain $\nabla\cdot\mathbf{V} = \mathbf{0}$ for all $\mathbf{x}\in\mathcal{X}$ and $\alpha\mathbf{V}\hat{\mathbf{n}} + (1-\alpha)(\mathbf{m}-\mathbf{b}) = 0$ for all $\mathbf{x}\in\partial\mathcal{X}$, i.e.,

$$\nabla\cdot(\mathbf{C}^{\mathrm{T}}\mathbf{C}\mathrm{D}\mathbf{m}) = \nabla\cdot(\mathbf{C}^{\mathrm{T}}\mathbf{P}), \qquad \mathbf{x}\in\mathcal{X}, \tag{7.13a}$$

$$(1-\alpha)\mathbf{m} + \alpha(\mathbf{C}^{\mathrm{T}}\mathbf{C}\mathrm{D}\mathbf{m})\hat{\mathbf{n}} = (1-\alpha)\mathbf{b} + \alpha\mathbf{C}^{\mathrm{T}}\mathbf{P}\hat{\mathbf{n}}, \qquad \mathbf{x}\in\partial\mathcal{X}. \tag{7.13b}$$

We solve this system of PDEs for $\mathbf{m}$ using a standard second-order finite volume method [68, Appendix B]. Because the system is coupled in terms of m_1 and m_2, we solve the system iteratively. For the calculation of $\mathbf{m}^{n+1}$, we first calculate m_1 where we approximate m_2 by the second component of $\mathbf{m}^n$. Next, we calculate m_2 using the newly found m_1. Subsequently, one can use the newly found m_2 and approximate m_1. This process can be repeated, however, we stop the iteration after the first calculation of m_2 and do not calculate m_1 a second time, which has experimentally been proven to be sufficient.

Lastly, we consider the minimization for u_1. To find the minimum of I we calculate the first variation of I, viz.

$$\begin{aligned}\delta I(u_1)(\eta) &= \lim_{\epsilon\to 0}\frac{I(u_1+\epsilon\eta) - I(u_1)}{\epsilon} \\ &= \lim_{\epsilon\to 0}\tfrac{1}{2}\iint_{\mathcal{X}}\Big(2(\nabla u_1 - \nabla_{\mathbf{x}} c(\cdot,\mathbf{m}))\cdot\nabla\eta + \epsilon|\nabla\eta|^2\Big)\,\mathrm{d}\mathbf{x} \\ &= \iint_{\mathcal{X}}(\nabla u_1 - \nabla_{\mathbf{x}} c(\cdot,\mathbf{m}))\cdot\nabla\eta\,\mathrm{d}\mathbf{x} \\ &= \oint_{\partial\mathcal{X}}\eta(\nabla u_1 - \nabla_{\mathbf{x}} c(\cdot,\mathbf{m}))\cdot\hat{\mathbf{n}}\,\mathrm{d}s - \iint_{\mathcal{X}}(\Delta u_1 - \nabla\cdot\nabla_{\mathbf{x}} c(\cdot,\mathbf{m}))\eta\,\mathrm{d}\mathbf{x},\end{aligned} \tag{7.14}$$

where in the last step we applied the divergence theorem. For u_1 to minimize $I(u_1)$, we have $\delta I(u_1)(\eta) = 0$ for all $\eta \in C^2(\mathcal{X})$. Applying the fundamental lemma of calculus of variations yields

$$\Delta u_1 = \nabla \cdot \nabla_{\mathbf{x}} c(\cdot, \mathbf{m}), \qquad \mathbf{x} \in \mathcal{X}, \tag{7.15a}$$

$$\nabla u_1 \cdot \hat{\mathbf{n}} = \nabla_{\mathbf{x}} c(\cdot, \mathbf{m}) \cdot \hat{\mathbf{n}}, \qquad \mathbf{x} \in \partial\mathcal{X}. \tag{7.15b}$$

We solve for u_1 using a standard second-order finite volume scheme, see for example [68, Appendix B]. The solution u_1 of (7.15) is determined up to an additive constant. We prescribe this constant by prescribing the average of u_1, see [68, § 6.1.4] for more details.

For the initial guess we choose $\mathbf{m}^0$ such that it maps $\mathcal{X}$ to the smallest bounding box of $\mathcal{Y}$. Therefore, let $\mathcal{X} = [x_{1,\min}, x_{1,\max}] \times [x_{2,\min}, x_{2,\max}]$ and let $[y_{1,\min}, y_{1,\max}] \times [y_{2,\min}, y_{2,\max}]$ be the smallest bounding box of $\mathcal{Y}$. The initial guess is given by

$$\begin{aligned} \sigma_1 &= \frac{x_1 - x_{1,\min}}{x_{1,\max} - x_{1,\min}}, \\ m_1^0(x_1, x_2) &= \sigma_1(i_2\, y_{1,\min} + i_1\, y_{1,\max}) + (1 - \sigma_1)(i_1\, y_{1,\min} + i_2\, y_{1,\max}), \end{aligned} \tag{7.16a}$$

$$\begin{aligned} \sigma_2 &= \frac{x_2 - x_{2,\min}}{x_{2,\max} - x_{2,\min}}, \\ m_2^0(x_1, x_2) &= \sigma_2(i_4\, y_{2,\min} + i_3\, y_{2,\max}) + (1 - \sigma_2)(i_3\, y_{2,\min} + i_4\, y_{2,\max}), \end{aligned} \tag{7.16b}$$

and corresponds to ∇U (up to multiplicative and additive constants) with the indices i_k and U given in Table 7.1. These choices for $\mathbf{m}$ are due to the symmetries of the elliptic and hyperbolic Monge-Ampère equation as discussed in Section 2.5. The functions U are themselves either convex, concave or saddle-shaped which, in part, gave rise to the naming convention. See [77, §5] for more details on c-convexity being a generalization of conventional convexity and for the relation with optical design see [68, §4.3]. The type of u_1 can also be characterised in terms of some derivatives of the mapping, as given in Table 7.1. This corresponds to $\mathbf{P}(\mathbf{x})$, as given by (7.3), being symmetric

	i_1	i_2	i_3	i_4	U	$\partial_{x_1} m_1$	$\partial_{x_2} m_2$
c-convex	1	0	1	0	$x_1^2 + x_2^2$	> 0	> 0
c-concave	0	1	0	1	$-x_1^2 - x_2^2$	< 0	< 0
c-saddle (type 1)	1	0	0	1	$x_1^2 - x_2^2$	> 0	< 0
c-saddle (type 2)	0	1	1	0	$-x_1^2 + x_2^2$	< 0	> 0

Table 7.1: The indices i_k and the function U corresponding to the type of solution.

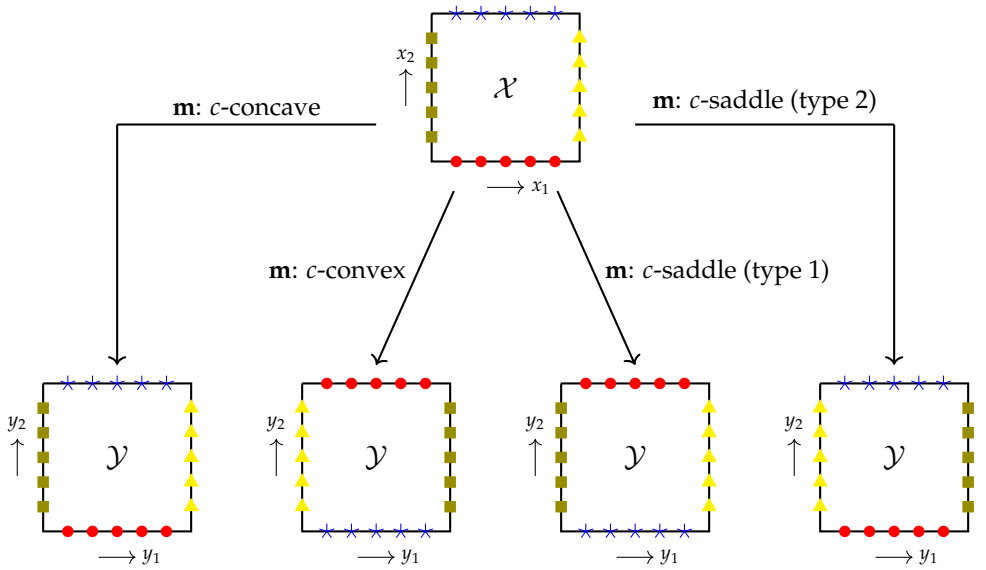


Figure 7.1: Schematic representation of the influence of the type of **m** on the boundaries of the source and target domain.

positive definite for a c-concave solution, symmetric negative definite for a c-convex solution [68, p. 113] and symmetric indefinite for c-saddle solutions. As a practical example, we consider a rectangular domain $\mathcal{X}$ and apply the optical map to $\partial\mathcal{X}$, as shown in Figure 7.1. The figure shows four different configurations depending on the type of the solution, in accordance with Table 7.1. Two implications follow. First, by the edge ray principle [67], rays adhering to the optical mapping do not cross in the case of a c-concave solution, cross in one point for the c-convex solutions, and cross in (two different lines) for the c-saddle solutions. Secondly, by fixing an orientation along $\partial\mathcal{X}$, an orientation along $\partial\mathcal{Y}$ is implied by the mapping. This orientation is the same between c-concave and c-convex solutions, and opposite for the c-saddle solutions, i.e., the orientation along $\partial\mathcal{Y}$ is flipped between elliptic and hyperbolic solutions.

7.2 Least-squares solutions

Here we will present numerical results for the four optical systems discussed in Section 6.2. For each of the optical systems considered there exist a c-convex, a c-concave, and two distinct c-saddle solutions. We will not discuss all possible solutions, but mainly focus on the c-saddle solutions, i.e., the solutions to the hyperbolic Monge-Ampère equation.

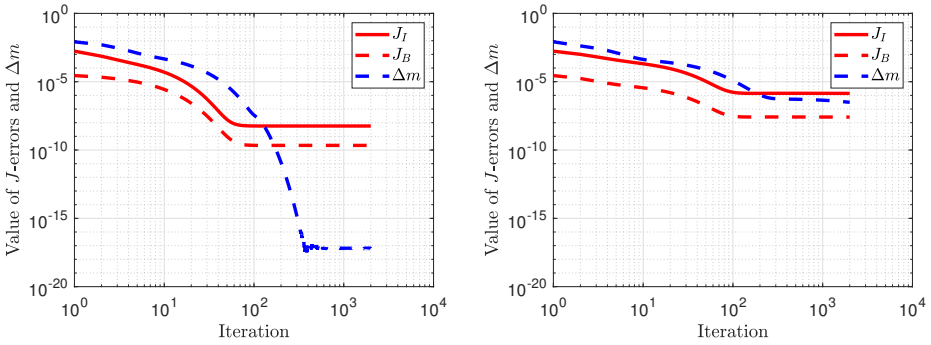


Figure 7.2: Convergence of the least-squares algorithm for the elliptic (left) and hyperbolic (right) Monge-Ampère equation.

7.2.1 Parallel-to-far-field reflector

For the first example we compare a c-concave (elliptic) and a c-saddle (type 1) (hyperbolic) solution for the parallel to far-field case. For the source and target intensity we choose the distributions

$$f(\mathbf{x}) = 1 + \cos(x_1)x_2, \qquad \mathbf{x} \in [-\tfrac{1}{2}, \tfrac{1}{2}]^2, \tag{7.17a}$$

$$g(\mathbf{y}) = \tfrac{1}{2} + y_1^2 \sin(y_2^2), \qquad \mathbf{y} \in [-\tfrac{1}{2}, \tfrac{1}{2}] \times [0, 1]. \tag{7.17b}$$

The functionals J_I, J_b and the update of $\mathbf{m}$ given by $\Delta m^n = \|\mathbf{m}^{n+1} - \mathbf{m}^n\|_2$ are shown in Figure 7.2 for the elliptic (left) and hyperbolic (right) example using a grid of 101×101 points. The figure shows that the functionals for the elliptic variant converge faster and to lower values than their hyperbolic counterparts. For the elliptic solution the convergence stagnates, in terms of J_I and J_b after 102 iterations with $J_\text{I} \approx 10^{-8}$ and $J_\text{b} \approx 10^{-10}$ and for the hyperbolic solution the functionals stagnate after 67 iterations with $J_\text{I} \approx 10^{-6}$ and $J_\text{b} \approx 10^{-8}$. Furthermore, after 2000 iterations the elliptic mapping has become stationary, due to Δm being of the order of computer precision, while this is not the case for the hyperbolic mapping. This difference in convergence behavior shows up for all optical systems presented. The calculated surfaces are shown in Figure 7.3. It is clearly observed that the sign of the curvature in the x_1-direction is opposite for the c-convex and c-saddle solution, but equal in the x_2-direction. The reflectors have been ray-traced using a quasi-Monte Carlo method using 101 bins in each direction and $5 \cdot 10^6$ rays. The obtained error in the target intensity w.r.t. to the desired intensity $g(\mathbf{y})$ is shown in Figure 7.4 and show no apparent patterns. The mean error for the c-convex and c-saddle (type 1) solutions are approximately $4 \cdot 10^{-3}$ and $5 \cdot 10^{-3}$, respectively.

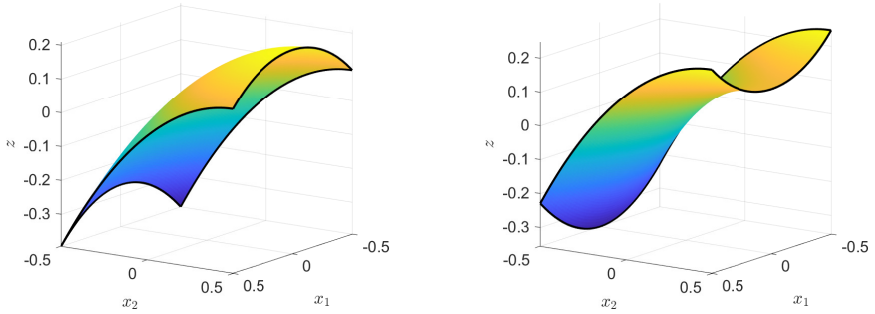


Figure 7.3: The reflector surface after 2000 iterations for c-convex (left) and c-saddle (type 1) (right) solutions.

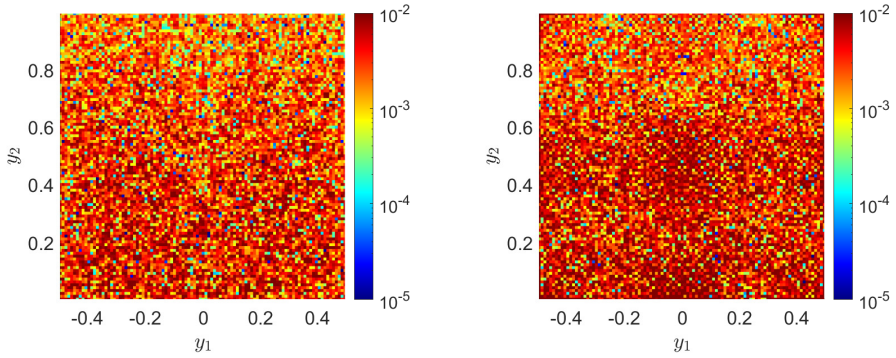


Figure 7.4: Bin-wise error in the ray-traced target distribution w.r.t. $g(\mathbf{y})$ for the elliptic (left) and hyperbolic (right) Monge-Ampère equation.

Furthermore, the root-mean squared errors (RMSs) are approximately $5 \cdot 10^{-3}$ and $7 \cdot 10^{-3}$. Using more iterations for the c-saddle (type 1) solution does not improve the mean target intensity error or RMS for this case.

Qualitative differences between the reflected rays for c-concvex, c-concave and c-saddle solutions exist. In Figures 7.5 we plot 49 rays for the c-convex and c-saddle solutions. For the c-convex solution, all rays are converging in one point, $(0, 0.5, 1)$, while for the c-saddle (type 1) solution the rays pass through the line $(x_1, 0.5, 1)$ with $x_1 \in \mathbb{R}$. This difference in behavior is typical for different solutions for each of the four optical system in this thesis. More precisely, an optical surface may have a optical axis implying refracted rays either converge or diverge relative to this axis. In our case, c-convex and c-concave surfaces have one (main) optical axis while c-saddle surfaces have two distinct optical axes such that refracted rays converge relative to one axis and diverge relative to the other.

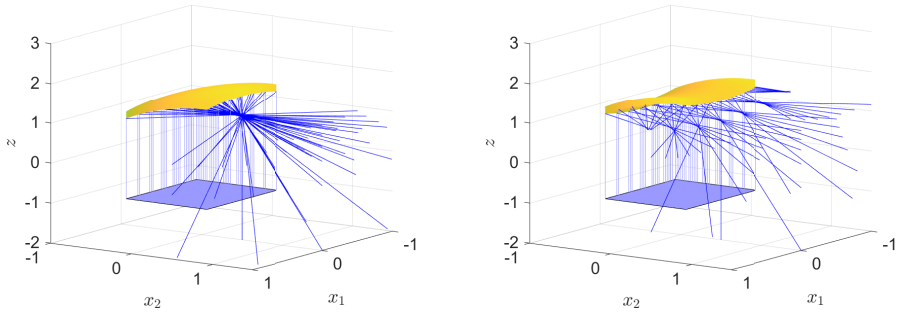


Figure 7.5: Plot of 49 rays for a c-convex (left) and c-saddle (type 1) (right) solution.

Lastly we note that in general the functionals J_{I} and J_{B} converge to lower values for the elliptic Monge-Ampère equation than for the hyperbolic variant. Furthermore, the number of iterations required for Δm to converge is in general higher for the hyperbolic Monge-Ampère equation than for the elliptic variant.

7.2.2 Parallel-to-far-field lens

For the next example we consider a lens with two surfaces. The first surface is flat and does not cause any refractions and the second surface is freeform. The target intensity is given in stereographic coordinates. We prescribe a refractive index of $n = 1.52$, and the source and target distributions read

$$f(\mathbf{x}) = 1 + x_1^2 x_2^3, \qquad \mathbf{x} \in [-\tfrac{1}{2}, \tfrac{1}{2}]^2, \tag{7.18a}$$

$$g(\mathbf{y}) = \exp(-\tfrac{y_1^2}{2} - 2y_2^2), \qquad \mathbf{y} \in [-\tfrac{1}{2}, \tfrac{1}{2}]^2. \tag{7.18b}$$

Using the least-squares algorithm with 101 grid points in each direction a numerical solution is obtained. The resulting optical system is shown on the left of Figure 7.6 for a c-saddle (type 2) pair. The figure shows that the refracted rays are diverging along the x_1-axis and converging along the x_2-axis. The error in the ray-traced target intensity is shown on the right of Figure 7.6, where $5 \cdot 10^6$ rays and 101 bins in each direction are used. Furthermore, the average error is approximately $4 \cdot 10^{-3}$ and the RMS approximately $5 \cdot 10^{-3}$. If instead we use 201 grid points in each direction for the least-squares algorithm, we obtain an average error in the ray-traced target intensity of $8 \cdot 10^{-4}$ and an RMS of 10^{-3} using the same ray-tracing parameters as before. This clearly indicates that the numerically obtained lens surface becomes more accurate

when using more discretization points.

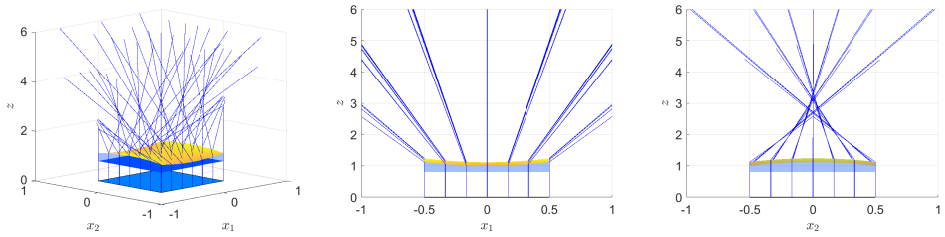


Figure 7.6: The *c*-saddle (type 2) solution with 49 rays from three different points of view.

7.2.3 Parallel-to-parallel reflectors

For the parallel-to-parallel system with two reflectors we consider the source and target distributions

$$f(\mathbf{x}) = 1 + x_1^2 + x_2, \qquad \mathbf{x} \in [-\tfrac{1}{2}, \tfrac{1}{2}]^2, \tag{7.19a}$$

$$g(\mathbf{y}) = 1 + (y_1 - \tfrac{7}{2})^2 + (y_2 - 3)^2, \qquad \mathbf{y} \in [\tfrac{5}{2}, \tfrac{7}{2}]^2. \tag{7.19b}$$

Furthermore, we choose the design parameters $\beta = 5 + \sqrt{10}$ and $L = 1$. Both the *c*-saddle (type 1 and 2) solutions have been calculated using a 101×101 grid. The mean error in the target intensity and the RMS, calculated after ray-tracing using 101 bins in each direction and $5 \cdot 10^6$ rays, are approximately $6 \cdot 10^{-3}$ and $7 \cdot 10^{-3}$ for both cases. The only real difference between the two solutions is the shape of the individual reflectors, and consequently their optical axes, as is shown in Figure 7.7. On the left we show the *c*-saddle (type 1) solution and on the right the *c*-saddle (type 2) solution with 49 rays. Again, the figure shows rays converging in one direction, and diverging in the other. Furthermore, the optical axis of convergence and divergence are swapped between the *c*-saddle (type 1) and *c*-saddle (type 2) solutions.

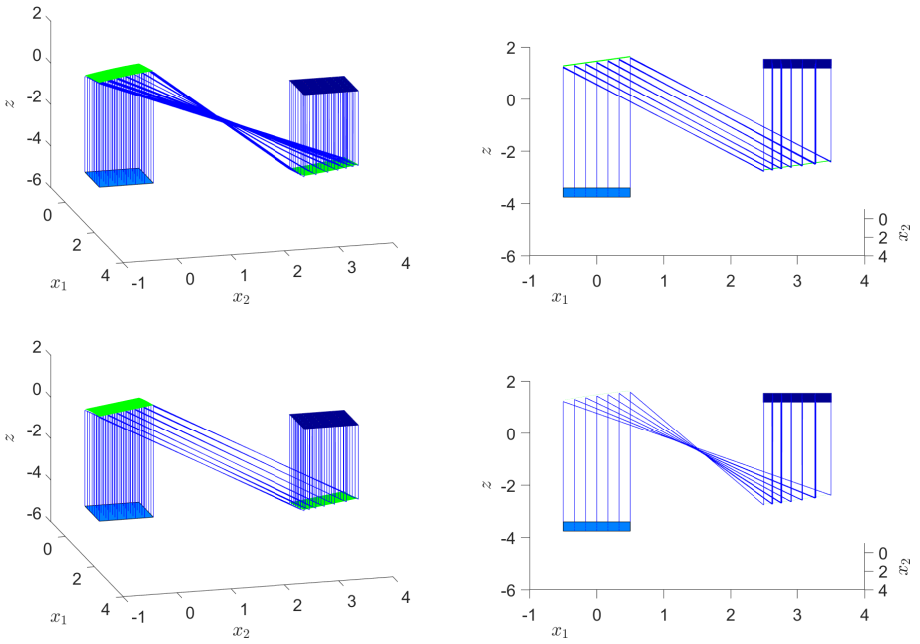


Figure 7.7: The optical systems for a c-saddle (type 1) (top) and c-saddle (type 2) (bottom) from two different points of view and 49 uniformly sampled rays.

7.2.4 Parallel-to-parallel lens

Lastly, we discuss the case for one lens with two freeform surfaces, a parallel source and parallel target. The refractive index chosen is $n = 1.52$ and the source and target distributions are given by:

$$f(\mathbf{x}) = 1 + \sqrt{x_1^2 + 2x_2^2}, \qquad \mathbf{x} \in [-\tfrac{1}{2}, \tfrac{1}{2}]^2, \tag{7.20a}$$

$$g(\mathbf{y}) = 2\frac{1 + y_1^2}{1 + y_2^2}, \qquad \mathbf{y} \in [\tfrac{5}{2}, \tfrac{7}{2}] \times [-\tfrac{1}{4}, 1]. \tag{7.20b}$$

The resulting c-saddle (type 2) optical system with design parameters $\beta = \pi$ and $L = 10$ is shown on the left of Figure 7.8, and the ray-traced target intensity on the right. Both the least-squares parameters and the ray-trace parameters are equal to those of Section 7.2.3. The resulting average error in the target intensity is approximately $6 \cdot 10^{-3}$ and the RMS $9 \cdot 10^{-3}$.

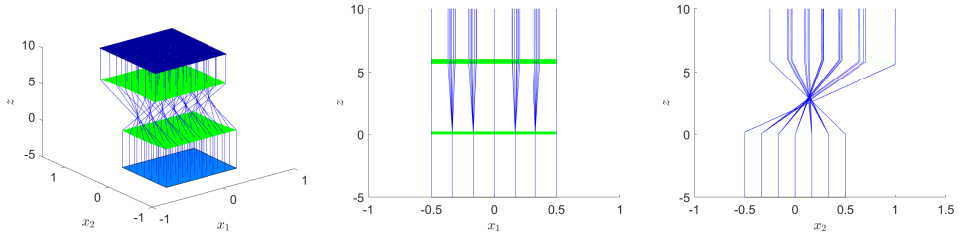


Figure 7.8: The numerically obtained *c*-saddle (type 2) solution with 49 uniform sampled rays on the left and the ray-traced target intensity on the right.

7.3 Summary

In this chapter we presented a least-squares method to calculate the optical surfaces of four optical systems. Each of the optical systems has multiple solutions, 2 elliptic (*c*-convex and *c*-concave) and 2 hyperbolic (*c*-saddle (type 1 and 2)). Various examples are given on which the numerical method has been tested. The calculated optical systems are subsequently ray-traced and error estimates are given. Differences in the ray-paths between elliptical and hyperbolic solutions are observed. Furthermore, although the least-squares algorithm works for both variants of the Monge-Ampère equation, it requires fewer iterations and converges to lower errors for the elliptic Monge-Ampère equation.

Chapter 8

Conclusions and Recommendations

8.1 Summary and conclusions

In this thesis we developed multiple methods for the hyperbolic Monge-Ampère equation. We started by introducing some general concepts regarding the Monge-Ampère equation, i.e, classification and exact solutions. These solutions are used to test the developed numerical methods.

The first algorithm is based on the method of characteristics. By classifying curves as either free or characteristic, we are able to determine the evolution of the solutions. The solution turns out to evolve exactly along the characteristics and is mathematically described by two mutually coupled systems of ordinary differential equations (ODEs). Because the systems are coupled, the characteristic (base) curve is dependent on the characteristic of the other family. This increases the difficulty of solving the systems of ODEs. By using explicit integrators, this dependency can be decoupled from the evolution along the characteristic. Each explicit integration step yields new approximations of the unknowns along a grid line. As the evolution of the direction of a characteristics is unknown along said characteristic, the dependency between characteristics need to be reconstructed. This is done by interpolating along grid lines using B-splines. Furthermore, by using a dynamic step-size, the numerical error can be controlled for both the integration and interpolation methods while also providing stability to the numerical method.

The direction of the characteristics at the boundary yields the required boundary conditions: where two characteristics enter the domain, Cauchy boundary conditions need to be prescribed, where one characteristic enters, a second-order derivative of the solution needs to prescribed and where no

characteristics enter the domain, no boundary conditions need to be prescribed. The use of an interpolation routine allows for incorporating the (varying) boundary conditions by simply adding the boundary condition as a known data point along a grid line. The forward Euler, modified Euler and Runge-Kutta methods have been demonstrated to work for various examples with the expected rate of convergence. We presented multiple test cases, each with a different number of required boundary conditions. We also presented a case where a discontinuity is propagated along a characteristic, which is in accordance to the theory. Lastly, convergence of the numerical methods for the general Monge-Ampère equation has been shown for multiple examples and multiple initial strips. In total, the method of characteristics provides a proper foundation for understanding the complex behavior of the hyperbolic Monge-Ampère equation. Furthermore, we used the method of characteristics to develop numerical methods for solving the hyperbolic Monge-Ampère equation which, as far as tested, are stable and convergent. The main downside of the method of characteristics is that the required boundary conditions are often unknown a priori.

For the design of optical systems, we require the so-called transport boundary condition. This condition requires the boundary of a source domain to be mapped to the boundary of the target domain. The method of characteristics does not easily lend itself to the transport boundary conditions. Therefore we considered a second numerical method. This method, a least-squares algorithm originally developed for the elliptic Monge-Ampère equation [66], has been adapted to work for the hyperbolic Monge-Ampère equation, requiring a few adaptations. First, we improved upon the original optimization routine for the **P**-matrix to find all possible minimizers. Second, the original method uses a projection method for approximating the boundary conditions. For the hyperbolic Monge-Ampère equation, the target domain can possibly be such that projections yield wrong results. Therefore we introduced the *segmented projection method* and the *segmented arc length method*. Both methods cut the boundary into segments and approximate boundary points per segment. For the *segmented projection method* the approximation is based on orthogonal projections. For the *segmented arc length method*, the approximation is based on distributing the boundary points by pairing the arc length of the mapping along the source boundary to the arc length of the target boundary. Both methods have been shown to outperform the projection method in both effectivity and computational efficiency. Lastly, we introduced two possible grid-shock corrections. We observed that locally putting more emphasis on the boundary error prevents grid-shocks which we forthrightly used. Using these adaptations, the least-squares solver has been demonstrated to work for a

wide range of examples. Moreover, it shows the expected rate of convergence as function of spatial discretization.

The design of optical systems with parallel light sources, far field and parallel targets in the geometrical optics regime can be posed as an inverse problem. This problem can be modeled by both the elliptic and hyperbolic Monge-Ampère equation. It was found that at least four distinct optical systems can be calculated for the same optical problem (mapping a fixed source distribution to a fixed target distribution). Two of these systems are solutions to the elliptic Monge-Ampère equation and two are solutions to the hyperbolic problem, yielding convex/concave or saddle surfaces, respectively. We presented multiple design problems and solutions and verified the obtained optical surfaces using ray tracers.

8.2 Future research

One of the key reasons for doing research is finding new and unforeseen mechanisms, consequences and opportunities. Here we present a few of the research opportunities which we have not yet explored but deem worthy of further investigation.

- Currently, our numerical methods based on the method of characteristics require a fixed step size per grid line. The factors κ^α and κ^β are coordinate dependent functions and consequently allow for skew grid lines, i.e., grid lines with $c_1 x + c_2 y = c_3$ for $c_1, c_2, c_3 \in \mathbb{R}$ and $c_1^2 + c_2^2 \neq 0$. Phrased differently, by choosing κ^α and κ^β the grid lines need not be parallel to the x- or y-axis and can possibly even be curvilinear coordinate lines. This possibly allows for efficient hyperbolic Monge-Ampère solvers on irregular domains.

- In Chapter 4 we showed a numerical example where we introduced a discontinuity in the boundary data. Due to the spline interpolation, the shock does not propagate precisely along the characteristic but rather smears out along the grid lines. This may be improved upon by introducing shock-capturing techniques and one-sided spline interpolation.

- In [68], sixteen optical design problems were formulated of which we presented four in this thesis. It seems likely that for each of these design problems (multiple) hyperbolic solutions exist, requiring further research.

- In Chapter 6 we found two elliptic and two hyperbolic solutions to the optical design problem. This is in accordance with the results obtained

in Chapter 2, i.e., for certain examples multiple symmetries exist in the Monge-Ampère equation and multiple solutions can exist when using transport boundary conditions. More specifically, in Chapter 2 we showed there can exist two elliptic solutions and an uncountable number of hyperbolic solutions, dependent on the chosen (co-)domain. Therefore, we speculate there may exist more solutions for the hyperbolic optical design problem than the two found thus far.

- The combination of convex, concave and saddle solutions allows for the construction of periodic and smooth optical elements, possibly reducing unwanted optical artefacts, like scattering or glaring, and increasing the efficacy of optical systems. As the least-squares solver has been shown to work for both the elliptic and hyperbolic Monge-Ampère equation, we conjecture the solver may be used for such periodic continuous optical surfaces.

- We have experimented, but not documented, combining a shooting-like algorithm with the method of characteristics while prescribing transport boundary conditions. To do so, we approximated the initial strip. By stepping in the x-direction, the transport boundary condition yields sufficient conditions to derive the boundary conditions required for the method of characteristics if at least one characteristic leaves the domain at all boundary points. By quantifying how well a numerical solution satisfies the transport boundary condition an optimization problem can be formulated. By iteratively updating the initial strip, this method seemed to work for most hyperbolic examples tried. A problem arises when no characteristics leave the domain at least one boundary point, because then the transport boundary condition no longer yields sufficient boundary conditions for mthe method of characteristics. So the question is: Does the transport boundary condition guarantee at least one characteristic leaving the domain for each boundary point? Or phrased more broadly, how are the characteristics and the transport boundary condition related?

- For the elliptic Monge-Ampère equation various uniqueness and existence results are established in the literature. These include, among others, results for classical, Aleksandrov and viscosity solutions and for equivalent formulations in optimal transport theory. Few such uniqueness and existence results exist for the hyperbolic Monge-Ampère equation and even fewer when combined with transport boundary conditions. Our research has experimentally shown that solutions exist and therefore further theoretical results are needed.

- Both the method of characteristics and the least-squares method have their own strengths and weaknesses. Wether these methods can be combined somehow requires further research. One can, for example, consider the evolution of the mapping along the characteristics to update the mapping locally within one iteration of the least-squares algorithm. Vice versa, the result of the least-squares method can serve as the initial strip and accompanying boundary conditions for the method of characteristics.

- In Section 7.1 we mentioned that the orientation of the target boundary, found by applying the optical mapping to the source boundary, differs between elliptic and hyperbolic solutions to the Monge-Ampère equation. This has been observed for all tried elliptic and hyperbolic examples, and does not seem to be confined to optical systems. Thus far no proof has been found for this, but examples are easily constructed by contracting a mapping with a reflection in one axis, e.g., if $\mathbf{m}$ solves an elliptic Monge-Ampère equation, then $\mathbf{m} \circ \mathbf{l}$ with $\mathbf{l}(x_1, x_2) = (-x_1, x_2)^{\mathrm{T}}$ solves an hyperbolic Monge-Ampère equation.

Appendix A

Interpolation for the numerical MOC

In this section we will briefly introduce splines, which we use for numerical interpolation. To understand spline interpolation, we first introduce knot sequences, which generate B-splines, and in term determine the spline interpolant. To this end consider a set of $N+1$ numbers $\boldsymbol{\xi} = \{\xi_0, \xi_1, \ldots, \xi_N\}$ with $\xi_0 \leq \xi_1 \leq \cdots \leq \xi_N$. Such a sequence is called a knot sequence and each member of the sequence is called a knot. B-splines b_k^n of degree n for the knot sequence $\boldsymbol{\xi}$ are recursively defined on the interval $[\xi_k, \xi_{k+n+1})$ by [41, p. 52]

$$b_k^n(t) = \gamma_k^n(t) b_k^{n-1}(t) + \Big(1 - \gamma_{k+1}^n(t)\Big) b_{k+1}^{n-1}(t), \qquad \gamma_k^n(t) = \frac{t - \xi_k}{\xi_{k+n} - \xi_k}, \quad \text{(A.1)}$$

for $0 \leq k, k+n+1 \leq N$, $0 \leq n \leq k$ with initial values

$$b_k^0(t) = \begin{cases} 1, & \text{if } \xi_k \leq t < \xi_{k+1}, \\ 0, & \text{otherwise.} \end{cases} \quad \text{(A.2)}$$

Each B-spline b_k^n is a polynomial, of degree $\leq n$ on its knot interval $[\xi_k, \xi_{k+n+1})$, and vanishes outside this interval.

Let $m \in \mathbb{N}_+$, let g be a sufficiently smooth function and let $t_0 \leq \cdots \leq t_{m-1}$ be a set of points such that $g(t_0), \ldots, g(t_{m-1})$ are known. Furthermore, let $N = m + n$, so $\boldsymbol{\xi} = \{\xi_0, \xi_1, \ldots, \xi_{m+n}\}$ and let $\boldsymbol{\xi}$ be such that

$$\xi_k < t_k < \xi_{k+n+1} \text{ for } k = 0, \ldots, m-1. \quad \text{(A.3)}$$

These conditions, also known as the Schoenberg-Whitney conditions [41, p.

91], imply that there exists a unique interpolating spline

$$P(t) = \sum_{k=0}^{m-1} c_k b_k^n(t), \tag{A.4}$$

of degree n, which interpolates g in the interval $t \in [t_0, t_{m-1})$. The coefficients c_k are calculated via the implicit relation

$$\mathbf{Ac} = \mathbf{g}, \qquad (a_{j,k}) = b_k^n(t_j), \qquad \mathbf{c} = (c_k), \qquad \mathbf{g} = (g(t_k)), \tag{A.5}$$

for $j, k = 0, 1, \ldots, m-1$. Furthermore, the associated error can be estimated by

$$|g(t) - P(t)| \leq C\Big(n, \|\mathbf{A}^{-1}\|_\infty\Big)\, \|g^{(n+1)}\|_{\infty,R}\, h^{n+1}, \qquad t \in R, \tag{A.6}$$

where $\|g^{(n+1)}\|_{\infty,R}$ is the maximum norm of the derivative of g of order $n+1$ on the interval $R = [\xi_n, \xi_{n+m}]$, $h = (\xi_{j+1} - \xi_j)$ and the constant C depends on the degree n of the B-splines and the infinity-norm of the inverse of $\mathbf{A}$. A convergence order for odd and even n has been established, where generally convergence for odd n is of order $\mathcal{O}(h^{n+1})$ as given by (A.6). Convergence for even n has been observed to be of order $\mathcal{O}(h^{n+2})$ instead of the theoretical established upper bound (A.6). This observed superior convergence for even n is not fully understood at the time of writing [78].

What remains is to construct a suitable knot sequence such that the Schoenberg-Whitney condition holds. Let data points $t_0 \leq \cdots \leq t_{m-1}$ be such that $m > n$, i.e., let the number of data points exceeds the degree of the B-splines used for interpolation. Then we choose the knot sequence $\boldsymbol{\xi}$ according to

$$\begin{cases} \xi_i = t_0 & i = 0, \ldots, n-1, \\ \xi_i = \frac{1}{n-1} \sum_{j=1}^{n-1} t_{i-n+j}, & i = n, \ldots, m, \\ \xi_i = t_m & i = m+1, \ldots, m+n-1. \end{cases} \tag{A.7}$$

The first and last terms of $\boldsymbol{\xi}$ are the original, possibly duplicated, starting and end values t_0 and t_m, respectively. The remaining components are running averages of size $n-1$ of $\{t_j\}$, which ensures (A.3) holds.

Example: Consider the ordered sequence of data points $t : \{1, 2, 4, 5, 7, 10\}$ and let the desired spline order be $n = 3$. Hence $m = 6$ and the knot sequence $\boldsymbol{\xi}$ according to (A.7) is given by $\xi : \{1, 1, 1, 3, 4.5, 6, 10, 10, 10\}$.

In the case that we require extrapolation at a point $t_e < t_0$ or $t_e > t_{m-1}$, we simply estimate $g(t_e) \approx P(t_e)$. This need for extrapolation does occur

for both the modified Euler and Runge-Kutta based methods, as shown in Figure A.1 for the modified Euler scheme, due to the predictor lying outside of Ω and missing either the value a, or b, as it cannot be determined along the characteristic. Let without loss of generality $a > b$, then in order to approximate $b(x_{i+\frac{1}{2}}, \tilde{y}^{\beta}_{i+\frac{1}{2}}(1))$ which is needed to calculated $\tilde{\mathbf{v}}^{\beta}_{i+1}(1)$ according to (4.7), we extrapolate using a spline based on $\tilde{\mathbf{v}}^{\alpha}_{i+1/2}(1)$ for $j = 1, \ldots, m+1$, known at the y-values $\tilde{y}^{\alpha}_{i+\frac{1}{2}}(1)$ inside the domain, to $\tilde{y}^{\beta}_{i+\frac{1}{2}}(1)$ outside the domain.

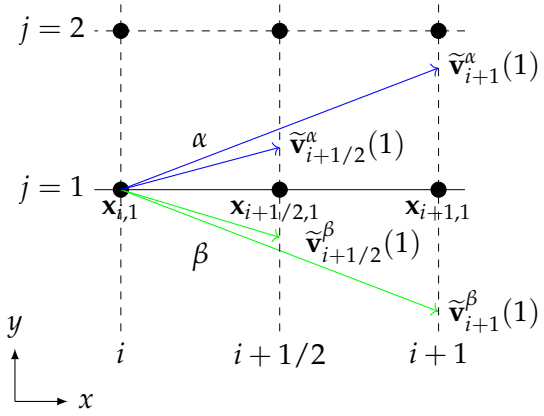


Figure A.1: Modified Euler based method near the lower boundary of the domain.

Bibliography

[1] S. Abdallah. Numerical solutions for the pressure Poisson equation with Neumann boundary conditions using a non-staggered grid, I. *J. Comput. Phys.*, 70(1):182–192, May 1987.

[2] A. Adams and C. Essex. *Calculus: a complete course (7th Ed.)*. Pearson, Toronto, 2009.

[3] M. J. H. Anthonissen, L. B. Romijn, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Unified mathematical framework for a class of fundamental freeform optical systems. *Opt. Express*, 29(20):31650–31664, Sep 2021.

[4] R. Beltman, J. ten Thije Boonkkamp, and W. IJzerman. A least–squares method for the inverse reflector problem in arbitrary orthogonal coordinates. *Journal of Computational Physics*, 367:347–373, Aug. 2018.

[5] J. 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.

[6] J. 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.

[7] J.-D. Benamou, F. Collino, and J.-M. Mirebeau. Monotone and consistent discretization of the Monge-Ampère operator. *Mathematics of Computation*, 85(302):2743–2775, Mar. 2016.

[8] J.-D. Benamou and V. Duval. Minimal convex extensions and finite difference discretisation of the quadratic Monge–Kantorovich problem. *European Journal of Applied Mathematics*, 30(6):1041–1078, Sept. 2018.

[9] M. W. M. C. Bertens, E. M. T. Vugts, M. J. H. Anthonissen, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman. Numerical methods for the hyper-

bolic Monge–Ampère equation based on the method of characteristics. *Partial Differential Equations and Applications*, 3(4):52, Jul 2022.

[10] G. Bonnet and J.-M. Mirebeau. Monotone discretization of the Monge–Ampère equation of optimal transport. *ESAIM: Mathematical Modelling and Numerical Analysis*, 56(3):815–865, Apr. 2022.

[11] M. Born and E. Wolf. *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Cambridge University Press, 7th expanded edition, 1999.

[12] F. Brickell and B. S. Westcott. Reflector design for two-variable beam shaping in the hyperbolic case. *Journal of Physics A: Mathematical and General*, 9(1):113–128, January 1976.

[13] C. J. Budd and J. F. Williams. Moving mesh generation using the parabolic Monge–Ampère equation. *SIAM Journal on Scientific Computing*, 31(5):3438–3465, 2009.

[14] A. Caboussat. Computation of portfolio hedging strategies using a reduced Monge-Ampère equation. In *Proceedings of the 20th International Conference on Computing in Economics and Finance, Oslo, June 22-24, 2014*, pages 1–13, 2014.

[15] 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.

[16] 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, Optimisation and Calculus of Variations*, 19(3):780–810, June 2013.

[17] 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.

[18] T. Chen. Smooth local solutions to degenerate hyperbolic Monge-Ampère equations. *Annals of PDE*, 5(1):1–47, January 2019.

[19] J. Cordova and T. Barth. Grid generation for general 2-D regions using hyperbolic equations. In *26th Aerospace Sciences Meeting*. American Institute of Aeronautics and Astronautics, Jan. 1988.

[20] R. Courant and D. Hilbert. *Methods of Mathematical Physics: Partial Differential Equations*. John Wiley & Sons, Ltd, Singapore, April 1989.

[21] R. Courant and D. Hilbert. *Methods of Mathematical Physics, volume 1*. John Wiley & Sons, Ltd, 1989.

[22] 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 Mathematique*, 339(12):887–892, Dec. 2004.

[23] 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.

[24] L. C. Evans. *Partial differential equations*, volume 19. American Mathematical Soc., 2010.

[25] J. J. Fahie. *A history of electric telegraphy, to the year 1837*. London: E. & FN Spon, 1884.

[26] X. Feng, R. Glowinski, and M. Neilan. Recent developments in numerical methods for fully nonlinear second order partial differential equations. *SIAM Review*, 55(2):205–267, Jan. 2013.

[27] X. Feng and M. Neilan. Vanishing moment method and moment solutions for fully nonlinear second order partial differential equations. *Journal of Scientific Computing*, 38(1):74–98, July 2008.

[28] X. Feng and M. Neilan. Mixed finite element methods for the fully nonlinear Monge–Ampère equation based on the vanishing moment method. *SIAM Journal on Numerical Analysis*, 47(2):1226–1250, Jan. 2009.

[29] 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.

[30] A. Figalli. *The Monge–Ampère Equation and Its Applications*. Zurich Lectures in Advanced Mathematics. European Mathematical Society, Zürich, Switzerland, Jan. 2017.

[31] B. D. Froese. A numerical method for the elliptic Monge–Ampère equation with transport boundary conditions. *SIAM Journal on Scientific Computing*, 34(3):A1432–A1459, Jan. 2012.

[32] 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.

[33] 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.

[34] 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.

[35] W. Gautschi. *Numerical Analysis*. Birkhäuser Boston, Boston, 2012.

[36] W. R. Hamilton. Theory of systems of rays. *The Transactions of the Royal Irish Academy*, pages 69–174, 1828.

[37] 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.

[38] 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.

[39] E. Hecht. *Optics*. Pearson, 5th edition, 2017.

[40] W. Heisenberg. Nonlinear problems in physics. *Physics Today*, 20(5):27–33, May 1967.

[41] K. Höllig and J. Hörner. *Approximation and Modeling with B-Splines*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013.

[42] J. Hong. The global smooth solutions of Cauchy problems for hyperbolic equation of Monge-Ampère type. *Nonlinear Analysis: Theory, Methods & Applications*, 24(12):1649–1663, June 1995.

[43] T. M. Howard. *Hyperbolic Monge-Ampère Equations*. PhD thesis, University of North Texas, 2006.

[44] IEA. (2022), Lighting, IEA, Paris `https://www.iea.org/reports/lighting`, License: CC BY 4.0. Accessed: 06-03-2023.

[45] F. James. *Michael Faraday: A Very Short Introduction*. Oxford University Press, 2010.

[46] D. Kahaner, C. Moler, G. Forsythe, S. Nash, S. Nash, and M. Malcolm. *Numerical Methods and Software*. Prentice-Hall series in Computational Mathematics. Prentice-Hall, Englewood cliffs, New Jersey, 1988.

[47] 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.

[48] M. N. Koleva and L. G. Vulkov. Numerical solution of the Monge-Ampère equation with an application to fluid dynamics. *AIP Conference Proceedings*, 2048(1):030002, 2018.

[49] A. Kushner, V. V. Lychagin, and J. Slovák. *Lectures on Geometry of Monge–Ampère Equations with Maple*, pages 53–94. Springer International Publishing, Cham, 2019.

[50] O. Lakkis and T. Pryer. A finite element method for second order nonvariational elliptic problems. *SIAM J. Sci. Comput.*, 33(2):786–801, 2011.

[51] O. Lakkis and T. Pryer. A finite element method for nonlinear elliptic problems. *SIAM J. Sci. Comput.*, 35(4):A2025–A2045, 2013.

[52] L. D. Landau and E. M. Lifshitz. *Course of Theoretical Physics: Vol. 1, Mechanics*. Course of Theoretical Physics. Butterworth-Heinemann, 3 edition, 1976.

[53] S. J. Leon. *Linear Algebra with Applications (9th Ed.)*. Pearson, Toronto, 2015.

[54] G. Loeper and F. Rapetti. Numerical solution of the Monge–Ampère equation by a Newton's algorithm. *Comptes Rendus Mathematique*, 340(4):319–324, Feb. 2005.

[55] J. D. Logan. *An introduction to nonlinear partial differential equations*. Pure and applied mathematics. Wiley-Interscience, 2nd ed edition, 2008.

[56] R. K. Luneburg. *Mathematical theory of optics*. University of California Press, Berkeley, CA, July 2021.

[57] J. C. Maxwell. VIII. A dynamical theory of the electromagnetic field. *Philosophical Transactions of the Royal Society of London*, 155:459–512, Dec. 1865.

[58] M. Neilan, A. J. Salgado, and W. Zhang. Numerical analysis of strongly nonlinear PDEs. *Acta Numerica*, 26:137–303, May 2017.

[59] 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.

[60] 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.

[61] V. I. Oliker. On reconstructing a reflecting surface from the scattering data in the geometric optics approximation. *Inverse Probl.*, 5(1):51–65, 1989.

[62] V. I. Oliker and L. D. Prussner. On the numerical solution of the equation $\frac{\partial^2 z}{\partial x^2}\frac{\partial^2 z}{\partial y^2} - \left(\frac{\partial^2 z}{\partial x \partial y}\right) = f$ and its discretizations, I. *Numerische Mathematik*, 54(3):271–293, May 1989.

[63] Philips. CLASSICTONE 40W E27 220-240V A55 CL 1CT/10X10F | 920053843329 | Philips lighting, `https://www.lighting.philips.com/main/prof/lamps/incandescent-lamps/standard-t-a-e-shape/classictone-standard/920053843329_EU/product`. Accessed: 31-03-2023.

[64] Philips. Philips, MAS LEDBulbND7.3-100WE27830 A70 FRG UE | 929003480202 | Philips lighting, `https://www.lighting.philips.com/main/prof/led-lamps-and-tubes/led-bulbs/master-ultraefficient-led-bulb/929003480202_EU/product`. Accessed: 31-03-2023.

[65] C. R. Prins. *Inverse Methods for Illumination Optics*. PhD thesis, Eindhoven University of Technology, 2014.

[66] 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 Journal on Scientific Computing*, 37(6):B937–B961, jan 2015.

[67] H. Ries and A. Rabl. Edge-ray principle of nonimaging optics. *Journal of the Optical Society of America A*, 11(10):2627, Oct. 1994.

[68] L. B. Romijn. *Generated Jacobian Equations in Freeform Optical Design: Mathematical Theory and Numerics*. PhD thesis, Eindhoven University of Technology, 2021.

[69] 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 Journal on Scientific Computing*, 43(2):B298–B322, January 2021.

[70] A. Stroud and D. Secrest. *Gaussian Quadrature Formulas: By A.H. Stroud and Don Secrest*. Prentice-Hall series in Automatic Computation. Prentice-Hall, Englewood cliffs, New Jersey, 1966.

[71] J.-P. Tignol. *Galois' theory of algebraic equations*. World Scientific Publishing, Singapore, Singapore, Apr. 2001.

[72] G. Tilmann and V. I. Oliker. Optical design of two-reflector systems, the Monge-Kantorovich mass transfer problem and Fermat's principle. *Indiana University Mathematics Journal*, 53(5):1255–1277, 2004.

[73] K. Toraichi, K. Katagishi, I. Sekita, and R. Mori. Computational complexity of spline interpolation. *International Journal of Systems Science*, 18(5):945–954, 1987.

[74] J. L. Troutman. *Variational Calculus and Optimal Control: Optimization with Elementary Convexity*. Undergraduate Texts in Mathematics. Springer, 2nd edition, 1995.

[75] D. V. Tunitskii. On the global solubility of the Monge–Ampère hyperbolic equations. *Izvestiya: Mathematics*, 61(5):1069–1111, oct 1997.

[76] N. K. Tuy. Numerical solution of hyperbolic equations and systems with two independent variables by a method of the Runge–Kutta type. 1. Technical report, 1968.

[77] C. Villani. *Optimal Transport: Old and New*. Number 338 in Grundlehren der Mathematischen Wissenschaften. Springer, Berlin, 2009.

[78] Y. S. Volkov. Study of the convergence of interpolation processes with splines of even degree. *Siberian Mathematical Journal*, 60(6):973–983, November 2019.

[79] B. S. Westcott and F. Brickell. Computation of reflector surfaces for two-variable beam shaping in the hyperbolic case. *Journal of Physics A: Mathematical and General*, 9(4):611–625, April 1976.

[80] W. Wong. Regular hyperbolicity, dominant energy condition and causality for Lagrangian theory of maps. *Classical and Quantum Gravity*, 28, November 2010.

[81] 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, Sep 2013.

[82] N. K. Yadav. *Monge–Ampère problems with non-quadratic cost function: application to freeform optics*. PhD thesis, Eindhoven University of Technology, Sept. 2018.

[83] 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. *Journal of Scientific Computing*, 80(1):475–499, 2019.

[84] Y. Zhang, J. Gao, J. Peng, and W. Han. A robust method of computing finite difference coefficients based on Vandermonde matrix. *Journal of Applied Geophysics*, 152:110–117, May 2018.

[85] 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.

Summary

Numerical Methods for the Hyperbolic Monge-Ampère Equation with Applications to Optical Design

Climate change, carbon emissions and the need for more sustainable technologies are familiar topics nowadays. One possible venue for reducing our carbon footprint is to improve the optical/lighting systems we use. In 2021 approximately 875.000 kg of CO_2 was globally emitted due to lighting applications [44]. Although this is 6.4% less than in 2010, the need for more sustainable optical systems remains. Since the introduction of light emitting diodes (LEDs) the luminous efficacy has already increased from 10 lumen per watt for the traditional incandescent light bulb to over 200 lumen per watt for LED bulbs. A big contributor for this is the fact that LEDs produce relatively little heat. This has in term led to the need to redesign many optical systems.

The design of optical systems can be modeled by a mathematical equation called the Monge-Ampère equation. The Monge-Ampère equation is a nonlinear partial differential equation (PDE), which is a class of extremely hard problems for which no complete mathematical theory exists yet. To intuitively understand this, consider one practical application of these PDEs; the forecasting of the weather. People often experience the weather forecasts to be wrong, especially when forecasting multiple days or even weeks ahead. It turns out these forecasts are indeed often objectively wrong, the reason being that the underlying mathematics is complicated and the physics is chaotic!

The design of optical systems, modeled by the Monge-Ampère equation, shares this complexity. Furthermore, the equation comes in two variants, an elliptic and a hyperbolic variant. The optical surfaces one can calculate using the elliptic variant are convex or concave while the hyperbolic surfaces resemble a saddle. Mathematically the elliptic surfaces are easier to calculate than the hyperbolic ones. In this thesis, we developed two algorithms to calculate the hyperbolic surfaces. The first algorithm uses the method of characteristics and the second uses a least-squares approach. We've shown that the hyperbolic surfaces are different from the elliptic ones. Now, combining elliptic and hyperbolic surfaces enables the design of new smooth optical elements which are more energy efficient as they can better steer the light.

Curriculum Vitae

Maikel Bertens was born on July 20th, 1992 in 's-Hertogenbosch, the Netherlands. After finishing his high school degree in 2011 at the d'Oultremont college in Drunen, he began studying applied mathematics at the Technische Universiteit Eindhoven. After a change of heart, he went on to study applied physics and eventually graduated with a bachelor degree in both applied physics and applied mathematics. Next he obtained a master's degree in applied mathematics with a focus on computational methods which he, a.o., applied as an intern at Philips Lighting (Signify).

In March 2018 Maikel started as a technology consultant at Alten NL. His work focused on the simulation of LIDAR sensors for autonomous vehicles, relying on his knowledge of numerical mathematics, physics and computer science. At this point, Maikel believed he could have obtained his PhD if he wanted to, so instead of solely believing this, he made obtaining a PhD his new goal. He then contacted Assoc. Prof. Jan ten Thije Boonkkamp, his former advisor for his master thesis, to obtain a reference and was promptly made aware of an open PhD position on the hyperbolic Monge-Ampère equation at Prof Wilbert IJzerman's and ten Thije Boonkkamp's group Computational Illumination Optics. A research position at the Centre for Analysis, Scientific Computing and Applications (CASA) within the Technische Universiteit Eindhoven followed and the results can be read in this thesis.

List of Publications

Journal articles

3. M. W. M. C. Bertens, M. J. H. Anthonissen, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman, Design of optical systems with freeform convex, concave or saddle surfaces. In preparation.

2. M. W. M. C. Bertens, M. J. H. Anthonissen, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman, An iterative least-squares method for the hyperbolic Monge-Ampère equation with transport boundary conditions, SIAM Journal on Scientific Computing. Submitted.

1. M. W. M. C. Bertens, E. M. T. Vugts, M. J. H. Anthonissen, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman, Numerical methods for the hyperbolic Monge-Ampère equation based on the method of characteristics, Partial Differential Equations and Applications 3, 52 (2022), https://doi.org/10.1007/s42985-022-00181-4.

Conference contributions

1. M. W. M. C. Bertens, M. J. H. Anthonissen, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman, Design of optical surfaces conform the hyperbolic Monge-Ampère equation, European Physical Journal Web of Conferences, Vol. 266. 2022.

Oral presentations at scientific conferences

5. M. W. M. C. Bertens, M. J. H. Anthonissen, W. L. IJzerman and J. H. M. ten Thije Boonkkamp, Numerical Methods for the Hyperbolic General Monge-Ampère Equation based on the Method of Characteristics. SIAM Conference on Computational Science and Engineering (CSE23), Amsterdam, Netherlands, 26 February - 3 March, 2023.

4. M. W. M. C. Bertens, M. J. H. Anthonissen, W. L. IJzerman and J. H. M. ten Thije Boonkkamp, Design of optical surfaces conform the hyperbolic Monge-Ampère equation. In TOM 2 - Computational, Adaptive and Freeform Optics, European Optical Society Biennial Meeting (EOSAM), Porto, Portugal, 12-16 September, 2022.

3. M. W. M. C. Bertens, M. J. H. Anthonissen, W. L. IJzerman and J. H. M. ten Thije Boonkkamp, Numerical Methods for the Hyperbolic Monge-Ampère Equation based on the Method of Characteristics. SciCade 2022, International Conference on Scientific Computation and Differential Equations, Reykjavík, Iceland, 25–29 July, 2022.

2. M. W. M. C. Bertens, M. J. H. Anthonissen, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman, A Least-Squares Solver for the Hyperbolic Monge-Ampère Equation. Freeform Scattering Optics program meeting, 1-2 November, 2021.

1. M. W. M. C. Bertens, M. J. H. Anthonissen, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman, Numerical Methods for the Hyperbolic Monge-Ampère Equation Based on the Method of Characteristics, International Conference on Numerical Analysis and Partial Differential Equations ICNAPDE, Online conference, 16-17 September, 2021.

Acknowledgments

Working towards a PhD knows many ups and downs and I am happy that I may have lived this experience at the TU/e.

I am extremely grateful for my supervisors Wilbert, Jan and Martijn for all the discussions and jokes we have shared. Specifically I am thankful for Wilbert's practical and insightful ideas, for Jan's scrutinizing and critical thinking and for Martijn's attention to professional and personal details.

Thanks go out to Enna and Diane, for their administrative support and Diane's help in organizing the CASA-days.

I would like to thank my CASA colleagues with whom I have shared many lunches, laughs and discussions. It has been enriching to meet so many genuine people with such varying, though vast, amounts of knowledge.

To my office mates, thank you for listening to my rants about failing software, computers, bureaucracies and all other nuances.

I am grateful for my peers in the Computational Illumination Optics group for helping me understand and generate new scientific ideas.

My sincere thanks goes to my family, Danique's family and the friends who took interest in my research even when they did not understand a word about it.

For making everything better, for both listening and pushing back, for the serious and the joyful conversations, for the love and support, for you just being you, thank you Danique.

The only true wisdom is in knowing you know nothing.

Socrates