Solving the Monge-Ampère Equation for a free-form reflector in arbitrary coordinate systems

Beltman, R.

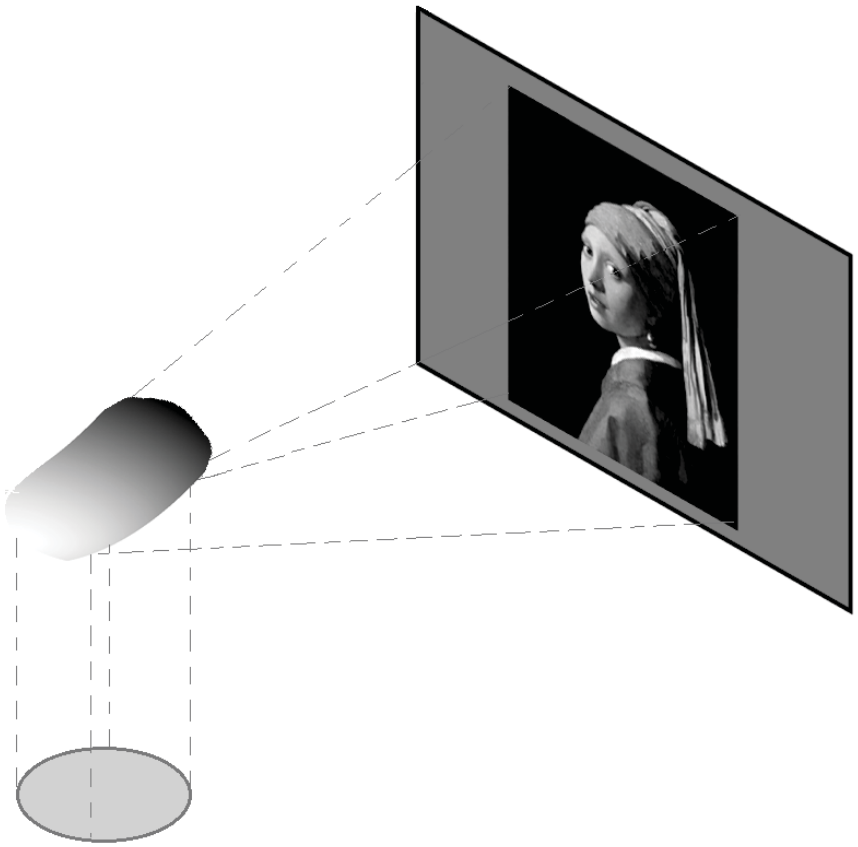
Award date:
2015

Technische Universiteit Eindhoven University of Technology

Where innovation starts

Master's Thesis

Solving the Monge-Ampère Equation for a Free-Form Reflector in Arbitrary Coordinate Systems



Author:
René Beltman

Supervisors:
Jan ten Thije Boonkkamp (CASA)
Wilbert IJzerman (Philips Lighting)

August 31, 2016

With this master's thesis I completed the track *Computational Science and Engineering* of the master's programma *Industrial and Applied Mathematics* at the *University of Technology, Eindhoven*. During my graduation project I worked at *Philips Lighting* and was supervised by Wilbert IJzerman (Philips Lighting) and Jan ten Thije Boonkkamp (CASA). I am very grateful for their support.

Contents

Chapter 1

Introduction

This thesis will be concerned with illumination optics, i.e. the design of optical systems for lighting. In the field of illumination optics there are essentially two approaches. The most common one of the two approaches determines the light output of a given optical system with light source. Methods of this kind are called *forward methods*. The most familiar of such methods is *ray-tracing*. In ray-tracing the optical system is modeled on a computer and light-rays are emitted randomly from the simulated light source in correspondence with the given light output of the light source. These light rays are then traced through the optical system and in this way the light output of the system is determined. Ray-tracing is a valuable tool for the optical designer, because it allows to check if the designed optical system has the desired output. In practice, the optical designer will be adapting his optical system until it gives a light output close to the demanded output. This way of designing optical system by ray-tracing has some disadvantages. Ray-tracing is very computational intensive if high precision is needed and therefore quite slow. Moreover, this way of optical designing relies very much on the creativity and skill of the optical designer.

Another technique contrasting with the forward methods are the so-called *inverse methods*. An inverse method determines an optical system which transforms a given intensity output of a light source to a desired light output distribution. So, when a light source with a specific intensity output and a desired light output distribution are prescribed, an inverse method will determine the shape of the refractor or reflector which converts the light output of the source in the desired output distribution. Inverse methods have the advantage over the forward methods that they are much less labour-intensive and demanding of the optical designer. Recent developments in diamond turning techniques have resulted in the fact that arbitrarily shaped lenses and reflectors can be made with much higher precision than before. This allows for inverse methods to come up with optical systems that would never have been achieved with forward methods, but can be physically realized nonetheless. This thesis will be concerned with an inverse method and the goal of this research is to physically produce a reflector that transforms a uniform parallel bundle of light into a highly nontrivial light output distribution to demonstrate the capabilities of this inverse method. To make this more concrete we will now describe the simple optical system that we will focus on for the rest of this thesis.

1.1 The optical system and the milling machine

In Figure 1.1 the optical system, that will be of interest for the rest of this thesis, is depicted. The system consists of a light source, a reflector surface and a projection screen. We will assume that the light source is circular and emits a parallel homogenous bundle of light in the direction perpendicular to its plane. Directly above this light source a reflecting surface is positioned which reflects the bundle of light in the direction of a projection screen. We will assume that this projection screen is situated in the far-field of the reflector surface. This means that we will assume that the reflected light rays originate from one point. As long as the distance between the

reflector and the projection screen is large enough, the error introduced by it will be small. The problem that interests us is the following. Suppose that the light source with its output intensity is given and, moreover, that a desired light intensity distribution on the projection screen is given, what then should the shape of the reflector surface be?

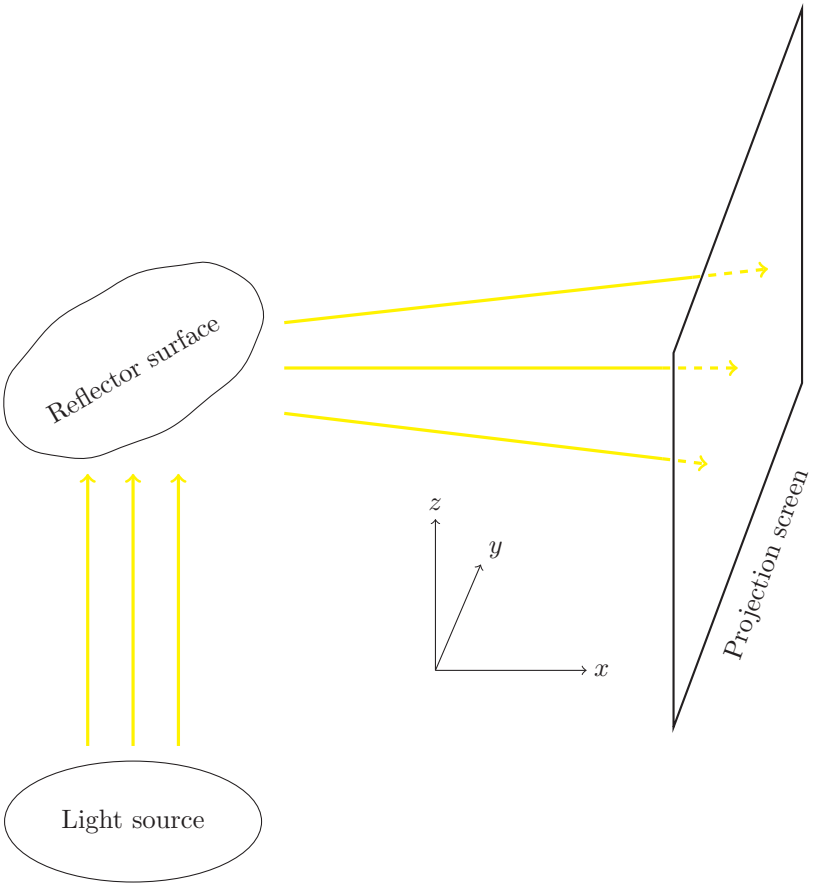


Figure 1.1: Schematic representation of the optical system

In Chapter 3 we will see that the reflector surface is described by a complicated partial differential equation involving the output intensity of the source and the desired intensity distribution on the projection screen. In [5] a numerical method was introduced to solve this partial differential equation. This numerical method we will from now on call the least-squares method. In the least-squares method the light source gets covered with a Cartesian grid and in each of the grid points the height of the reflector gets determined. The least-squares method is able to determine in this way the reflector surface corresponding to highly nontrivial light intensity distributions. For example the light intensity distribution corresponding to a painting was determined with the least-squares method. The reflector surface subsequently gets tested by using professional ray-tracing software and this gave the satisfactory result depicted in Fig 1.2.

The goal of this graduation project is to set the necessary steps in order to be able to physically produce the reflector that transforms the homogeneous parallel bundle of light into an outgoing intensity distribution that produces Figure 1.2 on the projection screen. To produce this reflector we will use a milling machine available at *Philips Lighting, Eindhoven*.

Crudely said, the milling machine constructs a reflector by removing material from the top surface of a cylindrical workpiece. In order to do this the chisel of the machine moves along concentric circles around the axis of the milling machine and at each point cuts to a specified depth. It starts with the circles with smallest radii and moves outward in a more or less smooth manner. The depth at each point corresponds to the height of the reflector. The machine needs

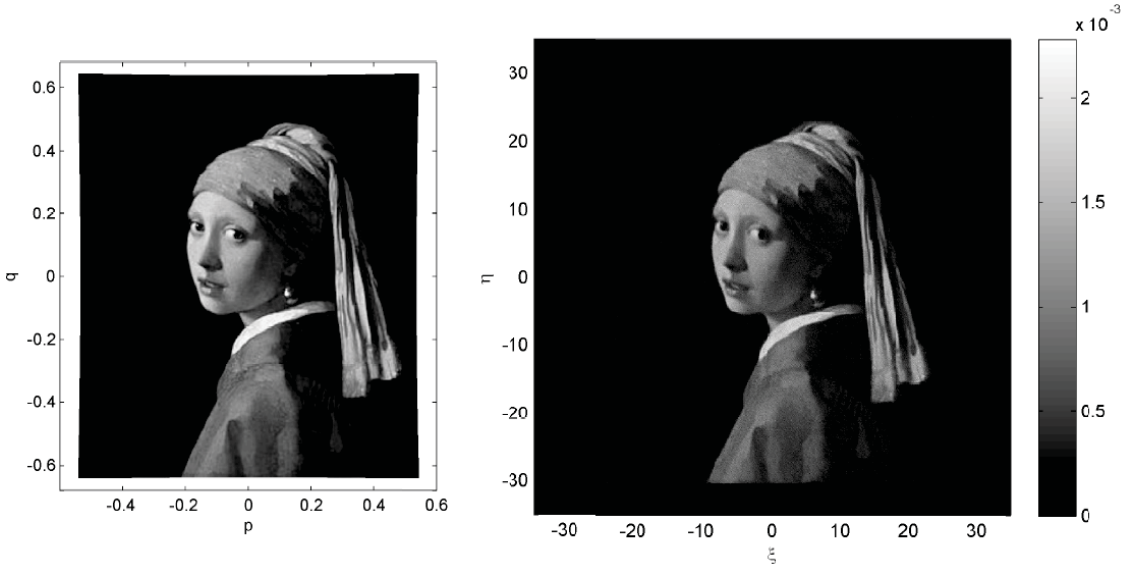


Figure 1.2: On the left the desired target distribution on the projection screen is shown. On the right the result is shown when using professional ray-trace software to determine the output distribution on the projection screen given by the reflector surface as determined by the least-squares method.

to be provided with the reflector height on a polar coordinate grid normal to the chisel axis. One thing that puts a constraint on the set of workable data is the fact that when the axis of the chisel moves along its path, the acceleration of the chisel in the direction of cutting is finite. To deduce the implications of this restriction on the reflector surfaces it produces we consider a circular path $\gamma : [t_0, t_1] \to \mathbb{R}^2$ in the plane normal to the axis of the milling machine. Let $v : \mathcal{M} \to \mathbb{R}$ be the function giving the height of the reflector for each position of the chisel in the set of possible positions, i.e. the set $\mathcal{M}$. Along the path described by γ the height of the reflector is given by $v(\gamma(t))$, $t \in [t_0, t_1]$. Using polar coordinates we have

$$\begin{aligned}\frac{\mathrm{d}^2 v(\gamma(t))}{\mathrm{d}t^2} &= \frac{\mathrm{d}}{\mathrm{d}t}\Big(\frac{\partial v}{\partial r}\frac{\mathrm{d}r}{\mathrm{d}t} + \frac{\partial v}{\partial \theta}\frac{\mathrm{d}\theta}{\mathrm{d}t}\Big) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\Big(\frac{\partial v}{\partial \theta}\frac{\mathrm{d}\theta}{\mathrm{d}t}\Big) \\ &= \frac{\partial v}{\partial \theta}\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + \frac{\partial^2 v}{\partial \theta^2}\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right) \\ &= \frac{\partial^2 v}{\partial \theta^2}\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right).\end{aligned}$$

Here we used that derivative of r with respect to time is zero, because the path is circular, and, the fact that second derivative of θ with respect to time is zero, because the chisel moves at constant speed along its path. From this we see that the acceleration of the chisel is proportional to the second partial derivative of v with respect to θ. Thus the fact that the chisel acceleration is restricted implies that the second derivative of the height of the reflector surface with respect to θ is restricted also.

In the least-squares method the light source is covered with a Cartesian grid and in the grid points the height of the reflector is determined. To provide the milling machine with workable data we must provide the machine with the reflector height on a polar coordinate grid. In Figure 1.1 it can be seen that the angle between the direction of the incoming light ray and the reflector surface will be approximately 45°. In order to reduce the chisel accelerations we need to reduce

$\partial^2 v/\partial\theta^2$. We can reduce $\partial^2 v/\partial\theta^2$ to a great extent to choose a cylindrical coordinate system such that the axis of the cylinder is approximately normal to the reflector surface. This coordinate system is depicted in purple in Figure 1.3.

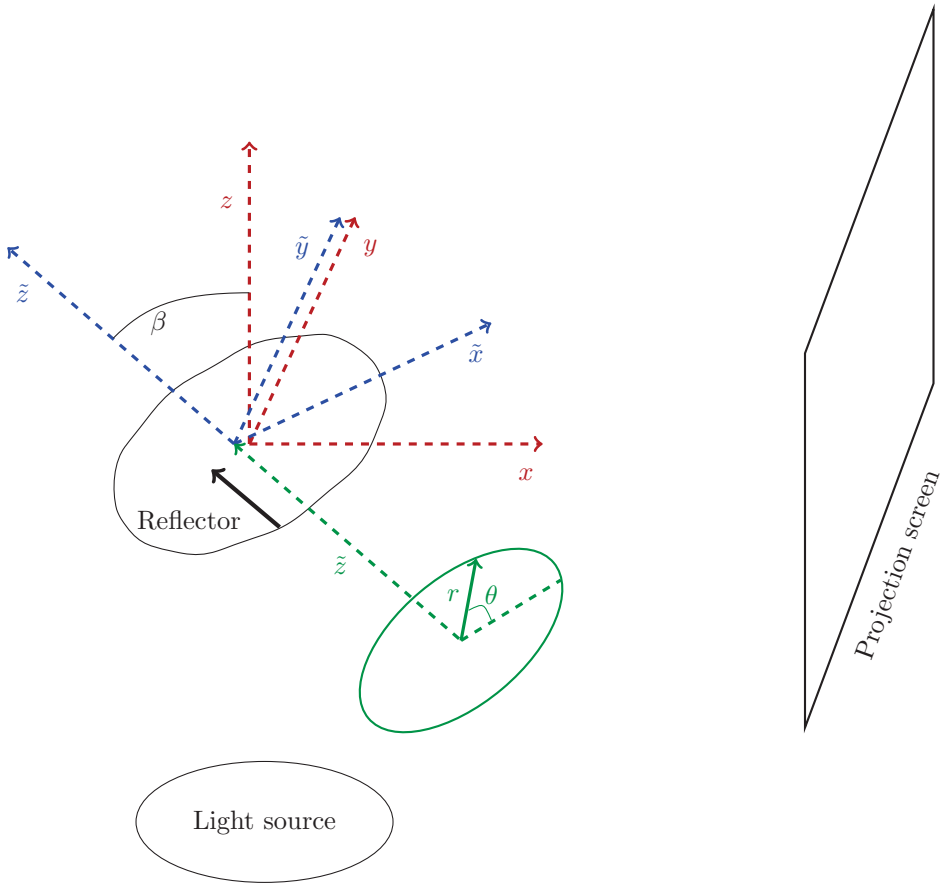


Figure 1.3: The Cartesian coordinate system is depicted in red, the rotated Cartesian coordinate system is depicted in blue and the cylindrical coordinate system is depicted in green. The coordinate system in red is the coordinate system in which the reflector heights are specified by the least-squares method. The black arrow represents the chisel of the milling machine, which is parallel to the $\tilde{z}$-axis. The origin of the three coordinate systems coincide on the reflector. This not immediately clear from the figure, because we plotted the cylindrical coordinate system with a origin beneath the reflector for clarity of the figure.

The least-squares method determines the reflector height in the Cartesian coordinate system with the $x-y$ plane parallel to the light source, which is depicted in red in Figure 1.3. To transform a point (x, y, z) to the corresponding point $(\tilde{x}, \tilde{y}, \tilde{z})$ in the rotated Cartesian coordinate system, which is depicted in blue in Figure 1.3, we rotate over an angle β. We have

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \cos(\beta) & 0 & -\sin(\beta) \\ 0 & 1 & 0 \\ \sin(\beta) & 0 & \cos(\beta) \end{pmatrix} \begin{pmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \end{pmatrix} = \begin{pmatrix} \tilde{x}\cos(\beta) - \tilde{z}\sin(\beta) \\ \tilde{y} \\ \tilde{x}\sin(\beta) + \tilde{z}\cos(\beta) \end{pmatrix}.$$

We can furthermore relate the point (x, y, z) in the original Cartesian coordinate system to a point in the cylindrical coordinate system of the milling machine, by defining the cylindrical coordinate system by the relations

$$\tilde{x} = r\cos(\theta) \quad \text{and} \quad \tilde{y} = r\sin(\theta).$$

With this we find that the relation between the Cartesian coordinates x, y, z and the cylindrical coordinates $r, \theta, \tilde{z}$ is given by

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} r\cos(\theta)\cos(\beta) - \tilde{z}\sin(\beta) \\ r\sin(\theta) \\ r\cos(\theta)\sin(\beta) + \tilde{z}\cos(\beta) \end{pmatrix}. \tag{1.1}$$

Now, suppose the reflector height is determined by the Least-Squares method in the coordinate system of the source, i.e. the coordinate system depicted in red in Figure 1.3. Let the reflector height be given by a function $u : \mathcal{D}_{R_1} \to \mathbb{R}$. We assume that the light source is a disk with radius $R_1 > 0$, i.e. the set $\mathcal{D}_{R_1}$, and that the reflector height is given by the function u for every point on the light source. We want to determine from this function $u : \mathcal{D}_{R_1} \to \mathbb{R}$ a function $v : \mathcal{M} \to \mathbb{R}$ that gives the reflector heights in the rotated cylindrical coordinate system. The set $\mathcal{M}$ depends on the function u and the angle of rotation β. The reflector is in the coordinate system of the light source given by points $(x, y, u(x, y)) \in \mathbb{R}^3$. Equation (1.1) tells us how x, y, z are related to the cylindrical coordinates $r, \theta, \tilde{z}$. From this we see that the coordinates $r, \theta, \tilde{z}$ are a point on the reflector surface if and only if they satisfy the relation

$$u\big(r\cos(\theta)\cos(\beta) - \tilde{z}\sin(\beta), r\sin(\theta)\big) = r\cos(\theta)\sin(\beta) + \tilde{z}\cos(\beta).$$

Using this we define $v : \mathcal{M} \to \mathbb{R}$ as: For each $(r, \theta) \in \mathcal{M}$, $v(r, \theta) := \tilde{z}$, where $\tilde{z}$ is the root of the function $f_{r,\theta}$, which is defined by

$$f_{r,\theta}(\tilde{z}) := u\big(r\cos(\theta)\cos(\beta) - \tilde{z}\sin(\beta), r\sin(\theta)\big) - \big(r\cos(\theta)\sin(\beta) + \tilde{z}\cos(\beta)\big). \tag{1.2}$$

Furthermore, we define the set $\mathcal{M}$ in this to be the set of points $(r, \theta) \in \mathbb{R}_{>0} \times [0, 2\pi)$ for which $f_{r,\theta}$ has a root. We will choose the angle of rotation β in such a way that the maximum value of $|\partial^2 v/\partial\theta^2|$ is as small as possible.

The function $v : \mathcal{D}_{R_1} \to \mathbb{R}$ is only given on the grid points used in the Least-Squares method. This means that we need to interpolate the function when searching for the roots of the functions $f_{r,\theta}$. To provide the milling machine with workable data we cover the set $\mathcal{M}$ with a polar coordinate grid, given by (r_i, θ_j), $1 \le i \le N_r$, $1 \le j \le N_\theta$. To find the height of the reflector surface for a grid point (r_i, θ_j) we determine the root $\tilde{z}_{i,j}$ of f_{r_i,θ_j} and have

$$v_{i,j} := v(r_i, \theta_j) = \tilde{z}_{i,j}.$$

The set $\mathcal{D}_{R_1}$ is by definition a disk, however, the set $\mathcal{M}$ is not disk-shaped anymore. As the reflector makes approximately an angle of 45° with the plane of the light source, β will be close to $\pi/4$. The set $\mathcal{M}$ will therefore be roughly shaped as an ellipse whose semi-major axis, by the Pythagorean theorem, has roughly $\sqrt{2}$ times the length of its semi-minor axis. However, the milling machine only produces disk-shaped reflectors and therefore the reflector needs to be extrapolated to a disk $\mathcal{D}_R$ containing the set $\mathcal{M}$.

In Figure 1.4 a possible shape of the set $\mathcal{M}$ and a disk $\mathcal{D}_R$ containing it are sketched. Besides this, in this figure, also some examples of chisel paths are shown. The function $v : \mathcal{M} \to \mathbb{R}$ needs to be extrapolated to the whole of $\mathcal{D}_R$ such that at the boundary between $\mathcal{M}$ and $\mathcal{D}_R \backslash \mathcal{M}$ the second derivative $|\partial^2 v/\partial\theta^2|$ does not get too large. As part of an internship project preceding this graduation project different ways of extrapolating the function v to $\mathcal{D}_R$ where considered. In the most fruitful of these attempts we extrapolated the reflector surface by minimizing $\int_{\mathcal{D}_R\backslash\mathcal{M}} (\Delta v)^2 \, \mathrm{d}A$, while demanding continuous differentiability of v over the boundary $\partial\mathcal{M}$. We minimized $\int_{\mathcal{D}_R\backslash\mathcal{M}} (\Delta v)^2 \, \mathrm{d}A$, because it is a functional treatable by the Calculus of Variations and $(\Delta v)^2$ is related to $(\partial^2 v/\partial\theta^2)^2$. The application of the Calculus of Variations resulted in a boundary value problem with two boundary conditions on $\partial\mathcal{M}$ and none on $\partial\mathcal{D}_R$. We were able to solve this boundary value problem and in this way extrapolate the reflector surface. However, there was an important issue that this approach did not take into consideration. Theoretically no light leaving the the light source as a parallel bundle should end up at the part $\mathcal{D}_R\backslash\mathcal{M}$ of the

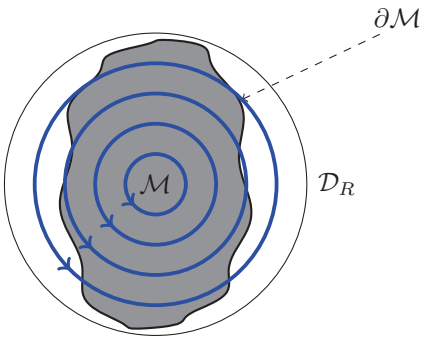


Figure 1.4: The disk $\mathcal{D}_R$ with four examples of chisel paths.

extrapolated reflector, but in practice, due to small alignment errors, this is unavoidable. This was not taken into account for when extrapolating the reflector surface in the way just described. Light that falls on these extrapolated parts of the reflector will therefore be reflected in unwanted directions and in this way ruin the image on the projection screen created by the $\mathcal{M}$ part of the reflector.

In this thesis we will therefore take a different approach. We will extrapolate the reflector surface already in the coordinate system of the source. We will extend the function $v : \mathcal{D}_{R_1} \to \mathbb{R}$ to a disk $\mathcal{D}_{R_2}$, with $R_2 > R_1$ and take R_2 large enough such that when we determine the reflector height function in the rotated cylindrical coordinate system, the support of the function contains a disk $\mathcal{D}_R$ that contains $\mathcal{M}$. We will first use the least-squares method to determine the function $v : \mathcal{D}_{R_1} \to \mathbb{R}$ and subsequently use an adaptation of the least-squares method to extrapolate this function to the set $\mathcal{D}_{R_2}$. By using an adaptation to the least-squares method we will be able to prescribe in which direction the extrapolated parts of the reflector should reflect light. Moreover, this extrapolated function $v : \mathcal{D}_{R_2} \to \mathbb{R}$ will also be continuously differentiable over the boundary $\partial\mathcal{D}_{R_1}$.

In order to be able adapt the least-squares method such that we can use it to extrapolate the reflector surface, we first also need to improve it for disk-shaped light sources. As Figure 1.2 shows, the least-squares method works quite satisfactory for rectangular sources. However, when a disk-shaped light source is used, the results are less ideal. This can be seen in Figure 1.5. The image has some strange features along its edges. In the middle of all four edges strange bulges appear and especially the lower corners of the image appear truncated. These abnormalities are the result of the fact that a Cartesian grid is not very suitable to a disk-shaped source. Although these abnormalities appear relatively small they will turn out to be detrimental when we need to extrapolate our reflector to the larger disk $\mathcal{D}_{R_2}$.

In order to deal with arbitrarily shaped light sources we will in this thesis introduce the least-squares method independent of the choice of coordinate system. Besides this we will implement a polar coordinate version of the least-squares method for the disk-shaped light source, that will outperform the least-squares method in Cartesian coordinates in this case.

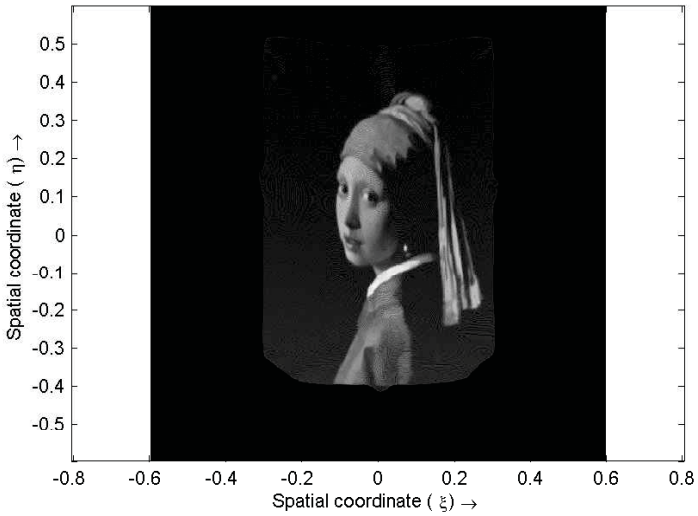


Figure 1.5: The image is determined by calculating the reflection of approximately 1.1 million evenly distributed light rays emitted by the source. The reflector surface was calculated with the Least-Squares method on a 800×800 Cartesian grid for a disk-shaped source.

1.2 Outline of this thesis

This thesis will start out with introducing the necessary concepts of the theory of Tensor Calculus in Chapter 2. We will in this chapter define tensors and show how the components of a tensor transform under a basis transformation. Besides this we will clarify the difference between so-called holonomic and anholonomic bases and see how they are related. Furthermore, we will consider the properties of the the directional derivative in Euclidean space. We will see that the directional derivative is a special case of the more general concept of covariant derivative. We will end this chapter with the definition of the covariant derivative, which is a convenient definition to rely on when deriving the energy conservation equation for our reflector system.

The derivation of this energy conservation equation, which is called the Monge-Ampère equation, is what will occupy us for most of Chapter 3. We will derive the Monge-Ampère equation in a coordinate independent manner, i.e. beforehand we will not make any assumption on the coordinate system in use besides the assumption that it is orthogonal. The derivation of the Monge-Ampère equation culminates in Theorem 3.3.4 in which we state the Monge-Ampère equation for the reflector system in coordinate independent form. Subsequently, we will derive in Section 3.4 some coordinate specific expressions for Monge-Ampère equation. We will derive from the coordinate independent form of the equation, the coordinate specific ones for polar coordinates with a holonomic basis, polar coordinates with an anholonomic basis and Cartesian coordinates. In this last coordinate system we retrieve the form of the Monge-Ampère equation as earlier derived in [5]. We will end Chapter 3 by deriving a boundary value problem for the Monge-Ampère equation when the output intensity of the light source and the desired light intensity distribution on the projection screen are prescribed. This will be the subject of Section 3.5 and Section 3.6.

In Chapter 4 we will generalize the Least-Squares method introduced in [5] to general coordinate systems. Each iteration of the Least-Squares method consists of three steps. These three steps will be treated in Sections 4.2, 4.3 and 4.4 The method as presented in [5] contains a minor flaw in the second of these three steps. In this step a constraint was forgotten when minimizing a certain functional. We will therefore in Section 4.3 consider this step quite extensively and show that with this extra constraint we can still minimize this functional algebraically. We will increase our intuition on this minimization problem by using a very indicative way to graphically represent the minimization problem.

In Chapter 5 we will discuss the implementation of the Least-Squares method for polar coordinates with an anholonomic basis. We will compare the Least-Squares method in polar coordinates with the one in Cartesian coordinates, as presented in [5], for a disk-shaped light source. We will see that for a disk-shaped light source the Least-Squares method in polar coordinates outperforms the Least-Squares method in Cartesian coordinates. Furthermore, we will see that the unwanted bulges appearing on the edges of the image on the projection screen, as shown in Figure 1.5, disappear when using the polar coordinate implementation of the Least-Squares method.

Finally, we will in Chapter 6 consider the extension of the reflector surface from an initial disk-shaped source $\mathcal{D}_{R_1}$ to a second larger disk-shaped source $\mathcal{D}_{R_2}$, with $R_2 > R_1$. In order to this we present a adapted version of the least-squares method and test it for several desired intensity distributions on the projection screen. Furthermore, we will study the discontinuities over the boundary between the $\mathcal{D}_{R_1}$ and $\mathcal{D}_{R_2}$.

We will end this thesis by summarizing the achievements and making some recommendations for further research by considering what final things need to be done to produce the reflector and thereby achieve the goal of this graduation project.

Chapter 2

Tensor Calculus

In the mathematical description of a physical system one often starts out with defining a coordinate system. This coordinate system is used to quantitatively describe the features of the system and allows for the application of the powerful tools of *Calculus*. However, the results obtained by these tools should not depend on the particular choice of coordinate system, because the coordinate system is not a feature of the physical system. The choice for a specific coordinate system is frequently based on the symmetries of the physical system. For example, a physical system which is rotationally symmetric with respect to a certain axis is easily described in terms of cylindrical coordinates. It could, however, be that one wants a mathematical description of such a degree of generality that the symmetries of this system are not known beforehand. In such cases one would like to postpone the choice of a specific coordinate system.

When one wants to avoid the choice of a specific coordinate system in describing a physical system, the natural mathematical language to use is that of *Tensor Calculus*. Tensor Calculus provides a way of applying the tools of Calculus without specifying a coordinate system and in this way ensures that the results obtained are indeed independent of coordinate systems as demanded. From the coordinate-free expressions derived, the coordinate specific ones easily follow.

In this chapter we will introduce all the necessary concepts of Tensor Calculus that will be used later on. Instead of referring to literature when using concepts of Tensor Calculus we decided to devote a chapter to it. We have several reasons for this. Firstly, it causes this thesis to be self-contained to a larger extent what makes for more pleasant and convenient reading. Furthermore, it might serve as an introduction to the concepts of Tensor Calculus and *Differential Geometry* for someone with a specific interest in the application of Tensor Calculus to *free form reflector design*. Part of the literature on free form reflector design relies heavily on differential geometry, see for example [9, 10]. This chapter should serve as a minimal introduction to Tensor Calculus and Differential Geometry such that these papers are understandable.

This chapter relies heavily on [1]. Besides this, also [2, 4, 3, 7, 8] have been consulted. If one wants a more detailed understanding of what is to follow, these texts should be studied. We will indicate throughout this chapter which parts of these sources have been used.

2.1 Euclidean spaces and manifolds

The optical system we would like to describe exists in a 3-dimensional space. The directions of light rays in this space will be indicated by vectors, hence our space needs a vector space structure. Besides this we need a concept of inclination between two vectors and in order to establish this the vector space will be furnished with an inner product. However, we do not want to give a point in this 3-dimensional space the special status of origin. Such a choice is arbitrary and we therefore avoid it. The mathematical concept which precisely encapsulates the aforementioned is the *Euclidean space*, which is defined as follows for an arbitrary dimension. (See [1, p.45].)

Definition 2.1.1. An n-dimensional Euclidean space E is a metric space, i.e. a set equipped with

a distance function $d : E \times E \to \mathbb{R}_{>0}$, furnished with a mapping $+ : E \times V \to E$, in which V is an n-dimensional Euclidean vector space* with inner product $(\cdot|\cdot) : V \times V \to \mathbb{R}$, such that

(i) $\forall x, y \in E\ \exists! v \in V :\quad y = x + v,\ d(x, y) = \sqrt{(v|v)}$;

(ii) $\forall x, y \in E,\ \forall v \in V :\quad d(x + v, y + v) = d(x, y)$;

(iii) $\forall x \in E,\ \forall u, v \in V :\quad (x + u) + v = x + (u + v)$.

The Euclidean space of dimension 3 is the minimal description of the space in which our optical system is situated. Nonetheless, it is immediately clear that the reflector surface due to its curvature does not have the linearity expressed by the vector space V and therefore is not a Euclidean space. The generalization of the concept of Euclidean space, which lends itself to the description of curved spaces, is the *manifold*. (See for example [3, p.7].)

Definition 2.1.2. An n-dimensional *topological manifold* is a second countable Hausdorff† topological space, say M, such that every point $p \in M$ is contained in some open set U_p that is homeomorphic to an open subset of the n-dimensional Euclidean space.

The trivial example of a manifold is the Euclidean space, because the Euclidean space is second countable and Hausdorff and obviously homeomorphic to itself. The manifolds of interest in this thesis besides Euclidean spaces are *surfaces* in $\mathbb{R}^3$. The surface is an example of a *submanifold of* $\mathbb{R}^3$. The following definition is from [4, p.4].

Definition 2.1.3. A subset $M \subset \mathbb{R}^N$ is said to be an n-dimensional submanifold of $\mathbb{R}^N$, with $n \leq N$‡, if locally M can be described by giving $N - n$ of the coordinates continuously in terms of the n remaining ones. This means that given $p \in M$, a neighborhood of p on M can be described in some coordinate system $(x^1, \ldots, x^n, y^1, \ldots, y^{N-n})$ of $\mathbb{R}^N$ by $N - n$ continuous functions

$$y^\alpha = y^\alpha(x^1, \ldots, x^n), \qquad \alpha = 1, \ldots, N - n.$$

If the functions y^α are k-times ($k \geq 1$) continuously differentiable, we will call M an n-dimensional C^k-submanifold of $\mathbb{R}^N$. If the functions y^α are k-times differentiable for every $k \in \mathbb{N}$, we call the submanifold *smooth*. An $(N - 1)$-dimensional submanifold of $\mathbb{R}^N$ we call a *surface* and a 1-dimensional submanifold we call a *curve*.

It is clear that submanifolds of $\mathbb{R}^N$ are examples of topological manifolds, because as subsets of the Euclidean space $\mathbb{R}^N$ they are certainly locally homeomorphic to $\mathbb{R}^N$. Moreover, because the Euclidean space satisfies the necessary topological conditions, a submanifold of the Euclidean space does also. The reflector surface which will be of much interest to us is an example of a surface. We can represent the reflector surface in the following way.

Example 2.1.4. Let $f : V \subset \mathbb{R}^2 \to \mathbb{R}$ be a continuously differentiable function. Then the set of points $M = \{(x, f(x)) \in \mathbb{R}^3 \mid x \in V\}$ describes a surface in $\mathbb{R}^3$.

A neighbourhood on a submanifold of $\mathbb{R}^N$ can be described by a coordinate system. (The following definition is from [1, p.46].)

Definition 2.1.5. Let M be an n-dimensional submanifold of $\mathbb{R}^N$ and $U \subset M$ an open subset. The couple (U, v), where v is a system of n functions $v : U \to v(U) \subset \mathbb{R}^n$, is called a *(local) (curvilinear) coordinate system* if it is a differentiable bijection, with non-degenerate Jacobian. The function $v : U \to v(U)$ is referred to as a *(coordinate) chart* of U and the inverse function $v^{-1} : v(U) \to U$ is referred to as a *parametrization* of $v(U)$.

*One often writes $\mathbb{R}^n$ for both E and V.

†A topological space X, with collection of open sets τ, is said to be second countable if it has a countable base, i.e. there exists a countable collection of open sets $\{U_i\}_{i \in I}$ such that every $U \in \tau$ is a union of elements of this collection. The topological space (X, τ) is said to be Hausdorff if for each distinct pair of points $x, y \in X$ there exists disjoint $U, V \in \tau$ such that $x \in U$ and $y \in V$. However, for all the manifolds used in this text, these topological notions will be satisfied.

‡From now on, when n and N are used in such a context, one may assume that $n \leq N$.

In second place behind the Cartesian coordinate system, the most used coordinates in 2-dimensional Euclidean space are probably the polar coordinates.

Example 2.1.6. Let $U = \mathbb{R}^2 \backslash \{(\lambda, 0) \mid \lambda \geq 0\}$. Following notational standards we set $v = (v^1, v^2) = (r, \theta)$ and $u = (u^1, u^2) = (x, y)$. The coordinate chart $u \mapsto v(u)$ of the polar coordinate system is given by

$$r(x,y) = \sqrt{x^2 + y^2},$$
$$\theta(x,y) = \tan^{-1}(x,y) := \begin{cases} \arctan(y/x) & (x > 0, y > 0), \\ \arctan(y/x) + 2\pi & (x > 0, y < 0), \\ \arctan(y/x) + \pi & (x < 0), \\ \pi/2 & (x = 0, y > 0), \\ 3\pi/2 & (x = 0, y < 0). \end{cases}$$

The parametrization $v \mapsto u(v)$ is the inverse of the coordinate chart and is given by

$$x(r,\theta) = r\cos(\theta),$$
$$y(r,\theta) = r\sin(\theta).$$

An example of a familiar parametrization for surfaces in $\mathbb{R}^3$ is the following.

Example 2.1.7. Consider the surface of Example 2.1.4. Let $p = (x, f(x))$ and let $u : V \to \mathbb{R}^2 : x \mapsto u(x)$ define some local coordinate system on V, with local coordinates u^1 and u^2. We define a coordinate system on M by the mapping $v : M \to \mathbb{R}^2$, given by

$$p = (x, f(x)) \mapsto v(p) = u(x). \tag{2.1}$$

The inverse of this mapping $v^{-1} : \mathbb{R}^2 \to M$, given by

$$y \mapsto (u^{-1}(y), f(u^{-1}(y))), \tag{2.2}$$

is called the *Monge parametrization* of the surface M.

In Figure 2.1 two examples of a Monge parametrization are shown. The Monge parametrization makes use of the fact that there is a one-to-one mapping between the subset V of the plane and the surface M. This gives a one-to-one mapping between a point on M and the coordinate pair (u^1, u^2) on V. Note, however, that such a bijection, and hence such a parametrization, is only possible if the surface has no "overhangs". From Figures 2.1 it is intuitively clear that such a parametrization is not possible for example for all the points of the sphere $\mathcal{S}^2$ at once. The reflector surface with which we will be concerned turns out to have no such overhangs and the Monge parametrization is therefore suitable to describe it. This is a restriction of the possible surfaces we can describe, but a justifiable one. The set V will represent our light source and we will assume it to radiate only in the direction perpendicular to V in the direction of the reflector surface. A light ray radiating from the point $x \in V$ will hit upon the reflector surface at the point $(x, f(x)) \in M$, because the light rays travel in straight lines. Thus, the Monge parametrization is a very natural way to describe the reflector surface if we take V to be the light source, and it does not pose a new restriction for our problem.

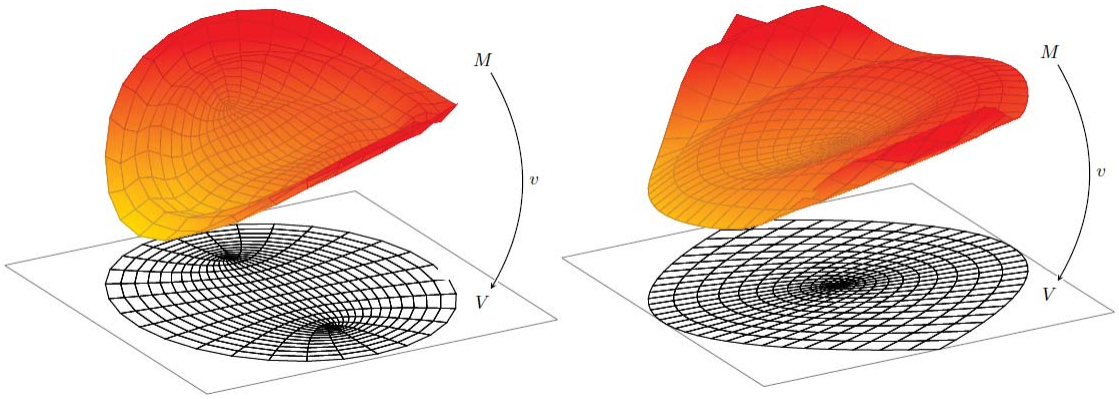


Figure 2.1: Two surfaces with a Monge parametrization. The coordinate system on the plane induces a coordinate system on the curved surface. On the left we have an elliptic coordinate system and on the right side a parabolic coordinate system. The resulting coordinate systems on the surfaces are indicated by coordinate lines.

2.2 Tangent space

In this section we will introduce the concept of a *tangent space*. The tangent space is the vector space of *tangent vectors* attached to each point of an n-dimensional submanifold of $\mathbb{R}^N$. Before introducing the tangent space we will first introduce tangent vectors and the difference between a holonomic basis and an anholonomic basis.

A tangent vector to an n-dimensional submanifold M of $\mathbb{R}^N$ can be considered as the velocity vector of some differentiable curve on M. Let us consider a point $p \in M$ and suppose we have a curve passing through this point. We can interpret the velocity vector of this curve in the point p as a vector in $\mathbb{R}^N$ originating from the point $p \in M \subset \mathbb{R}^N$. Let us formalize this in the following definition.

Definition 2.2.1. Consider an n-dimensional continuously differentiable submanifold M of $\mathbb{R}^N$. Let $p \in M$ and assume that a neighbourhood U_p of p can be described in some coordinate system $(V_p, (x, y))$ of $\mathbb{R}^N$ by $(x, y) = (x^1, \ldots, x^n, y^1, \ldots, y^{N-n})$, where

$$y^\alpha = y^\alpha(x^1, \ldots, x^n), \qquad \alpha = 1, \ldots, N-n,$$

and the functions y^α are continuously differentiable. Furthermore, $U_p \subset V_p$, where U_p is n-dimensional and V_p is N-dimensional. Let us assume without loss of generality that

$$p = \big(x^1 = 0, \ldots, x^n = 0, y^1(0, \ldots, 0), \ldots, y^{N-n}(0, \ldots, 0)\big).$$

A vector $\boldsymbol{v} \in \mathbb{R}^N$ originating at p is a tangent vector to M at p if there exists a curve $\gamma : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$, $\varepsilon > 0$, $\gamma(t) = (x^1(t), \ldots, x^n(t))$ such that $\gamma(0) = (x^1(0) = 0, \ldots, x^n(0) = 0)$ and

$$\frac{\mathrm{d}\overline{\gamma}(\gamma(t))}{\mathrm{d}t}\Big|_{t=0} = \boldsymbol{v},$$

where

$$\overline{\gamma}(\gamma(t))t := \Big(x^1(t), \ldots, x^n(t), y^1\big[x^1(t), \ldots, x^n(t)\big], \ldots, y^{N-n}\big[x^1(t), \ldots, x^n(t)\big]\Big) \in \mathbb{R}^N.$$

It is clear that the curve γ giving the tangent vector $\boldsymbol{v}$ at p is not unique. Notice that $\overline{\gamma}$ is a function of the n-variables $x^1, \ldots, x^n$. Applying the chain rule gives us

$$\frac{\mathrm{d}\overline{\gamma}(\gamma(t))}{\mathrm{d}t}\bigg|_{t=0} = \sum_{i=1}^{n} \frac{\mathrm{d}\overline{\gamma}}{\mathrm{d}x^i} \frac{\mathrm{d}x^i(t)}{\mathrm{d}t}\bigg|_{t=0}. \tag{2.3}$$

In this expression the derivatives of $\overline{\gamma}$ are the vectors given by

$$\frac{\mathrm{d}\overline{\gamma}}{\mathrm{d}x^i} = \left(0, \ldots, 0, 1, 0, \ldots, 0, \frac{\partial y^1}{\partial x^i}, \ldots, \frac{\partial y^{N-n}}{\partial x^i}\right), \quad i = 1, \ldots, n, \tag{2.4}$$

where the 1 on the right hand side is in the i-th position. These vectors do not depend on the choice of curve γ, but only on M and the choice of local coordinate system. Every tangent vector is a linear combination of these n vectors in $\mathbb{R}^N$.

Let us now consider a curve $\gamma_j : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$, where $\gamma_j(t) = (0, \ldots, 0, x^j(t) = t, 0, \ldots, 0)$, for $1 \leq j \leq n$. In this case we have

$$\left.\frac{\mathrm{d}\overline{\gamma}(\gamma_j(t))}{\mathrm{d}t}\right|_{t=0} = \sum_{i=1}^{n} \frac{\mathrm{d}\overline{\gamma}}{\mathrm{d}x^i} \left.\frac{\mathrm{d}x^i(t)}{\mathrm{d}t}\right|_{t=0} = \frac{\mathrm{d}\overline{\gamma}}{\mathrm{d}x^j}.$$

From this we see that the vectors (2.4) are the velocity vectors of the curves $\overline{\gamma}_j$ give by

$$\overline{\gamma}(\gamma_j(t))(t) := \Big(0, \ldots, x^j(t) = t, \ldots, 0, y^1\big[0, \ldots, x^j(t) = t, \ldots, 0\big], \ldots, y^{N-n}\big[0, \ldots, x^j(t) = t, \ldots, 0\big]\Big).$$

The curves $\overline{\gamma}(\gamma_j(t))$ are called the *coordinate lines*, because these are the curves obtained by varying one of the n coordinates, while keeping the others constant. Let us denote the n vectors in (2.4) by $\boldsymbol{e}_i$, i.e.

$$\boldsymbol{e}_i := \frac{\mathrm{d}\overline{\gamma}}{\mathrm{d}x^i}.$$

Lemma 2.2.2. *The tangent vectors to M at a point $p \in M$ constitute an n-dimensional subspace of $\mathbb{R}^N$. The n vectors in $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ span this vector space.*

Proof. The vectors $\boldsymbol{e}_i$ have a 1 in the i-th position and all of the other first n components are zero. From this fact it follows that they are linearly independent. Furthermore, (2.3) implies that every tangent vector can be written as a linear combination of the n vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$. The vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ are vectors in $\mathbb{R}^N$ which is a vector space, hence it follows that the tangent vectors constitute an n-dimensional subspace of $\mathbb{R}^N$ and the vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ are a basis for this vector space. □

Definition 2.2.3. The subspace in Lemma 2.2.2 is called the *tangent space* to M at p.

The vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ are a natural choice for a basis for T_pM. The basis $\{\boldsymbol{e}_i\}$ for T_pM is called a *coordinate basis* or *holonomic basis*, because it consists of the velocity vectors to the coordinate lines in the point p. The tangent space T_pM is a concept independent of the choice of coordinate system. Suppose we choose another coordinate system $(W_p, (z, \tilde{y}))$ instead of $(V_p, (x, y))$, where $\tilde{y}^1, \ldots, \tilde{y}^{N-n}$ are continuously differentiable functions of $z^1, \ldots, z^n$. In this new coordinate system $\overline{\gamma}$ is a function of the coordinates $z^1, \ldots, z^n$. By the chain rule it follows that

$$\frac{\partial\overline{\gamma}}{\partial z^j} = \sum_{i=1}^{n} \frac{\partial\overline{\gamma}}{\partial x^i} \frac{\partial x^i}{\partial z^j}.$$

Thus, we see that when changing to another coordinate system the new coordinate basis vectors are linear combinations of the old coordinate basis vectors:

$$\boldsymbol{f}_j := \frac{\partial\overline{\gamma}}{\partial z^j} = \sum_{i=1}^{n} \frac{\partial x^i}{\partial z^j} \boldsymbol{e}_i.$$

It is also possible to choose a basis for T_pM such that there does not exists a coordinate system for which it is the coordinate basis. Such a basis is called *anholonomic*.

We remarked earlier that the n-dimensional Euclidean space is the trivial example of a submanifold of $\mathbb{R}^n$. From the preceding definition of tangent space it follows that the tangent space to $\mathbb{R}^n$ in a point $p \in \mathbb{R}^n$ is an n-dimensional subspace of $\mathbb{R}^n$. However, there is only one possible n-dimensional subspace of $\mathbb{R}^n$ and that is $\mathbb{R}^n$ itself, hence the tangent space $T_p\mathbb{R}^n$ to the point p in $\mathbb{R}^n$ is $\mathbb{R}^n$. Although, the basis for the space $T_p\mathbb{R}^n$ does, in general, depend on p. In Cartesian coordinates the coordinate lines run in straight lines and this results in the fact that in Cartesian coordinates the coordinate basis vectors of $T_p\mathbb{R}^n$ have the same orientation and length for every $p \in \mathbb{R}^n$. In general this is not true as we will see in the following example in which we consider the polar coordinate system.

Example 2.2.4. Let us consider the 2-dimensional Euclidean space described by polar coordinates as in Example 2.1.6. In order to determine the coordinate basis for $T_p\mathbb{R}^2$, we first determine the coordinate lines. The coordinate lines are obtained by varying one of the coordinates while keeping the others constant. Keeping the radius r constant while varying the angle θ we obtain

$$\gamma_\theta(t) = r\cos(t+\theta)\boldsymbol{e}_x + r\sin(t+\theta)\boldsymbol{e}_y,$$

where $\{\boldsymbol{e}_x, \boldsymbol{e}_y\}$ is the trivial Cartesian coordinate basis. Varying the radius r while keeping the angle θ constant we obtain the coordinate lines

$$\gamma_r(t) = (t+r)\cos(\theta)\boldsymbol{e}_x + (t+r)\sin(\theta)\boldsymbol{e}_y.$$

From these coordinate lines we can find how the coordinate basis for the polar coordinate system relates to the trivial coordinate basis of the Cartesian coordinate system. We find

$$\boldsymbol{e}_\theta = \left.\frac{\mathrm{d}\gamma_\theta(t)}{\mathrm{d}t}\right|_{t=0} = -r\sin(\theta)\boldsymbol{e}_x + r\cos(\theta)\boldsymbol{e}_y,$$
$$\boldsymbol{e}_r = \left.\frac{\mathrm{d}\gamma_r(t)}{\mathrm{d}t}\right|_{t=0} = \cos(\theta)\boldsymbol{e}_x + \sin(\theta)\boldsymbol{e}_y.$$

It is clear that the orientation of the coordinate basis for the polar coordinate system depends on θ and the length of $\boldsymbol{e}_\theta$ depends on r, therefore, for the polar coordinate system, the coordinate basis for $T_p\mathbb{R}^2$ depends on $p \in \mathbb{R}^2$.

2.3 The dual space and tensors

In last section we considered tangent vectors to a submanifold M of Euclidean space. We saw that at a point $p \in M$ they constitute the tangent space T_pM. In this section we will construct the dual space to T_pM. This will be necessary to eventually define tensor fields over M. Tensors will play a crucial role in tensor calculus and differential geometry. The tensors as introduced in this section can be introduced in this way for any finite dimensional vector space V. We will present the concepts of this section mainly for an arbitrary finite dimensional vector space V and not constantly refer to T_pM, although the results obviously also apply to T_pM. A vector field on M assigns to each point $p \in M$ a vector in T_pM. In a similar way tensor fields will be introduced. Tensor fields assign to each point $p \in M$ a tensor. Finally, at the end of this section, we will introduce some specific tensors of use to us.

In this section we will heavily rely on the second chapter of [1]. Furthermore, we will from now on follow the *Einstein summation convention*, which states that summation is implied over indices which occur once as lower index and once as upper index.

Let us assume a vector space V over the real numbers with a basis $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$. A vector $\boldsymbol{v} \in V$ can be written as a linear combination of the basisvectors, i.e. $\boldsymbol{v} = v^i\boldsymbol{e}_i$. The real numbers v^i are called the *contravariant* components of the vector $\boldsymbol{v}$ relative to the basis $\{\boldsymbol{e}_i\}$. When we change to another basis for V, say $\{\bar{\boldsymbol{e}}_i\}$, with $\bar{\boldsymbol{e}}_i = A^j_i\boldsymbol{e}_j$, then the contravariant components of the vector $\boldsymbol{v}$ change according to $\bar{v}^i = B^i_j v^j$, where $A^i_k B^k_j = \delta^i_j$ with δ^i_j the familiar *Kronecker symbol*. This is called the *vector transformation law*.

In order to define the dual vector space, we must first define *linear functionals*.

Definition 2.3.1. A linear functional on V is a linear mapping from V into the real numbers, i.e. $\hat{\boldsymbol{f}} : V \to \mathbb{R}$ is a linear functional if and only if

$$\forall \boldsymbol{v}, \boldsymbol{w} \in V,\ \forall \lambda, \mu \in \mathbb{R} : \qquad \hat{\boldsymbol{f}}(\lambda \boldsymbol{v} + \mu \boldsymbol{w}) = \lambda \hat{\boldsymbol{f}}(\boldsymbol{v}) + \mu \hat{\boldsymbol{f}}(\boldsymbol{w}).$$

The set of all linear functionals on V, which we will denote by V^*, turns out to be a vector space when we define the addition of linear functionals by

$$\forall \hat{\boldsymbol{f}}, \hat{\boldsymbol{g}} \in V^*,\ \forall \lambda, \mu \in \mathbb{R},\ \forall \boldsymbol{v} \in V : \qquad (\lambda \hat{\boldsymbol{f}} + \mu \hat{\boldsymbol{g}})(\boldsymbol{v}) := \lambda \hat{\boldsymbol{f}}(\boldsymbol{v}) + \mu \hat{\boldsymbol{g}}(\boldsymbol{v}).$$

The vector space V^* of all linear functionals on V we call the *dual space* and such a linear functional is also called a *covector* or *1-form*. We place a hat ($\hat{}$) above covectors to distinguish them from vectors. The basis $\{\boldsymbol{e}_i\}$ for the vector space V induces a basis $\{\hat{\boldsymbol{e}}^i\}$ for the dual space V^* by demanding that $\hat{\boldsymbol{e}}^i(\boldsymbol{e}_j) = \delta^i_j$. To see that the set of covectors $\{\hat{\boldsymbol{e}}^i\}$ indeed forms a basis for V^*, we check that this set spans V^* and is linearly independent. For any $\hat{\boldsymbol{f}} \in V^*$ and $\boldsymbol{v} = v^i \boldsymbol{e}_i \in V$ we have

$$\hat{\boldsymbol{f}}(\boldsymbol{v}) = \hat{\boldsymbol{f}}(v^i \boldsymbol{e}_i) = v^i \hat{\boldsymbol{f}}(\boldsymbol{e}_i) = \hat{\boldsymbol{e}}^i(\boldsymbol{v}) \hat{\boldsymbol{f}}(\boldsymbol{e}_i) = (\hat{\boldsymbol{f}}(\boldsymbol{e}_i) \hat{\boldsymbol{e}}^i)(\boldsymbol{v}),$$

which implies that each covector in V^* is a linear combination of elements of the set $\{\hat{\boldsymbol{e}}^i\}$. Now suppose that we have some linear combination $\lambda_i \hat{\boldsymbol{e}}^i = 0$, then

$$\forall 1 \leq j \leq n : \qquad 0 = (\lambda_i \hat{\boldsymbol{e}}^i)(\boldsymbol{e}_j) = \lambda_i \delta^i_j = \lambda_j$$

and hence we see that the set $\{\hat{\boldsymbol{e}}^i\}$ is indeed linearly independent. The coefficients $v_i \in \mathbb{R}$ in $\hat{\boldsymbol{v}} = v_i \hat{\boldsymbol{e}}^i$ are called the *covariant components* of the covector $\hat{\boldsymbol{v}}$.

We can of course also consider the dual space to V^*, i.e. $(V^*)^*$. It turns out that for a finite dimensional vector space V, the vector space $(V^*)^*$ is isomorphic to V. The natural isomorphism between $(V^*)^*$ and V is given by the map $\boldsymbol{v} \to \boldsymbol{\psi}(\boldsymbol{v})$, where $\boldsymbol{v} \in V,\ \boldsymbol{\psi} \in (V^*)^*$ and $\boldsymbol{\psi}(\boldsymbol{v})$ is defined by

$$\boldsymbol{\psi}(\boldsymbol{v})(\hat{\boldsymbol{f}}) := \hat{\boldsymbol{f}}(\boldsymbol{v}).$$

We can therefore interpret the vectors as linear functionals on V^* and have

$$\boldsymbol{v}(\hat{\boldsymbol{f}}) = \hat{\boldsymbol{f}}(\boldsymbol{v}).$$

To further emphasize this fact often the following notation is used:

$$\langle \hat{\boldsymbol{f}}, \boldsymbol{v} \rangle := \hat{\boldsymbol{f}}(\boldsymbol{v}) = \boldsymbol{v}(\hat{\boldsymbol{f}}).$$

The covector $\hat{\boldsymbol{f}}$ can in this way also be written as $\langle \hat{\boldsymbol{f}}, \cdot \rangle$. Similarly the vector $\boldsymbol{v}$ can also be written as $\langle \cdot, \boldsymbol{v} \rangle$. This is called the *bracket formalism*. Note that by the definition of the dual basis we have

$$\langle \hat{\boldsymbol{e}}^i, \boldsymbol{e}_j \rangle = \hat{\boldsymbol{e}}^i(\boldsymbol{e}_j) = \delta^i_j.$$

We have the following *covector transformation law*. (See [1, p.10].)

Theorem 2.3.2. *Consider the change of basis $\boldsymbol{f}_j = A^i_j \boldsymbol{e}_i$. This induces a change of dual vector basis given by $\hat{\boldsymbol{f}}^j = B^j_i \hat{\boldsymbol{e}}^i$, in which $A^k_j B^i_k = \delta^i_j$. Consequently if $\hat{\boldsymbol{v}} = v_i \hat{\boldsymbol{e}}^i = \bar{v}_i \hat{\boldsymbol{f}}^i$, then $v_i = B^j_i \bar{v}_j$ and $\bar{v}_i = A^j_i v_j$.*

The proof of this theorem can be found in [1, p.10] and will be left out for brevity.

We have seen that we can interpret vectors and covectors as linear mappings from V^* and V, respectively, to the real numbers. By allowing for more general multi-linear mappings from multiple Cartesian products of V and V^* to the real numbers we get the concept of a *tensor* ([1, p.17]).

Definition 2.3.3. A tensor is a multi-linear mapping

$$\boldsymbol{T} : \underbrace{V^* \times \cdots \times V^*}_{p} \times \underbrace{V \times \cdots \times V}_{q} \to \mathbb{R},$$

for some p and q in $\mathbb{N} \cup \{0\}$. For a definite p and q we denote the space of all such tensors by $\boldsymbol{T}^p_q(V)$. The ordering of the arguments in this definition matters.

We say that a tensor $\boldsymbol{T} \in \boldsymbol{T}^p_q(V)$ has *contravariant rank* p and *covariant rank* q. It is clear that we have $\boldsymbol{T}^1_0(V) = V$ and $\boldsymbol{T}^0_1(V) = V^*$. Moreover we identify $\boldsymbol{T}^0_0(V)$ with $\mathbb{R}$.

Example 2.3.4. The bracket formalism above already provides an example of a tensor. The mapping

$$\langle \cdot, \cdot \rangle : V^* \times V \to \mathbb{R} : (\hat{\boldsymbol{f}}, \boldsymbol{v}) \mapsto \langle \hat{\boldsymbol{f}}, \boldsymbol{v} \rangle,$$

is clearly a multi-linear mapping. This tensor is called the *Kronecker tensor*.

All tensors can be constructed from the vectors and covectors by an operation called the tensor product. To show this we first need to define the *outer product*.

Definition 2.3.5. The outer product $f \otimes g : X \times Y \to \mathbb{R}$ of two real-valued functions $f : X \to \mathbb{R}$ and $g : Y \to \mathbb{R}$ is defined by

$$\forall x \in X,\ \forall y \in Y : \qquad (f \otimes g)(x, y) := f(x)g(y).$$

The vectors and covectors are linear maps from V^* and V to $\mathbb{R}$, respectively, therefore we can use the outer product to define a tensor of arbitrary type. For example for the $(p+q)$-tuple of indices $(i_1, \ldots, i_p, j_1, \ldots, j_q)$ the tensor

$$\boldsymbol{e}_{i_1} \otimes \cdots \otimes \boldsymbol{e}_{i_p} \otimes \hat{\boldsymbol{e}}^{j_1} \otimes \cdots \otimes \hat{\boldsymbol{e}}^{j_q} \tag{2.5}$$

is an element of the space $\boldsymbol{T}^p_q(V)$ and we have

$$\begin{aligned} &\left(\boldsymbol{e}_{i_1} \otimes \cdots \otimes \boldsymbol{e}_{i_p} \otimes \hat{\boldsymbol{e}}^{j_1} \otimes \cdots \otimes \hat{\boldsymbol{e}}^{j_q}\right)(\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_p, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_q) \\ &\quad = \langle \hat{\boldsymbol{v}}_1, \boldsymbol{e}_{i_1} \rangle \cdots \langle \hat{\boldsymbol{v}}_p, \boldsymbol{e}_{i_p} \rangle \langle \hat{\boldsymbol{e}}^{j_1}, \boldsymbol{w}_1 \rangle \cdots \langle \hat{\boldsymbol{e}}^{j_q}, \boldsymbol{w}_q \rangle \end{aligned}$$

for all $\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_p \in V^*$ and $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_q \in V$.

In fact the tensors of the form of equation (2.5) constitute a basis for the space $\boldsymbol{T}^p_q(V)$. This fact is expressed by the following theorem. (See [1, p.9] for the theorem and proof.)

Theorem 2.3.6. *If $\boldsymbol{T} \in \boldsymbol{T}^p_q(V)$, then there exists a set of n^{p+q} numbers $t^{i_1 \ldots i_p}_{j_1 \ldots j_q} \in \mathbb{R}$ such that*

$$\boldsymbol{T} = t^{i_1 \ldots i_p}_{j_1 \ldots j_q} \boldsymbol{e}_{i_1} \otimes \cdots \otimes \boldsymbol{e}_{i_p} \otimes \hat{\boldsymbol{e}}^{j_1} \otimes \cdots \otimes \hat{\boldsymbol{e}}^{j_q}.$$

The collection of coefficients $t^{i_1 \ldots i_p}_{j_1 \ldots j_q}$ is known as the *holor* of the tensor $\boldsymbol{T}$. Note that if we have two tensors $\boldsymbol{T} \in \boldsymbol{T}^p_q(V)$ and $\boldsymbol{S} \in \boldsymbol{T}^r_s(V)$, these are linear mappings to the real numbers and we can take the outer product of the two. The outer product of two tensors is often referred to as the *tensor product*. If the tensor $\boldsymbol{T}$ has holor $t^{i_1 \ldots i_p}_{j_1 \ldots j_q}$ with respect to the basis $\{\boldsymbol{e}_{i_1} \otimes \cdots \otimes \boldsymbol{e}_{i_p} \otimes \hat{\boldsymbol{e}}^{j_1} \otimes \cdots \otimes \hat{\boldsymbol{e}}^{j_q}\}$ and tensor $\boldsymbol{S}$ has holor $s^{i_1 \ldots i_r}_{j_1 \ldots j_s}$ with respect to the basis $\{\boldsymbol{e}_{i_1} \otimes \cdots \otimes \boldsymbol{e}_{i_r} \otimes \hat{\boldsymbol{e}}^{j_1} \otimes \cdots \otimes \hat{\boldsymbol{e}}^{j_s}\}$, then we have

$$\boldsymbol{T} \otimes \boldsymbol{S} = t^{i_1 \ldots i_p}_{j_1 \ldots j_q} s^{i_{p+1} \ldots i_{p+r}}_{j_{q+1} \ldots j_{q+s}} \boldsymbol{e}_{i_1} \otimes \cdots \otimes \boldsymbol{e}_{i_{p+r}} \otimes \hat{\boldsymbol{e}}^{j_1} \otimes \cdots \otimes \hat{\boldsymbol{e}}^{j_{q+s}},$$

and $\boldsymbol{T} \otimes \boldsymbol{S} \in \boldsymbol{T}^{p+r}_{q+s}(V)$. From the transformation laws for vectors and covectors a transformation law for tensors follows. Let us consider a basis transformation given by $\boldsymbol{f}_i = A^j_i \boldsymbol{e}_j$ and let $A^i_k B^k_j = \delta^i_j$. Then we have

$$\boldsymbol{T} = t^{i_1 \ldots i_p}_{j_1 \ldots j_q} \boldsymbol{e}_{i_1} \otimes \cdots \otimes \boldsymbol{e}_{i_p} \otimes \hat{\boldsymbol{e}}^{j_1} \otimes \cdots \otimes \hat{\boldsymbol{e}}^{j_q} = \bar{t}^{i_1 \ldots i_p}_{j_1 \ldots j_q} \boldsymbol{f}_{i_1} \otimes \cdots \otimes \boldsymbol{f}_{i_p} \otimes \hat{\boldsymbol{f}}^{j_1} \otimes \cdots \otimes \hat{\boldsymbol{f}}^{j_q}$$

if and only if the holor adheres to the *tensor transformation law*

$$\bar{t}^{i_1\ldots i_p}_{j_1\ldots j_q} = A^{l_1}_{j_1}\cdots A^{l_q}_{j_q} B^{i_1}_{k_1}\cdots B^{i_p}_{k_p} t^{k_1\ldots k_p}_{l_1\ldots l_q}.$$

We are now in the position to introduce the notion of *tensor fields* on a submanifold of Euclidean space. Let us take for the vector space V in the above the tangent space at some point $x \in M$ for an n-dimensional submanifold of $\mathbb{R}^N$, i.e. T_xM. A (p,q)-*tensor field* on M assigns to each $x \in M$ an element from $\boldsymbol{T}^p_q(T_xM)$. So, for example, a *vector field* on M assigns to each point $x \in M$ a tangent vector in T_xM. We will denote the space of all k-times continuously differentiable vector fields on M by TM_{C^k} and the space of k-times continuously differentiable tensor fields of type (p,q), we denote by $\boldsymbol{T}^p_q(TM)_{C^k}$.

Example 2.3.7. Suppose we have a manifold M and a coordinate system (U, v) for $U \subset M$. The coordinate basis vectors $\boldsymbol{e}_i$ of this coordinate system are examples of vector fields, because $\boldsymbol{e}_i$ assigns to each point $p \in U$ the vector

$$\boldsymbol{e}_i\big|_p \in T_pM.$$

Given an inner product, $(\cdot|\cdot) : V \times V \to \mathbb{R}$, and a basis $\{\boldsymbol{e}_i\}$ for V then the coefficients $g_{ij} := (\boldsymbol{e}_i|\boldsymbol{e}_j)$ are the components of a tensor $\boldsymbol{g}$ called the *metric tensor*. The matrix with components g_{ij} is called the *Gram matrix* G. The symmetry of the inner product implies that G is symmetric. The components of the *inverse Gram matric* G^{-1} are denoted by g^{ij}, i.e. $g^{ik}g_{kj} = \delta^i_j$. Suppose we have two vectors $\boldsymbol{v} = v^i\boldsymbol{e}_i$ and $\boldsymbol{w} = w^i\boldsymbol{e}_i$ then by the linearity of the inner product we have

$$(\boldsymbol{v} \mid \boldsymbol{w}) = (v^i\boldsymbol{e}_i \mid w^j\boldsymbol{e}_j) = v^iw^j(\boldsymbol{e}_i \mid \boldsymbol{e}_j) = v^iw^jg_{ij}.$$

With the use of an inner product we can establish an important bijection between V and V^*. (From [1, p.13].)

Theorem 2.3.8. *There exists a linear bijection $\boldsymbol{G} : V \to V^*$ such that*

(i) $\forall \boldsymbol{v}, \boldsymbol{w} \in V : \quad (\boldsymbol{v}|\boldsymbol{w}) = \langle \boldsymbol{G}(\boldsymbol{v}), \boldsymbol{w}\rangle$*, and,*

(ii) $\forall \hat{\boldsymbol{v}} \in V^*,\ \forall \boldsymbol{w} \in V : \quad (\boldsymbol{G}^{-1}(\hat{\boldsymbol{v}})|\boldsymbol{w}) = \langle \hat{\boldsymbol{v}}, \boldsymbol{w}\rangle$.

The matrix representations of $\boldsymbol{G}$ and $\boldsymbol{G}^{-1}$ are given by the Gram matrix G and its inverse G^{-1}, respectively.

The proof of this theorem can be found in [1]. This bijection gives rise to the following useful operators.

Definition 2.3.9. The *conversion operators* $\sharp : V \to V^*$ and $\flat : V^* \to V$ are defined by $\sharp\boldsymbol{v} = \boldsymbol{G}(\boldsymbol{v})$ and $\flat\hat{\boldsymbol{v}} = \boldsymbol{G}^{-1}(\hat{\boldsymbol{v}})$. These operators are called the *sharp* and *flat* operators, respectively. The conversion operators are also called the *musical isomorphisms*.

In terms of the components we have for a vector $\boldsymbol{v} = v^i\boldsymbol{e}_i$ and covector $\hat{\boldsymbol{w}} = w_i\hat{\boldsymbol{e}}^i$, respectively, that

$$\sharp\boldsymbol{v} = g_{ij}v^i\hat{\boldsymbol{e}}^j \quad \text{and} \quad \flat\hat{\boldsymbol{w}} = g^{ij}w_i\boldsymbol{e}_j.$$

For Cartesian coordinates in Euclidean space, the Gram matrix is just the identity and we have $v_j = g_{ij}v^i = v^j$ and $w^i = g^{ij}w_j = w_i$. Thus for Cartesian coordinates in Euclidean space the vector $\boldsymbol{v}$ and the covector $\sharp\boldsymbol{v}$ have the same components (with respect to different bases) and therefore the distinction between vectors and covectors is not really apparent in this coordinate system.

Lastly, we will define some tensors that will be useful later on. We first define the completely anti-symmetric symbol, then the Levi-Civita tensor and then the generalized Kronecker tensor.

Definition 2.3.10. The completely anti-symmetric symbol, which we denote by $[i_1, \ldots, i_n]$, is defined by

$$[i_1, \ldots, i_n] := \begin{cases} 1 & \text{if } (i_1, \ldots, i_n) \text{ is an even permutation of } (1, \ldots, n), \\ -1 & \text{if } (i_1, \ldots, i_n) \text{ is an odd permutation of } (1, \ldots, n), \\ 0 & \text{otherwise.} \end{cases}$$

We can use the completely anti-symmetric symbol to determine the determinant of square matrices. Let (A_{ij}) be a square matrix then by developing with respect to the first column we find that

$$\det(A_{ij}) = \sum_{i_1,\ldots,i_n=1}^{n} [i_1, \ldots, i_n] A_{1i_1} \cdots A_{ni_n},$$

or equivalently by developing with respect to the first row we find that

$$\det(A_{ij}) = \sum_{i_1,\ldots,i_n=1}^{n} [i_1, \ldots, i_n] A_{i_1 1} \cdots A_{i_n n},$$

By taking an arbitrary permutation of the rows of the matrix we find that

$$[j_1, \ldots, j_n] \det(A_{ij}) = \sum_{i_1,\ldots,i_n=1}^{n} [i_1, \ldots, i_n] A_{i_1 j_1} \cdots A_{i_n j_n}. \tag{2.6}$$

The determinant of the metric tensor we will denote by the letter g, i.e.

$$g := \det(g_{ij}).$$

It is clear that $\det(g^{ij}) = 1/g$. With use of (2.6) it follows that

$$g^{i_1 j_1} \cdots g^{i_n j_n} [j_1, \ldots, j_n] = \det(g^{ij}) [i_1, \ldots, i_n].$$

Let us now use the anti-symmetric symbol to define the Levi-Civita tensor.

Definition 2.3.11. Consider a n-dimensional vector space with metric $\boldsymbol{g}$. The Levi-Civita tensor is the tensor with contravariant rank 0 and covariant rank n and components given by

$$\epsilon_{i_1 \ldots i_n} = \sqrt{g} [i_1, \ldots, i_n].$$

The contravariant representation of the Levi-Civita tensor is given by

$$\epsilon^{i_1 \ldots i_n} = \frac{1}{\sqrt{g}} \, [i_1, \ldots, i_n]$$

The definition of the contravariant and covariant representations of the Levi-Civita tensor are consistent, i.e. if we raise the components of covariant representation with the metric tensor we get the contravariant representation:

$$\begin{aligned} \epsilon^{i_1 \ldots i_n} &= g^{i_1 j_1} \cdots g^{i_n j_n} \epsilon_{j_1 \ldots j_n} \\ &= \sqrt{g} \; g^{i_1 j_1} \cdots g^{i_n j_n} [j_1, \ldots, j_n] \\ &= \sqrt{g} \det(g^{ij}) [i_1, \ldots, i_n] \\ &= \frac{1}{\sqrt{g}} [i_1, \ldots, i_n]. \end{aligned}$$

Here we used that $g^{i_1 j_1} \cdots g^{i_n j_n} [j_1, \ldots, j_n] = \det(g^{ij}) [i_1, \ldots, i_n]$.

The Levi-Civita tensor can be used to determine the *cross product* in $\mathbb{R}^3$. Suppose we have two vectors $\boldsymbol{v} = v^i \boldsymbol{e}_i$ and $\boldsymbol{w} = w^i \boldsymbol{e}_i$. The cross product of these two vectors is given by

$$\boldsymbol{v} \times \boldsymbol{w} = \epsilon_{ijk} v^i w^j g^{kl} \boldsymbol{e}_l,$$

for any coordinate system on $\mathbb{R}^3$ with metric $\boldsymbol{g}$ and basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$.

Related to the Levita-Civita tensor is the *generalized Kronecker tensor*. The following definition is from [1, p.40].

Definition 2.3.12. The generalized Kronecker tensor is the tensor with components given by

$$\delta^{i_1 \dots i_n}_{j_1 \dots j_n} := \begin{cases} +1 & \text{if } (i_1, \dots, i_n) \text{ is an even permutation of } (j_1, \dots, j_n), \\ -1 & \text{if } (i_1, \dots, i_n) \text{ is an odd permutation of } (j_1, \dots, j_n), \\ 0 & \text{otherwise.} \end{cases}$$

The generalized Kronecker tensor can be written as the product of the covariant and contravariant Levi-Civita tensor:

$$\delta^{i_1 \dots i_n}_{j_1 \dots j_n} = \epsilon_{j_1 \dots j_n} \epsilon^{i_1 \dots i_n}.$$

2.4 The covariant derivative

In this section we will focus on differentiation of scalars, vectors and more generally tensors. When working in Euclidean space with Cartesian coordinates, differentiation of tensors can be performed component-wise, because the basis of $T_p\mathbb{R}^n$ in Cartesian coordinates does not depend on $p \in \mathbb{R}^n$. However, when working in a different coordinate system or in non-Euclidean space we cannot just differentiate component-wise. In this section we will introduce the *covariant derivative* which is a coordinate independent way of differentiating tensors, because a covariant derivative of a tensor is again a tensor. We will start by considering the directional derivative of scalar fields on Euclidean spaces and look closer into the differences between holonomic bases and anholonomic bases. Then, we will consider the directional derivative of vector fields in Euclidean space for an arbitrary coordinate system and extend this to the directional derivative of tensors of any type. In the process of doing this, we will introduce Christoffel symbols and commutation symbols. The commutation symbols will give us more insight in the difference between holonomic bases and anholonomic bases. We will closely examine the properties of the directional derivative in Euclidean space. The covariant derivative is then introduced as a directional derivative operator for general, possibly non-Euclidean, spaces. It then becomes clear that the directional derivative is the covariant derivative for the special case that the space under consideration is a Euclidean space. We will not use the covariant derivative in the non-Euclidean context, but the definition of the covariant derivative is a convenient one to rely on in Chapter 3.

Let us consider the scalar function $u : U \subset \mathbb{R}^n \to \mathbb{R}$ on a subspace of n-dimensional Euclidean space. Suppose that we have a coordinate system $x^1, \dots, x^n$ on U with coordinate basis $\{\boldsymbol{e}_i\}$. This basis depends on the position $p \in U$. To determine how u changes in a certain direction $\boldsymbol{v} = v^i \boldsymbol{e}_i \in T_p\mathbb{R}^n$ we determine the directional derivative:

$$\nabla_{\boldsymbol{v}} u := \left. \frac{\mathrm{d}(u \circ \gamma_{\boldsymbol{v}})(t)}{\mathrm{d}t} \right|_{t=0},$$

where $\gamma_{\boldsymbol{v}} : (-\varepsilon, \varepsilon) \to U$ is a curve such that $\gamma_{\boldsymbol{v}}(0) = p$ and

$$\left. \frac{\mathrm{d}\gamma_{\boldsymbol{v}}(t)}{\mathrm{d}t} \right|_{t=0} = \boldsymbol{v}.$$

By applying the chain rule we find

$$\nabla_{\boldsymbol{v}} u = \left. \frac{\mathrm{d}(u \circ \gamma_{\boldsymbol{v}})(t)}{\mathrm{d}t} \right|_{t=0} = \left. \frac{\partial u}{\partial x^j} \right|_p \left. \frac{\mathrm{d}\gamma^j_{\boldsymbol{v}}}{\mathrm{d}t} \right|_{t=0} = \left. \frac{\partial u}{\partial x^j} \right|_p v^j. \tag{2.7}$$

If we take $\boldsymbol{v} = \boldsymbol{e}_i$ we find how u changes along the coordinate line of the i-th coordinate. The components of $\boldsymbol{e}_i$ are given by δ_i^j, hence it follows by (2.7) that

$$\nabla_{\boldsymbol{e}_i} u = \left.\frac{\partial u}{\partial x^i}\right|_p. \tag{2.8}$$

From (2.7) and (2.8) it follows that we have

$$\nabla_{\boldsymbol{v}} u = v^i \nabla_{\boldsymbol{e}_i} u.$$

This implies that the directional derivative $\nabla_{\boldsymbol{v}}$ is not only linear in u but also in $\boldsymbol{v}$, i.e. it holds that

$$\nabla_{\lambda\boldsymbol{v}+\mu\boldsymbol{w}}(u) = \lambda\nabla_{\boldsymbol{v}} u + \mu\nabla_{\boldsymbol{w}} u.$$

However, it should be noticed that (2.8) only holds for the coordinate basis vector $\boldsymbol{e}_i$ corresponding to the coordinate line of the coordinate x^i. If we work in an anholonomic basis $\{\boldsymbol{f}_i\}$ different from the coordinate basis $\{\boldsymbol{e}_i\}$ for the coordinates $x^1, \ldots, x^n$, then

$$\nabla_{\boldsymbol{v}} u = \nabla_{v^i \boldsymbol{f}_i} u \neq v^i \left.\frac{\partial u}{\partial x^i}\right|_p.$$

Suppose that the relation between the general basis and the coordinate basis is given by $\boldsymbol{f}_i = A_i^j \boldsymbol{e}_j$, then it holds that

$$\nabla_{\boldsymbol{v}} u = \nabla_{v^i \boldsymbol{f}_i} u = \nabla_{v^i A_i^j \boldsymbol{e}_j} u = v^i A_i^j \left.\frac{\partial u}{\partial x^j}\right|_p.$$

For coordinate bases $\{\boldsymbol{e}_i\}$, by definition, there always exists a system of coordinates $x^1, \ldots, x^n$ such that the $\boldsymbol{e}_i$ are the velocity vectors to the coordinate lines and hence

$$\forall u \in C^1(U): \qquad \nabla_{\boldsymbol{e}_i} u = \left.\frac{\partial u}{\partial x^i}\right|_p.$$

From this it follows that

$$\nabla_{\boldsymbol{e}_i}(\nabla_{\boldsymbol{e}_j} u) = \left.\frac{\partial^2 u}{\partial x^i \partial x^j}\right|_p = \left.\frac{\partial^2 u}{\partial x^j \partial x^i}\right|_p = \nabla_{\boldsymbol{e}_j}(\nabla_{\boldsymbol{e}_i} u).$$

This only holds for coordinate basis vectors and for general vectors this property does not hold. The fact that directional derivatives do not commute is expressed by the *Lie derivative*, which we will now define for the more general setting of a submanifold of Euclidean space.

Definition 2.4.1. Assume M to be a submanifold of a Euclidean space and let $\boldsymbol{v}, \boldsymbol{w} \in TM$. The Lie derivative of $\boldsymbol{w}$ with respect to $\boldsymbol{v}$ is defined as

$$\mathcal{L}_{\boldsymbol{v}} \boldsymbol{w} := [\nabla_{\boldsymbol{v}}, \nabla_{\boldsymbol{w}}],$$

in which the *Lie bracket* is defined by the following commutator

$$\forall u \in C^1(M): \qquad [\nabla_{\boldsymbol{v}}, \nabla_{\boldsymbol{w}}] u = \nabla_{\boldsymbol{v}}(\nabla_{\boldsymbol{w}} u) - \nabla_{\boldsymbol{w}}(\nabla_{\boldsymbol{v}} u).$$

We will denote the partial derivative of u with respect to x^i by $\partial_i u$. A quick calculation shows that the Lie derivative is again a directional derivative:

$$\begin{aligned}
[\nabla_{\boldsymbol{v}}, \nabla_{\boldsymbol{w}}] u &= \nabla_{\boldsymbol{v}}(\nabla_{\boldsymbol{w}} u) - \nabla_{\boldsymbol{w}}(\nabla_{\boldsymbol{v}} u) \\
&= \nabla_{\boldsymbol{v}}(w^i \partial_i u) - \nabla_{\boldsymbol{w}}(v^i \partial_i u) \\
&= v^j \partial_j (w^i \partial_i u) - w^j \partial_j (v^i \partial_i u) \\
&= v^j w^i \partial_j \partial_i u + v^j \partial_j (w^i) \partial_i (u) - w^j v^i \partial_j \partial_i u - w^j \partial_j (v^i) \partial_i (u) \\
&= v^j \partial_j (w^i) \partial_i (u) - w^j \partial_j (v^i) \partial_i (u) \\
&= \nabla_{(v^j \partial_j w^i - w^j \partial_j v^i) \boldsymbol{e}_i} u.
\end{aligned}$$

We remarked on page 24 that a basis $\{\boldsymbol{e}_i\}$ is a coordinate basis if and only if $[\nabla_{\boldsymbol{e}_i}, \nabla_{\boldsymbol{e}_j}] = 0$. This implies that for an anholonomic basis $\{\boldsymbol{f}_i\}$ we always have $[\nabla_{\boldsymbol{f}_i}, \nabla_{\boldsymbol{f}_j}] \neq 0$ for at least one pair of basis vectors. The Lie derivative is again a directional derivative and therefore there exist coefficients c_{ij}^k such that

$$[\nabla_{\boldsymbol{f}_i}, \nabla_{\boldsymbol{f}_j}] = c_{ij}^k \nabla_{\boldsymbol{f}_k}.$$

The coefficients c_{ij}^k are called the *commutation symbols*. It is clear that $c_{ij}^k = -c_{ji}^k$. Thus, for an anholonomic basis there exist always at least two nonzero commutation symbols. If the relation between a coordinate basis and an anholonomic basis is known then we can derive the commutation symbols for the anholonomic basis. This we show in the following lemma.

Lemma 2.4.2. *Suppose that we have an anholonomic basis $\{\boldsymbol{f}_i\}$ and a coordinate basis $\{\boldsymbol{e}_i\}$ corresponding to a coordinate system (U, x). Furthermore, assume that the coordinate basis and the anholonomic basis are related by $\boldsymbol{f}_j = A_j^i \boldsymbol{e}_i$ and that (B_j^i) is the inverse of (A_j^i), i.e. $A_j^k B_k^i = \delta_j^i$ and $B_j^k A_k^i = \delta_j^i$. Then the commutation symbols are given by*

$$c_{ij}^k = \left[A_i^l \left(\frac{\partial A_j^s}{\partial x^l} \right) - A_j^l \left(\frac{\partial A_i^s}{\partial x^l} \right) \right] B_s^k.$$

Proof. Let $u \in C^1(U)$ be arbitrary. By linearity of the directional derivative and the commutativity of partial derivatives it follows that

$$\begin{aligned}
[\nabla_{\boldsymbol{f}_i}, \nabla_{\boldsymbol{f}_j}]u &= \nabla_{\boldsymbol{f}_i}(\nabla_{\boldsymbol{f}_j} u) - \nabla_{\boldsymbol{f}_j}(\nabla_{\boldsymbol{f}_i} u) \\
&= A_i^l \partial_l \big(A_j^s (\partial_s u) \big) - A_j^l \partial_l \big(A_i^s (\partial_s u) \big) \\
&= A_i^l \partial_l (A_j^s) \partial_s(u) + A_i^l A_j^s \partial_l \partial_s u - A_j^l \partial_l (A_i^s) \partial_s(u) - A_j^l A_i^s \partial_l \partial_s u \\
&= A_i^l \partial_l (A_j^s) \partial_s(u) + A_i^l A_j^s \partial_l \partial_s u - A_j^l \partial_l (A_i^s) \partial_s(u) - A_j^s A_i^l \partial_l \partial_s u \\
&= \big[A_i^l \partial_l (A_j^s) - A_j^l \partial_l (A_i^s) \big] \partial_s u \\
&= \big[A_i^l \partial_l (A_j^s) - A_j^l \partial_l (A_i^s) \big] B_s^k \nabla_{A_k^t \boldsymbol{e}_t} u \\
&= \big[A_i^l \partial_l (A_j^s) - A_j^l \partial_l (A_i^s) \big] B_s^k \nabla_{\boldsymbol{f}_k} u.
\end{aligned}$$

□

We will now consider an example of an anholonomic basis in $\mathbb{R}^2$ that we will use later on in this thesis.

Example 2.4.3. Consider the coordinate basis $\{\boldsymbol{e}_r, \boldsymbol{e}_\theta\}$ for the polar coordinate system, which we derived in Example 2.2.4. We noticed that the vector $\boldsymbol{e}_\theta$ has length r in the point $(r, \theta) \in \mathbb{R}^2$. Often it is chosen to work in the orthonormal basis $\{\overline{\boldsymbol{e}}_r := \boldsymbol{e}_r, \overline{\boldsymbol{e}}_\theta = \boldsymbol{e}_\theta / r\}$. We number the basis vectors corresponding to r by 1 and the basis vectors corresponding to θ by 2 in order to be able to conveniently apply the index notation. The two bases are related by $\overline{\boldsymbol{e}}_j = A_j^i \boldsymbol{e}_i$ and $\boldsymbol{e}_j = B_j^i \overline{\boldsymbol{e}}_i$, where

$$(A_j^i) = \begin{pmatrix} 1 & 0 \\ 0 & 1/r \end{pmatrix} \quad \text{and} \quad (B_j^i) = \begin{pmatrix} 1 & 0 \\ 0 & r \end{pmatrix}.$$

We can now apply Lemma 2.4.2 to determine the commutation symbols for the basis $\{\overline{\boldsymbol{e}}_i\}$. Doing this, we find that there are two nonzero commutation symbols:

$$\begin{aligned}
c_{r\theta}^\theta = -c_{\theta r}^\theta &= \left[A_r^l \left(\frac{\partial A_\theta^s}{\partial x^l} \right) - A_\theta^l \left(\frac{\partial A_r^s}{\partial x^l} \right) \right] B_s^\theta \\
&= \left[A_r^r \left(\frac{\partial A_\theta^\theta}{\partial r} \right) - A_\theta^\theta \left(\frac{\partial A_r^\theta}{\partial \theta} \right) \right] r \\
&= A_r^r \left(\frac{\partial A_\theta^\theta}{\partial r} \right) r \\
&= -\frac{1}{r}.
\end{aligned}$$

Thus, we see that the orthonormal basis $\{\overline{\boldsymbol{e}}_r, \overline{\boldsymbol{e}}_\theta\}$ is an anholonomic basis.

Let us now turn our attention to the directional derivatives of vector fields. Suppose we have two vector fields $\boldsymbol{v} \in T\mathbb{R}^n,\ \boldsymbol{w} \in T\mathbb{R}^n_{C^1}$ and we want to know how the vector field $\boldsymbol{w}$ changes in the direction of $\boldsymbol{v}$, then we look at $\nabla_{\boldsymbol{v}}\boldsymbol{w}$. Let us assume that we have a basis $\{\boldsymbol{e}_i\}$, which is not necessarily holonomic. By linearity of the directional derivative we find

$$\nabla_{\boldsymbol{v}}\boldsymbol{w} = v^i \nabla_{\boldsymbol{e}_i}\boldsymbol{w}.$$

In order to evaluate the directional derivative $\nabla_{\boldsymbol{e}_i}\boldsymbol{w}$ we use the product rule and obtain

$$\nabla_{\boldsymbol{e}_i}\boldsymbol{w} = \nabla_{\boldsymbol{e}_i}(w^j)\boldsymbol{e}_j + w^j \nabla_{\boldsymbol{e}_i}(\boldsymbol{e}_j).$$

We know from the directional derivative in Cartesian coordinates that the directional derivative of a vector is again a vector and hence $\nabla_{\boldsymbol{e}_i}\boldsymbol{e}_j$ needs to be an element of $T\mathbb{R}^n$ again. It follows that there must exist coefficients Γ^k_{ij} such that

$$\nabla_{\boldsymbol{e}_i}\boldsymbol{e}_j = \Gamma^k_{ji}\boldsymbol{e}_k. \tag{2.9}$$

These coefficients are called the *Christoffel symbols*. If we work in the Cartesian coordinate basis, then the basis vectors do not depend on position and therefore the directional derivatives of the basis vectors will be 0. It follows that for the Cartesian coordinate basis the Christoffel symbols will all be 0. When we take the directional derivative of the vector field $\boldsymbol{w}$ in the direction of $\boldsymbol{v}$ for all points $p \in \mathbb{R}^n$ we again end up with a vector field. Furthermore, the directional derivative field is given by

$$\nabla_{\boldsymbol{v}}\boldsymbol{w} = v^i \nabla_{\boldsymbol{e}_i}\boldsymbol{w} = v^i D_i(w^k)\boldsymbol{e}_k,$$

where we have used $D_i(w^k)$ to denote the components of the directional derivative of $\boldsymbol{w}$ in the direction of $\boldsymbol{e}_i$. These components are given by

$$D_i(w^k) = \nabla_{\boldsymbol{e}_i} w^k + \Gamma^k_{ji} w^j. \tag{2.10}$$

So far, we have seen that the directional derivative in Euclidean space has certain properties. Given two vector fields $\boldsymbol{v} \in T\mathbb{R}^n,\ \boldsymbol{w} \in T\mathbb{R}^n_{C^1}$ the directional derivative field $\nabla_{\boldsymbol{v}}\boldsymbol{w}$ is again a vector field, i.e. $\nabla_{\boldsymbol{v}}\boldsymbol{w} \in T\mathbb{R}^n$. Furthermore, we noticed that it is linear in both of its arguments, i.e. the vector that is differentiated and the direction of differentiation. Moreover, the directional derivative clearly also satisfies

$$\forall u \in C^1(\mathbb{R}^n): \qquad \nabla_{\boldsymbol{v}}(u\boldsymbol{w}) = (\nabla_{\boldsymbol{v}} u)\boldsymbol{w} + u\nabla_{\boldsymbol{v}}\boldsymbol{w}. \tag{2.11}$$

Let us consider the Cartesian coordinate basis $\{\boldsymbol{e}_i\}$ and two vector fields $\boldsymbol{v}, \boldsymbol{w} \in T\mathbb{R}^n_{C^1}$. In this basis all derivatives of basis vectors are zero and the metric is the identity, therefore by the product rule it follows that

$$\begin{aligned}
\nabla_{\boldsymbol{e}_i}(\boldsymbol{v} \mid \boldsymbol{w}) &= \nabla_{\boldsymbol{e}_i}\left(\sum_{j=1}^n v^j w^j\right) \\
&= \frac{\partial}{\partial x^i}\left(\sum_{j=1}^n v^j w^j\right) \\
&= \sum_{j=1}^n (\partial_i v^j) w^j + \sum_{j=1}^n v^j (\partial_i w^j) \\
&= (\partial_i \boldsymbol{v} \mid \boldsymbol{w}) + (\boldsymbol{v} \mid \partial_i \boldsymbol{w}) \\
&= (\nabla_{\boldsymbol{e}_i}\boldsymbol{v} \mid \boldsymbol{w}) + (\boldsymbol{v} \mid \nabla_{\boldsymbol{e}_i}\boldsymbol{w}).
\end{aligned}$$

From the above and the linearity of the inner product and the directional derivative we see that $\nabla_{\boldsymbol{u}}(\boldsymbol{v} \mid \boldsymbol{w}) = (\nabla_{\boldsymbol{u}}\boldsymbol{v} \mid \boldsymbol{w}) + (\boldsymbol{v} \mid \nabla_{\boldsymbol{u}}\boldsymbol{w})$. This expression only involves vectors, which are objects independent of coordinate systems, and therefore makes no reference to a particular coordinate system or basis and therefore holds in all. We can rewrite this expression component-wise for a general, possibly anholonomic, basis $\{\boldsymbol{e}_i\}$ as

$$D_k(g_{ij}v^i w^j) = g_{ij}D_k(v^i)w^j + g_{ij}v^i D_k(w^j), \tag{2.12}$$

where $g_{ij} = (\boldsymbol{e}_i \mid \boldsymbol{e}_j)$ is the metric, $D_k(v^i)$ and $D_k(w^j)$ are the components of the directional derivative, in the direction of $\boldsymbol{e}_k$, of the vectors $\boldsymbol{v}$ and $\boldsymbol{w}$, respectively. Moreover, we have introduced here for consistency the notation $D_k(u) := \nabla_{\boldsymbol{e}_k}u$, for the directional derivative of a scalar. It can be shown that $\nabla_{\boldsymbol{e}_i}\hat{\boldsymbol{e}}^j = -\Gamma^j_{ki}\hat{\boldsymbol{e}}^k$. (See for example [7, ex.8.12].) From this it follows that the directional derivative of a covector field $\hat{\boldsymbol{v}} \in T(\mathbb{R}^n)^*$ is given by

$$\nabla_{\boldsymbol{e}_i}\hat{\boldsymbol{v}} = \nabla_{\boldsymbol{e}_i}(v_j\hat{\boldsymbol{e}}^j) = \nabla_{\boldsymbol{e}_i}(v_j)\hat{\boldsymbol{e}}^j + v_j\nabla_{\boldsymbol{e}_i}(\hat{\boldsymbol{e}}^j) = \left(\nabla_{\boldsymbol{e}_i}(v_k) - \Gamma^j_{ki}v_j\right)\hat{\boldsymbol{e}}^k = D_i(v_k)\hat{\boldsymbol{e}}^k, \tag{2.13}$$

where we use $D_i(v_k)$ to denote the components of the directional derivative $\nabla_{\boldsymbol{e}_i}\hat{\boldsymbol{v}}$ An example of a covector is the *differential* of a scalar function. Suppose we have a function $u \in C^1(\mathbb{R}^n)$, then the gradient of u is defined as the covector

$$\mathrm{d}u := (\nabla_{\boldsymbol{e}_i}u)\hat{\boldsymbol{e}}^i.$$

It is also shown in [7, p.214] that the directional derivative has the property

$$\nabla_{\boldsymbol{e}_i}(\boldsymbol{T} \otimes \boldsymbol{S}) = (\nabla_{\boldsymbol{e}_i}\boldsymbol{T}) \otimes \boldsymbol{S} + \boldsymbol{T} \otimes (\nabla_{\boldsymbol{e}_i}\boldsymbol{S}),$$

for general tensor fields on $\mathbb{R}^n$. Combining this fact with $\nabla_{\boldsymbol{e}_i}\hat{\boldsymbol{e}}^j = -\Gamma^j_{ki}\hat{\boldsymbol{e}}^k$ and $\nabla_{\boldsymbol{e}_i}\boldsymbol{e}_j = \Gamma^k_{ji}\boldsymbol{e}_k$ it follows analogously to (2.13) that for a tensor $\boldsymbol{T}$ of covariant rank 1 and contravariant rank 1 it holds that $\nabla_{\boldsymbol{e}_i}\boldsymbol{T} = D_i(T^l_k)\hat{\boldsymbol{e}}^k \otimes \boldsymbol{e}_l$, where the components of the directional derivative are given by

$$D_i(T^l_k) = \nabla_{\boldsymbol{e}_i}T^l_k + \Gamma^l_{ti}T^t_k - \Gamma^t_{ki}T^l_t.$$

This reasoning can be extended to any type of tensor. For every contravariant index we get an extra Christoffel symbol and for every covariant index we get an extra Christoffel symbol with a minus sign in front. Notice that the differentiation index is always in the second lower slot of the Christoffel symbol. Now that we know how to differentiate tensors of every type we are in the position to define a tensor that will be of particular interest to us and will eventually play a role in the energy conservation equation that we will derive in next chapter. This tensor is the *Hessian tensor* and it is defined as follows.

Definition 2.4.4. Given a twice continuously differentiable function u, the Hessian tensor is the tensor with covariant rank 2 and contravariant rank 0 which is given by

$$\boldsymbol{H} := \nabla_{\boldsymbol{e}_j}\mathrm{d}u \otimes \hat{\boldsymbol{e}}^j = \left(\nabla_{\boldsymbol{e}_j}(\nabla_{\boldsymbol{e}_i}u) - \Gamma^k_{ij}\nabla_{\boldsymbol{e}_k}u\right)\hat{\boldsymbol{e}}^i \otimes \hat{\boldsymbol{e}}^j.$$

In Cartesian coordinates, the Hessian tensor of a function u is given by $\boldsymbol{H} = \partial_j\partial_i u(\hat{\boldsymbol{e}}^i \otimes \hat{\boldsymbol{e}}^j)$ and the matrix representation of the Hessian tensor is just the ordinary Hessian matrix

$$\begin{pmatrix} \dfrac{\partial^2 u}{\partial x^2} & \dfrac{\partial^2 u}{\partial x \partial y} \\ \dfrac{\partial^2 u}{\partial x \partial y} & \dfrac{\partial^2 u}{\partial y^2} \end{pmatrix}.$$

However, in other coordinate systems the components are not given by the second order partial derivatives as we will soon see in the following examples. We see that in the Cartesian coordinates

with holonomic basis the Hessian tensor is a symmetric tensor. Symmetry of a tensor is a coordinate independent property. To see this assume that H_{ij} are the components of the Hessian in one coordinate system, $\overline{H}_{ij}$ are the components in another coordinate system and that the coordinate systems are related by the transformation A^i_j. Thus, for example, H_{ij} are the components with respect to the basis $\{\hat{\boldsymbol{e}}^i \otimes \hat{\boldsymbol{e}}^j\}$ and $\overline{H}_{ij}$ are the components with respect to another basis $\{\hat{\boldsymbol{f}}^i \otimes \hat{\boldsymbol{f}}^j\}$ and these different bases are related by $\boldsymbol{f}_j = A^i_j \boldsymbol{e}_i$. If $H_{ij} = H_{ji}$, then it follows that also $\overline{H}_{ij} = \overline{H}_{ji}$, because

$$\overline{H}_{ij} = A^k_i A^l_j H_{kl} = A^k_i A^l_j H_{lk} = A^l_i A^k_j H_{kl} = \overline{H}_{ji}.$$

The symmetry of the Hessian tensor implies that we have

$$\forall u \in C^1(\mathbb{R}^n): \qquad \big(\nabla_{\boldsymbol{e}_j}(\nabla_{\boldsymbol{e}_i} u) - \Gamma^k_{ij}\nabla_{\boldsymbol{e}_k} u\big) - \big(\nabla_{\boldsymbol{e}_i}(\nabla_{\boldsymbol{e}_j} u) - \Gamma^k_{ji}\nabla_{\boldsymbol{e}_k} u\big) = 0. \tag{2.14}$$

If we futher elaborate (2.14), we find

$$\begin{aligned} 0 &= \big(\nabla_{\boldsymbol{e}_j}(\nabla_{\boldsymbol{e}_i} u) - \Gamma^k_{ij}\nabla_{\boldsymbol{e}_k} u\big) - \big(\nabla_{\boldsymbol{e}_i}(\nabla_{\boldsymbol{e}_j} u) - \Gamma^k_{ji}\nabla_{\boldsymbol{e}_k} u\big) \\ &= [\nabla_{\boldsymbol{e}_j}, \nabla_{\boldsymbol{e}_i}]u - \Gamma^k_{ij}\nabla_{\boldsymbol{e}_k} u + \Gamma^k_{ji}\nabla_{\boldsymbol{e}_k} u \\ &= \big(c^k_{ji} - \Gamma^k_{ij} + \Gamma^k_{ji}\big)\nabla_{\boldsymbol{e}_k} u. \end{aligned}$$

This implies that the linear combination of commutation symbols and Christoffel symbols

$$T^k_{ij} := c^k_{ji} - \Gamma^k_{ij} + \Gamma^k_{ji} \tag{2.15}$$

needs to be zero for every choice of coordinate system and basis. This implies that the numbers T^k_{ij} transform according to the tensor transformation law and hence are the components of a tensor. This tensor is called the *torsion tensor*. Let us now define the torsion tensor in a coordinate-free way for the more general setting of submanifolds of Euclidean space.

Definition 2.4.5. Assume M to be a submanifold of a Euclidean space and let $\boldsymbol{v}, \boldsymbol{w} \in TM$. The torsion tensor is defined by*

$$\boldsymbol{T}(\boldsymbol{v}, \boldsymbol{w}) = \nabla_{\nabla_{\boldsymbol{v}}\boldsymbol{w}} - \nabla_{\nabla_{\boldsymbol{w}}\boldsymbol{v}} - [\nabla_{\boldsymbol{v}}, \nabla_{\boldsymbol{w}}]\,. \tag{2.16}$$

At this point the directional derivative for a submanifold of Euclidean space has not yet been defined but we will come to this soon. For now this M in this definition may assumed to be a Euclidean space. The directional derivatives of the vector fields $\boldsymbol{v}$ and $\boldsymbol{w}$ are again vector fields, therefore $\boldsymbol{T}(\boldsymbol{v}, \boldsymbol{w})$ is again a directional derivative. The numbers (2.15) are indeed the components of (2.16). This can be seen by evaluating the three terms in (2.16), while letting them act on a test function $u \in C^1(\mathbb{R}^n)$. For the first term we find that

$$\nabla_{\nabla_{\boldsymbol{v}}\boldsymbol{w}}(u) = v^i(\nabla_{\boldsymbol{e}_i} w^j)(\nabla_{\boldsymbol{e}_j} u) + v^i w^j \nabla_{\nabla_{\boldsymbol{e}_i}(\boldsymbol{e}_j)}(u) = v^i(\nabla_{\boldsymbol{e}_i} w^j)(\nabla_{\boldsymbol{e}_j} u) + v^i w^j \Gamma^k_{ji}\boldsymbol{e}_k(u).$$

Similarly, for the second term we find

$$\nabla_{\nabla_{\boldsymbol{w}}\boldsymbol{v}} = w^j(\nabla_{\boldsymbol{e}_j} v^i)(\nabla_{\boldsymbol{e}_i} u) + v^i w^j \Gamma^k_{ij}\nabla_{\boldsymbol{e}_k}(u).$$

For the last term we have

$$\begin{aligned} [\nabla_{\boldsymbol{v}}, \nabla_{\boldsymbol{w}}] &= v^i\nabla_{\boldsymbol{e}_i}(w^j\nabla_{\boldsymbol{e}_j} u) - w^j\nabla_{\boldsymbol{e}_j}(v^i\nabla_{\boldsymbol{e}_i} u) \\ &= v^i(\nabla_{\boldsymbol{e}_i} w^j)(\nabla_{\boldsymbol{e}_j} u) + v^i w^j(\nabla_{\boldsymbol{e}_i}\nabla_{\boldsymbol{e}_j} u) - w^j(\nabla_{\boldsymbol{e}_j} v^i)(\nabla_{\boldsymbol{e}_i} u) - v^i w^j(\nabla_{\boldsymbol{e}_j}\nabla_{\boldsymbol{e}_i} u) \\ &= v^i(\nabla_{\boldsymbol{e}_i} w^j)(\nabla_{\boldsymbol{e}_j} u) - w^j(\nabla_{\boldsymbol{e}_j} v^i)(\nabla_{\boldsymbol{e}_i} u) + v^i w^j[\nabla_{\boldsymbol{e}_i}, \nabla_{\boldsymbol{e}_j}]u. \end{aligned}$$

*The torsion tensor is more commonly defined as $T(\boldsymbol{v}, \boldsymbol{w}) := \nabla_{\boldsymbol{v}}\boldsymbol{w} - \nabla_{\boldsymbol{w}}\boldsymbol{v} - [\boldsymbol{v}, \boldsymbol{w}]$. However, if this definition is used then the tangent vectors have been identified with the directional derivative operators in the direction of that tangent vector. We have not made this identification, because it requires a justification which is beyond the scope of this text. The interested reader should consult [2, Ch.3] or [3, Ch.2].

When we add these tree terms as in (2.16) the terms $v^i(\nabla_{\boldsymbol{e}_i} w^j)(\nabla_{\boldsymbol{e}_j} f)$ and $w^j(\nabla_{\boldsymbol{e}_j} v^i)(\nabla_{\boldsymbol{e}_i} f)$ vanish and we end up with

$$\begin{aligned}\boldsymbol{T}(\boldsymbol{v},\boldsymbol{w})u &= v^i w^j \Gamma^k_{ji} \nabla_{\boldsymbol{e}_k}(u) - v^i w^j \Gamma^k_{ij} \nabla_{\boldsymbol{e}_k}(u) - v^i w^j [\boldsymbol{e}_i, \boldsymbol{e}_j] u \\ &= v^i w^j \big(c^k_{ji} - \Gamma^k_{ij} + \Gamma^k_{ji}\big) \nabla_{\boldsymbol{e}_k} u.\end{aligned}$$

Thus, we find that the components of (2.16) are indeed given by (2.15). So far, we have seen that the directional derivative on an n-dimensional Euclidean space is a function that takes two vector fields $\boldsymbol{v} \in T\mathbb{R}^n,\ \boldsymbol{w} \in T\mathbb{R}^n_{C^1}$ and produces a third vector field $\nabla_{\boldsymbol{v}}\boldsymbol{w} \in T\mathbb{R}^n$ that has certain properties. We have seen that the directional derivative is linear in the direction of differentiation and also in the argument to be differentiated. Furthermore we have seen that it satisfies a product rule, concerning the product of a scalar field and a vector field, i.e. (2.11), and concerning the inner product, i.e. (2.12). Lastly we have seen that the directional derivative is such that the torsion tensor equals the zero tensor, which also is equivalent to the fact that the Hessian tensor is symmetric.

It turns out that the directional derivative can be generalized to non-Euclidean spaces, while still having the aforementioned properties. This generalization is called the *Levi-Civita connection* and it is defined as follows. (See [1, Ch.3].)

Definition 2.4.6. Let M be a submanifold of a Euclidean space and let $\boldsymbol{g}$ be the metric induced by the ordinary Euclidean inner product on the ambient Euclidean space. A Levi-Civita connection on M is a mapping $\nabla : TM \times TM_{C^1} \to TM$ that takes two vector fields $\boldsymbol{v} \in T\mathbb{R}^n,\ \boldsymbol{w} \in T\mathbb{R}^n_{C^1}$ and produces a third vector field $\nabla_{\boldsymbol{v}}\boldsymbol{w} \in TM$ that has the following properties:

(i) $\forall u_1, u_2 \in C^1(M),\ \forall \boldsymbol{v}_1, \boldsymbol{v}_2 \in TM,\ \forall \boldsymbol{w} \in TM_{C^1}:\quad \nabla_{u_1\boldsymbol{v}_1 + u_2\boldsymbol{v}_2}\boldsymbol{w} = u_1 \nabla_{\boldsymbol{v}_1}\boldsymbol{w} + u_2 \nabla_{\boldsymbol{v}_2}\boldsymbol{w},$

(ii) $\forall \lambda_1, \lambda_2 \in \mathbb{R},\ \forall \boldsymbol{v} \in TM,\ \forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in TM_{C^1}:\quad \nabla_{\boldsymbol{v}}(\lambda_1\boldsymbol{w}_1 + \lambda_2\boldsymbol{w}_2) = \lambda_1 \nabla_{\boldsymbol{v}}\boldsymbol{w}_1 + \lambda_2 \nabla_{\boldsymbol{v}}\boldsymbol{w}_2,$

(iii) $\forall u \in C^1(M),\ \forall \boldsymbol{v} \in TM,\ \forall \boldsymbol{w} \in TM_{C^1}:\quad \nabla_{\boldsymbol{v}}(u\boldsymbol{w}) = u \nabla_{\boldsymbol{v}}\boldsymbol{w} + (\nabla_{\boldsymbol{v}} u)\boldsymbol{w},$

(iv) $\forall \boldsymbol{v} \in TM,\ \forall \boldsymbol{w}, \boldsymbol{z} \in TM_{C^1}:\quad \nabla_{\boldsymbol{z}}(\boldsymbol{g}(\boldsymbol{v},\boldsymbol{w})) = \boldsymbol{g}(\nabla_{\boldsymbol{z}}\boldsymbol{v}, \boldsymbol{w}) + \boldsymbol{g}(\boldsymbol{v}, \nabla_{\boldsymbol{z}}\boldsymbol{w}),$

(v) $\forall \boldsymbol{v} \in TM,\ \forall \boldsymbol{w} \in TM_{C^1}:\quad \boldsymbol{T}(\boldsymbol{v},\boldsymbol{w}) = 0.$

We will often use the name *Covariant derivative* to refer to the Levi-Civita connection.

We see that the Covariant derivative has all the properties that the directional derivative in Euclidean space has. When we take M in Definition 2.4.6 to be Euclidean space, this definition is just the definition of the directional derivative. So, the directional derivative is the covariant derivative for Euclidean space. For a local coordinate system on M with basis $\{\boldsymbol{e}_i\}$ for TM we again define the Christoffel symbols and commutation symbols by

$$\begin{aligned}\Gamma^k_{ij} &:= \langle \hat{\boldsymbol{e}}^k, \nabla_{\boldsymbol{e}_j}\boldsymbol{e}_i \rangle, \\ c^k_{ij} &:= [\nabla_{\boldsymbol{e}_i}, \nabla_{\boldsymbol{e}_j}].\end{aligned}$$

Notice that with these definitions the components of $\nabla_{\boldsymbol{v}}\boldsymbol{w}$ are still given by (2.10). From this expression it is clear that the Levi-Civita connection is uniquely determined if the Christoffel symbols are given. It turns out that the Christoffel symbols for the Levi-Civita connection can be uniquely determined from the metric tensor $\boldsymbol{g}$. This implies that there is a unique Levi-Civita connection for a submanifold of a Euclidean space. This we will prove in the following theorem.

Theorem 2.4.7. *The Levi-Civita connection is uniquely given by the Christoffel symbols*

$$\Gamma^k_{ij} = \frac{1}{2} g^{kl} \left(\nabla_{\boldsymbol{e}_i} g_{lj} + \nabla_{\boldsymbol{e}_j} g_{li} - \nabla_{\boldsymbol{e}_l} g_{ij} + c_{ilj} + c_{jli} - c_{lij} \right), \tag{2.17}$$

where $c_{lij} := g_{lk} c^k_{ij}$. In the case of a holonomic bases these expressions simplify to

$$\Gamma^k_{ij} = \frac{1}{2} g^{kl} (\partial_i g_{lj} + \partial_j g_{li} - \partial_l g_{ij}). \tag{2.18}$$

Proof. By the first property of Definition 2.4.6 it follows that for arbitrary $\boldsymbol{v}, \boldsymbol{w} \in TM$,

$$D_l(g_{ij}v^i w^j) = g_{ij}(D_l v^i)w^j + g_{ij}v^i(D_l w^j), \tag{2.19}$$

while by the product rule

$$D_l(g_{ij}v^i w^j) = (D_l g_{ij})v^i w^j + g_{ij}(D_l v^i)w^j + g_{ij}v^i(D_l w^j). \tag{2.20}$$

The fact that the product rule applies is shown in [7, ex.8.9]. Equations (2.19) and (2.20) together imply that $D_l g_{ij} = 0$. The components g_{ij} are the components of a tensor and therefore also

$$D_l g_{ij} = \nabla_{\boldsymbol{e}_l} g_{ij} - \Gamma^m_{il} g_{mj} - \Gamma^m_{jl} g_{im},$$

hence we have $\nabla_{\boldsymbol{e}_l} g_{ij} = \Gamma^m_{il} g_{mj} + \Gamma^m_{jl} g_{im}$. If we cyclically permute the free indices in this expression and subtract the resulting two expressions from the given expression we find

$$\nabla_{\boldsymbol{e}_l} g_{ij} - \nabla_{\boldsymbol{e}_i} g_{jl} - \nabla_{\boldsymbol{e}_j} g_{li} = \left(\Gamma^m_{il} - \Gamma^m_{li}\right) g_{mj} + \left(\Gamma^m_{jl} - \Gamma^m_{lj}\right) g_{mi} - \left(\Gamma^m_{ji} + \Gamma^m_{ij}\right) g_{ml}.$$

The connection is torsion free, therefore the components of the torsion tensor $T^k_{ij} = \Gamma^k_{ji} - \Gamma^k_{ij} - c^k_{ij}$ equal zero. Using this we find that

$$\nabla_{\boldsymbol{e}_l} g_{ij} - \nabla_{\boldsymbol{e}_i} g_{jl} - \nabla_{\boldsymbol{e}_j} g_{li} = c^m_{li} g_{mj} + c^m_{lj} g_{mi} - \left(2\Gamma^m_{ij} + c^m_{ij}\right) g_{ml}.$$

Introducing the short hand notation $c_{jli} := c^m_{li} g_{mj}$, where the first of the three indices on the left hand side corresponds to the upper index on the right hand side, it follows that

$$\Gamma^k_{ij} = \frac{1}{2} g^{kl} \left(\nabla_{\boldsymbol{e}_i} g_{lj} + \nabla_{\boldsymbol{e}_j} g_{li} - \nabla_{\boldsymbol{e}_l} g_{ij} + c_{ilj} + c_{jli} - c_{lij}\right).$$

In the case of a holonomic basis we have $\nabla_{\boldsymbol{e}_l} = \partial_l$, all the commutation symbols are zero and we end up with the simplified expression

$$\Gamma^k_{ij} = \frac{1}{2} g^{kl} (\partial_i g_{lj} + \partial_j g_{li} - \partial_l g_{ij}).$$

□

We will now put the equations derived in Theorem 2.4.7 to use. We will use it to determine the Christoffel symbols for polar coordinates in $\mathbb{R}^2$. We will first consider polar coordinates with the corresponding coordinate basis and then we will consider polar coordinates with a different, anholonomic, basis. Subsequently, we will derive the coordinate and basis dependent expression for the Hessian tensor for these two cases. The Hessian tensor in polar coordinates will be of importance in Chapter 3.

Example 2.4.8. We consider $\mathbb{R}^2$, with the ordinary inner product. If we use Cartesian coordinates then the metric is given by

$$(g_{ij}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{2.21}$$

By Equation 2.18 it follows that all the Christoffel symbols are zero in this case, because the basis vectors do not depend on position.

Let us now assume a polar coordinate system on $\mathbb{R}^2$ as described in Example 2.1.6. If we write $\overline{x}^1 = r, \overline{x}^2 = \theta$ and $x^1 = x, x^2 = y$, then the new holonomic basis vectors are given by

$$\overline{\boldsymbol{e}}_i = \frac{\partial x^j}{\partial \overline{x}^i} \boldsymbol{e}_j. \tag{2.22}$$

A simple calculation shows that the Jacobian is given by

$$\left(\frac{\partial x^j}{\partial \overline{x}^i}\right) = \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix}. \tag{2.23}$$

Using this we find that

$$\begin{aligned} \boldsymbol{e}_r &= \overline{\boldsymbol{e}}_1 = \cos(\theta)\boldsymbol{e}_1 + \sin(\theta)\boldsymbol{e}_2 = \cos(\theta)\boldsymbol{e}_x + \sin(\theta)\boldsymbol{e}_y, \\ \boldsymbol{e}_\theta &= \overline{\boldsymbol{e}}_2 = -r\sin(\theta)\boldsymbol{e}_1 + r\cos(\theta)\boldsymbol{e}_2 = -r\sin(\theta)\boldsymbol{e}_x + r\cos(\theta)\boldsymbol{e}_y. \end{aligned}$$

The metric transforms according to the tensor transformation laws and thereby we find

$$\overline{g}_{ij} = \frac{\partial x^r}{\partial \overline{x}^i}\frac{\partial x^s}{\partial \overline{x}^j} g_{rs}.$$

The matrix representation of (g_{rs}) is just the identity therefore we find

$$\begin{aligned} (\overline{g}_{ij}) &= \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix}^T \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}. \end{aligned}$$

We can now use equation (2.18) to calculate the Christoffel symbols. Note that the only nonzero derivative of the metric is $\partial_r \overline{g}_{\theta\theta} = \overline{\partial}_1 \overline{g}_{22} = 2r$. From this and equation (2.18) it follows that we have only three nonzero Christoffel symbols, namely

$$\overline{\Gamma}^1_{22} = \Gamma^r_{\theta\theta} = -r, \quad \overline{\Gamma}^2_{12} = \Gamma^\theta_{r\theta} = \frac{1}{r}, \quad \overline{\Gamma}^2_{21} = \Gamma^\theta_{\theta r} = \frac{1}{r}.$$

Now that we know the Christoffel symbols, we are able to determine the Hessian tensor. We work in a holonomic basis and hence the components of the Hessian tensor are given by

$$H_{ij} = \partial_j(\partial_i u) - \Gamma^k_{ij}\partial_k u.$$

Substituting the earlier derived Christoffel symbols in this equation we obtain

$$(H_{ij}) = \begin{pmatrix} u_{rr} & u_{r\theta} - u_\theta/r \\ u_{r\theta} - u_\theta/r & u_{\theta\theta} + ru_r \end{pmatrix}, \tag{2.24}$$

where the subscripts r and θ denote partial derivatives.

The metric in the coordinate basis indicates that this basis is not an orthonormal basis. The length of the basis vector $\boldsymbol{e}_\theta$ is equal to r and therefore depends on the location on $\mathbb{R}^2$. Often this basis vector is rescaled to obtain an orthonormal basis. This we will consider in the next example.

Example 2.4.9. Let us again assume the Euclidean space which was considered in Example 2.4.8. Let us define the basis vectors $\overline{\boldsymbol{e}}_r = \boldsymbol{e}_r$ and $\overline{\boldsymbol{e}}_\theta = r^{-1}\boldsymbol{e}_\theta$. From the preceding example it is clear that these vectors form an orthonormal basis. The metric, we denote by $\boldsymbol{g}$ and is given by

$$(g_{ij}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

which again indicates the orthonormality. This orthonormal basis is an anholonomic basis, thus to calculate the Christoffel symbols we need to use equation (2.17). The metric is constant over $\mathbb{R}^2$, hence only the commutation symbols contribute to the Christoffel symbols. In Example 2.4.3 we determined these commutation symbols. We found that there were only two nonzero commutation symbols: $c^\theta_{r\theta} = -c^\theta_{\theta r} = -r^{-1}$. Using this to evaluate equation (2.17) we find that in this basis we have only two nonzero Christoffel symbols, $\overline{\Gamma}^r_{\theta\theta} = -r^{-1}$ and $\overline{\Gamma}^\theta_{r\theta} = r^{-1}$. Note that in this anholonomic case the Christoffel symbols are not symmetric in the lower indices. This results in a different Hessian tensor than the one in holonomic polar coordinates. We now have

$$(H_{ij}) = \begin{pmatrix} u_{rr} & (ru_{r\theta} - u_\theta)/r^2 \\ (ru_{r\theta} - u_\theta)/r^2 & (u_{\theta\theta} + ru_r)/r^2 \end{pmatrix}. \tag{2.25}$$

Let us briefly summarize, what we have done in this chapter. We have started out in Section 2.1 by formally defining the space in which are optical system is situated, i.e. a Euclidean space. Then we went on to argue that, in general, the reflector surface is of our optical system is curved and can therefore not be described by a Euclidean space. In order to find a satisfactory way to describe the curved surface of the reflector, we introduced the notion of a submanifold of Euclidean space. Furthermore we defined coordinate systems on such submanifolds of Euclidean space and we showed that the reflector surface can be described by the Monge parametrization. This parametrization is the one we will indeed use in Chapter 3 to describe the reflector surface.

In Section 2.2, we discussed tangent vectors to a point on an n-dimensional submanifold of $\mathbb{R}^N$. We showed that the tangent vectors form a n-dimensional subspace of $\mathbb{R}^N$ and that a natural basis for this subspace is given by the tangent vectors to the coordinate lines. This basis is called the coordinate basis.

This vector space we took as the basis of our discussion in Section 2.3. In this section, we showed that we can add a lot of extra structure to such a vector space by considering the dual space to this vector space. We introduced the tensors, which are multi-linear maps from arbitrary Cartesian products of the vector space and its dual space to the real numbers. We showed that vectors and dual vectors are just special cases of tensors and we showed how all of these transform under a change of basis. Furthermore, we showed that when the vector space has an inner product, there exists a linear bijection between the vectors and covectors. We saw that when we have a vector-covector pair under this bijection, the components of the covector are obtained from the components of the vector by lowering the index with the metric. In a similar way the components of the vector are obtained from the components of the covector. This relationship between such a vector-covector pair is expressed by the musical isomorphisms $\sharp$ and $\flat$. Lastly, we introduced in this section some convenient tensor that will be of use later on, like the Levi-Civita tensor and the generalized Kronecker tensor.

We ended this Chapter with Section 2.4, which was on differentiation of scalars, vectors and more generally tensors. We discussed the properties of the directional derivative. Moreover, we further clarified the difference between holonomic bases, i.e. a basis such that the Lie derivative of pairs of basis vectors vanishes, and more generally anholonomic bases. We also introduced the Hessian tensor and showed that it is symmetric if and only if the components of the torsion tensor vanish. We ended the discussion by given a definition of the covariant derivative which generalises the directional derivative to non-Euclidean spaces. This definition, nicely summarizes the properties of the directional derivative and will therefore be convenient to rely on in Chapter 3, the chapter to which we will now proceed.

Chapter 3

Monge-Ampère Equation

In this chapter we will derive the energy conservation equation for the optical system described in Chapter 1, which will equate the energy output of the source with the energy output of the reflector. This equation will turn out to be of Monge-Ampère type. We will show that the equation obtained is independent of the choice of coordinate system. Subsequently, we will show some coordinate specific expressions for the Monge-Ampère equation and show that for Cartesian coordinates we recover the Monge-Ampère equation as it was given in [5]. Lastly, we will formulate the boundary value problem corresponding to the Monge-Ampère equation. This is the boundary value problem that we will try to solve numerically in Chapter 4 and onwards.

3.1 Source and reflector surface

The optical system that we will consider consists of nothing more than a light source and a reflector surface. Let $\mathcal{E}$ be a convex open subset of the two-dimensional Euclidean plane $\mathbb{R}^2$, which is a plane in the three-dimensional Euclidean space $\mathbb{R}^3$. We will equip $\mathbb{R}^3$ with the usual inner product $(\cdot|\cdot) : \mathbb{R}^3 \to \mathbb{R}$. The convex set $\mathcal{E}$ represents the source. Let us describe $\mathcal{E}$ with an orthogonal coordinate system with coordinates x^1 and x^2. Furthermore, we assume that for each tangent space $T_x\mathcal{E}$ at $x \in \mathcal{E}$, we have an orthogonal but not necessarily holonomic, basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$. The source $\mathcal{E}$ is a subset of the Euclidean plane and therefore we can make the identification $T_x\mathcal{E} = \mathcal{E}$. At each point $x \in \mathcal{E}$ the emittance [lm/m^2] of the source is given by a strictly positive function $E : \mathcal{E} \to \mathbb{R}_{>0}$. We will extend the basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$ to a basis of $\mathbb{R}^3$ by defining*

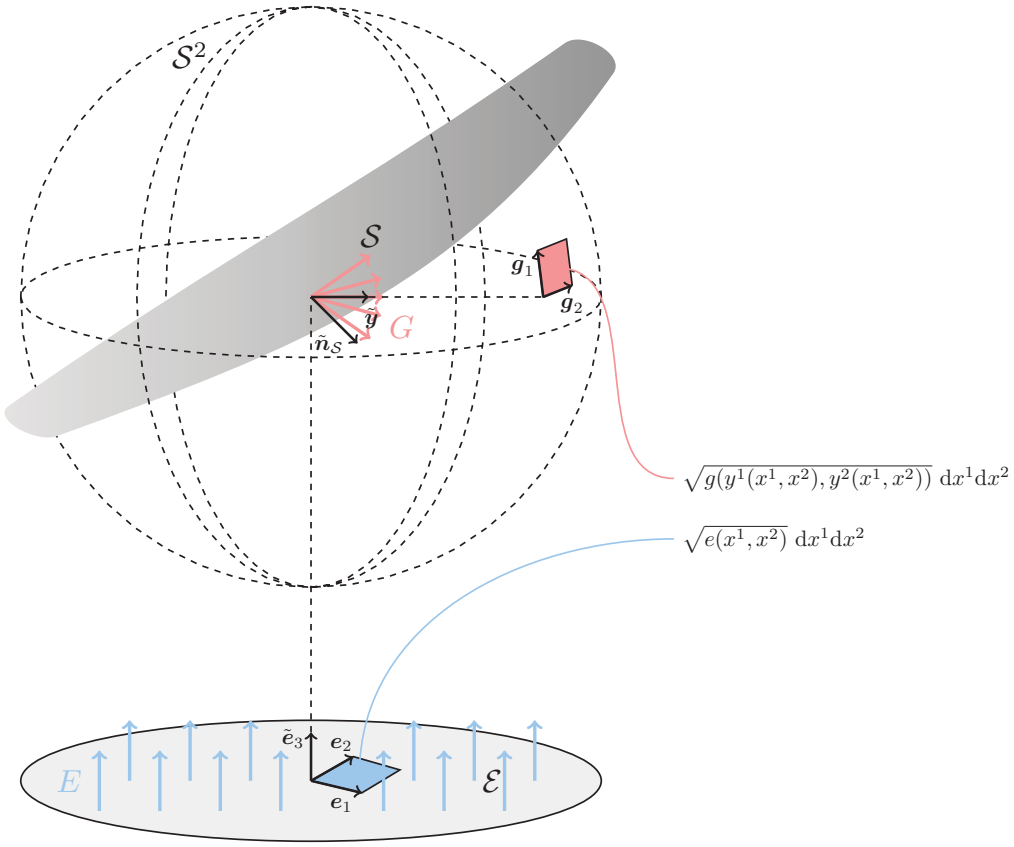
$$\tilde{\boldsymbol{e}}_3 := \frac{\boldsymbol{e}_1 \times \boldsymbol{e}_2}{\|\boldsymbol{e}_1 \times \boldsymbol{e}_2\|},$$

where the norm is defined by $\|\boldsymbol{v}\| := \sqrt{(\boldsymbol{v}|\boldsymbol{v})}$. Thus the third basis vector is perpendicular to the plane of $\mathcal{E}$ and has unit length. The set $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \tilde{\boldsymbol{e}}_3\}$ is an orthogonal basis for $\mathbb{R}^3$. We will denote the components of the metric on $\mathcal{E}$ by $e_{ij} := (\boldsymbol{e}_i|\boldsymbol{e}_j)$, for $i, j = 1, 2$. To clarify the notation, if we speak about the basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$ or the basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \tilde{\boldsymbol{e}}_3\}$ we will from now on use latin indices if we sum over $1, 2$ and use greek indices if we sum over $1, 2, 3$. The matrix representations $(e_{\mu\nu})$ and $(e^{\mu\nu})$ of the metric for $\mathbb{R}^3$ and its inverse are thus given by

$$(e_{\mu\nu}) = \begin{pmatrix} e_{11} & 0 & 0 \\ 0 & e_{22} & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad (e^{\mu\nu}) = \begin{pmatrix} 1/e_{11} & 0 & 0 \\ 0 & 1/e_{22} & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{3.1}$$

We will now consider a function $u : \mathcal{E} \to \mathbb{R}_{>0}$. We assume this function to be twice continuously

*We already use ˆ to indicate covectors, therefore we will use ˜ to indicate that a vector is of unit length.



differentiable and also strictly convex. The set $\mathcal{S} := \{(x, u(x)) \in \mathbb{R}^2 \times \mathbb{R} \mid x \in \mathcal{E}\}^*$ describes a surface in $\mathbb{R}^3$. We demand u to be a positive function in order to ensure that $\mathcal{S}$ lies above $\mathcal{E}$. When we say that $\mathcal{S}$ lies above $\mathcal{E}$ we mean that to get from $x \in \mathcal{E}$ to $(x, u(x)) \in \mathcal{S}$ one has to travel in the direction $\tilde{\boldsymbol{e}}_3$. In Example 2.1.7, we saw that a convenient parametrization of this surface is the Monge parametrization. We now use the basis for the tangent space $T_x\mathcal{E}$ to define a basis for the tangent space $T_{(x,u(x))}\mathcal{S}$:

$$\begin{aligned}\boldsymbol{s}_1 &:= \boldsymbol{e}_1 + (\nabla_{\boldsymbol{e}_1} u)\tilde{\boldsymbol{e}}_3,\\ \boldsymbol{s}_2 &:= \boldsymbol{e}_2 + (\nabla_{\boldsymbol{e}_2} u)\tilde{\boldsymbol{e}}_3.\end{aligned}$$

Taking the cross product of these vectors we get the normal vector to $\mathcal{S}$, i.e. $\boldsymbol{n}_{\mathcal{S}} := \boldsymbol{s}_2 \times \boldsymbol{s}_1$. Notice that we have defined $\boldsymbol{n}_{\mathcal{S}}$ such that $(\boldsymbol{n}_{\mathcal{S}} \mid \tilde{\boldsymbol{e}}_3) < 0$. This means that the vector $\boldsymbol{n}_{\mathcal{S}}$ points downwards from its point on $\mathcal{S}$. This choice is most natural when considering light rays emitted by $\mathcal{E}$ and reflected by $\mathcal{S}$. We can rewrite $\boldsymbol{n}_{\mathcal{S}}$ in terms of the basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \tilde{\boldsymbol{e}}_3\}$ as follows,

$$\boldsymbol{n}_{\mathcal{S}} = \boldsymbol{s}_2 \times \boldsymbol{s}_1 = \epsilon_{\alpha\beta\gamma}(\boldsymbol{s}_2)^\alpha(\boldsymbol{s}_1)^\beta e^{\gamma\sigma}\boldsymbol{e}_\sigma = \sqrt{e}\left[(\nabla_{\boldsymbol{e}_1} u)e^{11}\boldsymbol{e}_1 + (\nabla_{\boldsymbol{e}_2} u)e^{22}\boldsymbol{e}_2 - \tilde{\boldsymbol{e}}_3\right]. \tag{3.2}$$

The gradient is given by $\mathrm{d}u = (\nabla_{\boldsymbol{e}_i} u)\hat{\boldsymbol{e}}^i$ and the vector associated with it is given by $\nabla u = \flat \mathrm{d}u = (\nabla_{\boldsymbol{e}_i} u)e^{ij}\boldsymbol{e}_j$. Note that for an orthogonal basis we have $\nabla u = (\nabla_{\boldsymbol{e}_i} u)e^{ij}\boldsymbol{e}_j = (\nabla_{\boldsymbol{e}_1} u)e^{11}\boldsymbol{e}_1 + (\nabla_{\boldsymbol{e}_2} u)e^{22}\boldsymbol{e}_2$ and hence

$$\boldsymbol{n}_{\mathcal{S}} = \sqrt{e}(\nabla u - \tilde{\boldsymbol{e}}_3), \tag{3.3}$$

*The 2-tuple is not a vector it only serves to denote the point of $\mathcal{S}$ that one arrives at after covering a distance $u(x)$ in the direction of $\tilde{\boldsymbol{e}}_3$ from the point x on $\mathcal{E}$. The surface $\mathcal{S}$ is an object independent of the choice of coordinate system on $\mathcal{E}$ and we wish to define it in such a way. Moreover it allows for a coordinate independent way of denoting the tangent space to $\mathcal{S}$ in the point $(x, u(x))$ as $T_{(x,u(x))}\mathcal{S}$ instead of denoting it as $T_{x^i\boldsymbol{e}_i + u(x^1,x^2)\tilde{\boldsymbol{e}}_3}\mathcal{S}$.

where e is the determinant of the metric (e_{ij}). The norm of the normal vector is given by

$$\begin{aligned}\|\boldsymbol{n}_{\mathcal{S}}\| &= \left[e_{11}e_{22}\left(\left((\nabla_{\boldsymbol{e}_1}u)e^{11}\right)^2(\boldsymbol{e}_1|\boldsymbol{e}_1) + \left((\nabla_{\boldsymbol{e}_2}u)e^{22}\right)^2(\boldsymbol{e}_2|\boldsymbol{e}_2) + (\tilde{\boldsymbol{e}}_3|\tilde{\boldsymbol{e}}_3)\right)\right]^{\frac{1}{2}} \\ &= \left[(\nabla_{\boldsymbol{e}_1}u)^2e_{22} + (\nabla_{\boldsymbol{e}_2}u)^2e_{11} + e_{11}e_{22}\right]^{\frac{1}{2}}.\end{aligned} \tag{3.4}$$

We denote the components of the metric on $\mathcal{S}$ by $s_{ij} := (\boldsymbol{s}_i|\boldsymbol{s}_j)$ and the determinant of (s_{ij}) by s. We can relate the determinant of the metric of a surface to the norm of the normal vector to that surface. This is expressed by the following lemma.

Lemma 3.1.1. *Assume we have a basis $\{\boldsymbol{e}_\alpha\}$ and corresponding metric $e_{\mu\nu}$ for $\mathbb{R}^3$, a surface $\mathcal{W}$ in $\mathbb{R}^3$ with basis $\{\boldsymbol{w}_1, \boldsymbol{w}_2\}$ for the tangent space to $\mathcal{W}$ and furthermore a corresponding metric w_{ij} on $\mathcal{W}$. Then it holds that $\|\boldsymbol{n}_{\mathcal{S}}\|^2 = w$, where w is the determinant of the metric w_{ij} and $\boldsymbol{n} = \boldsymbol{w}_1 \times \boldsymbol{w}_2$, the normal to $\mathcal{S}$. For the cross product the vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are interpreted as vectors in $\mathbb{R}^3$.*

Proof. By definition of the cross product

$$\begin{aligned}\|\boldsymbol{n}_{\mathcal{S}}\|^2 &= (\boldsymbol{w}_1 \times \boldsymbol{w}_2|\boldsymbol{w}_1 \times \boldsymbol{w}_2) \\ &= \left(\epsilon_{\alpha\beta\gamma}(\boldsymbol{w}_1)^\alpha(\boldsymbol{w}_2)^\beta e^{\gamma\mu}\boldsymbol{e}_\mu|\epsilon_{\rho\sigma\tau}(\boldsymbol{w}_1)^\rho(\boldsymbol{w}_2)^\sigma e^{\tau\nu}\boldsymbol{e}_\nu\right) \\ &= \epsilon_{\alpha\beta\gamma}\epsilon_{\rho\sigma\tau}(\boldsymbol{w}_1)^\alpha(\boldsymbol{w}_2)^\beta(\boldsymbol{w}_1)^\rho(\boldsymbol{w}_2)^\sigma e^{\gamma\mu}e^{\tau\nu}e_{\mu\nu} \\ &= \epsilon_{\alpha\beta\gamma}\epsilon_{\rho\sigma\tau}(\boldsymbol{w}_1)^\alpha(\boldsymbol{w}_2)^\beta(\boldsymbol{w}_1)^\rho(\boldsymbol{w}_2)^\sigma e^{\gamma\mu}\delta^\tau_\mu \\ &= \epsilon_{\alpha\beta\gamma}\epsilon_{\rho\sigma\mu}(\boldsymbol{w}_1)^\alpha(\boldsymbol{w}_2)^\beta(\boldsymbol{w}_1)^\rho(\boldsymbol{w}_2)^\sigma e^{\gamma\mu} \\ &= \epsilon_{\alpha\beta\gamma}\epsilon^{\nu\tau\gamma}(\boldsymbol{w}_1)^\alpha(\boldsymbol{w}_2)^\beta(\boldsymbol{w}_1)^\rho(\boldsymbol{w}_2)^\sigma e_{\rho\nu}e_{\sigma\tau} \\ &= \delta^{\nu\tau\gamma}_{\alpha\beta\gamma}(\boldsymbol{w}_1)^\alpha(\boldsymbol{w}_2)^\beta(\boldsymbol{w}_1)^\rho(\boldsymbol{w}_2)^\sigma e_{\rho\nu}e_{\sigma\tau} \\ &= \delta^{\nu\tau}_{\alpha\beta}(\boldsymbol{w}_1)^\alpha(\boldsymbol{w}_2)^\beta(\boldsymbol{w}_1)^\rho(\boldsymbol{w}_2)^\sigma e_{\rho\nu}e_{\sigma\tau} \\ &= (\boldsymbol{w}_1)^\alpha(\boldsymbol{w}_2)^\beta(\boldsymbol{w}_1)^\rho(\boldsymbol{w}_2)^\sigma e_{\rho\alpha}e_{\sigma\beta} - (\boldsymbol{w}_1)^\alpha(\boldsymbol{w}_2)^\beta(\boldsymbol{w}_1)^\rho(\boldsymbol{w}_2)^\sigma e_{\sigma\alpha}e_{\rho\beta}.\end{aligned}$$

The inner products of the basis vectors are given by $w_{ij} = (\boldsymbol{w}_i)^\alpha(\boldsymbol{w}_j)^\beta e_{\alpha\beta}$. This implies that the determinant of w_{ij} is given by

$$\begin{aligned}w &= \left((\boldsymbol{w}_1)^\alpha(\boldsymbol{w}_1)^\beta e_{\alpha\beta}\right)\left((\boldsymbol{w}_2)^\rho(\boldsymbol{w}_2)^\sigma e_{\rho\sigma}\right) - \left((\boldsymbol{w}_1)^\alpha(\boldsymbol{w}_2)^\beta e_{\alpha\beta}\right)\left((\boldsymbol{w}_2)^\rho(\boldsymbol{w}_1)^\sigma e_{\rho\sigma}\right) \\ &= (\boldsymbol{w}_1)^\alpha(\boldsymbol{w}_2)^\beta(\boldsymbol{w}_1)^\rho(\boldsymbol{w}_2)^\sigma e_{\rho\alpha}e_{\sigma\beta} - (\boldsymbol{w}_1)^\alpha(\boldsymbol{w}_2)^\beta(\boldsymbol{w}_1)^\rho(\boldsymbol{w}_2)^\sigma e_{\sigma\alpha}e_{\rho\beta}.\end{aligned}$$

We see that we indeed have $\|\boldsymbol{n}\|^2 = w$. □

This lemma will be of much use later when we need to calculate the determinant of the metric for a different surface. We can also use this lemma to obtain a convenient expression for s, the metric on $\mathcal{S}$, in terms of e, the metric on $\mathcal{E}$ and the gradient of the function u.

Lemma 3.1.2. *The determinant of the metric s_{ij} on $\mathcal{S}$ can be expressed in terms of $\|\nabla u\|^2 = (\nabla_{\boldsymbol{e}_i}u)(\nabla_{\boldsymbol{e}_j}u)e^{ij}$ and e, the determinant of the metric on $\mathcal{E}$, as*

$$s = e(\|\nabla u\|^2 + 1).$$

Proof. By Lemma 3.1.1 it follows that $s = \|\boldsymbol{n}_{\mathcal{S}}\|^2$. We earlier obtained the expression (3.4) for $\|\boldsymbol{n}_{\mathcal{S}}\|$, hence we have

$$s = (\nabla_{\boldsymbol{e}_1}u)^2e_{22} + (\nabla_{\boldsymbol{e}_2}u)^2e_{11} + e_{11}e_{22}.$$

From equation (3.1) it is clear that $e = e_{11}e_{22}$ and furthermore that $e_{22} = e\ e^{11}$ and $e_{11} = e\ e^{22}$. This implies that

$$\begin{aligned}s &= e\left((\nabla_{\boldsymbol{e}_1}u)^2e^{11} + (\nabla_{\boldsymbol{e}_2}u)^2e^{22} + 1\right) \\ &= e\left((\nabla_{\boldsymbol{e}_i}u)(\nabla_{\boldsymbol{e}_j}u)e^{ij} + 1\right) \\ &= e\left(\|\nabla u\|^2 + 1\right).\end{aligned}$$

□

3.2 Law of reflection

Before we proceed we make some additional assumptions on our optical system. First we assume that the light rays emitted by the source $\mathcal{E}$ all leave in the direction $\tilde{\boldsymbol{e}}_3$. Thus the light from $\mathcal{E}$ constitutes a parallel bundle and hits the reflector surface $\mathcal{S}$ with incoming direction $\tilde{\boldsymbol{e}}_3$. The very familiar law of reflection tells us in what direction a light ray proceeds after reflection. The law of reflection in vector form, which can be found in [12], is given by

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{e}}_3 - 2(\tilde{\boldsymbol{e}}_3|\tilde{\boldsymbol{n}}_{\mathcal{S}})\tilde{\boldsymbol{n}}_{\mathcal{S}}. \tag{3.5}$$

We assume that the dimensions of the reflector surface are small enough that we can neglect it when compared to the distance between the reflector and projection screen. We will assume that the reflected rays all originate from one point. This is called the *far field approximation* and greatly reduces the difficulty of the problem, while the error introduced by the approximation is very small. This will become clear once we test our method for some examples.

By Equation (3.5), every $x \in \mathcal{E}$ gets mapped to a direction of reflection $\tilde{\boldsymbol{y}}(x)$, hence we have a mapping $x \mapsto \tilde{\boldsymbol{y}}(x)$, from $\mathcal{E}$ to the unit sphere $\mathcal{S}^2$. We denote the image under this mapping by $\mathcal{G}$, i.e. $\mathcal{G} := \tilde{\boldsymbol{y}}(\mathcal{E})$. The following lemma shows that a point $x \in \mathcal{E}$ corresponds to a unique direction of reflection $\tilde{\boldsymbol{y}}(x)$.

Lemma 3.2.1. *The map given by $x \mapsto \tilde{\boldsymbol{y}}(x)$, where*

$$\tilde{\boldsymbol{y}}(x) = \tilde{\boldsymbol{e}}_3 - 2(\tilde{\boldsymbol{e}}_3|\tilde{\boldsymbol{n}}_{\mathcal{S}}(x))\tilde{\boldsymbol{n}}_{\mathcal{S}}(x),$$

is a bijection from $\mathcal{E}$ to $\mathcal{G}$.

Proof. By definition of $\mathcal{G}$ the map is surjective. Rest us to show that the map is injective. We will argue by contradiction. We take $x_1, x_2 \in \mathcal{E}$ such that $\tilde{\boldsymbol{y}}(x_1) = \tilde{\boldsymbol{y}}(x_2)$ and $x_1 \neq x_2$ and show that this leads to a contradiction. From $\tilde{\boldsymbol{y}}(x_1) = \tilde{\boldsymbol{y}}(x_2)$ it follows that

$$(\tilde{\boldsymbol{e}}_3|\tilde{\boldsymbol{n}}_{\mathcal{S}}(x_1))\tilde{\boldsymbol{n}}_{\mathcal{S}}(x_1) = (\tilde{\boldsymbol{e}}_3|\tilde{\boldsymbol{n}}_{\mathcal{S}}(x_2))\tilde{\boldsymbol{n}}_{\mathcal{S}}(x_2),$$

which implies that $\tilde{\boldsymbol{n}}_{\mathcal{S}}(x_1)$ and $\tilde{\boldsymbol{n}}_{\mathcal{S}}(x_2)$ are parallel. However, because both have unit length this implies that $\tilde{\boldsymbol{n}}_{\mathcal{S}}(x_1) = \pm\tilde{\boldsymbol{n}}_{\mathcal{S}}(x_2)$. From the fact that we have

$$(\tilde{\boldsymbol{e}}_3|\tilde{\boldsymbol{n}}_{\mathcal{S}}(x)) = \left(\tilde{\boldsymbol{e}}_3\middle|\sqrt{\frac{e(x)}{s(x)}}(\nabla u - \tilde{\boldsymbol{e}}_3)\right) = -\sqrt{\frac{e(x)}{s(x)}} < 0, \tag{$\star$}$$

it follows that we cannot have $(\tilde{\boldsymbol{e}}_3|\tilde{\boldsymbol{n}}_{\mathcal{S}}(x_1)) = -(\tilde{\boldsymbol{e}}_3|\tilde{\boldsymbol{n}}_{\mathcal{S}}(x_2))$, hence we must have $\tilde{\boldsymbol{n}}_{\mathcal{S}}(x_1) = \tilde{\boldsymbol{n}}_{\mathcal{S}}(x_2)$. Moreover, it follows from $(\star)$ that $(\tilde{\boldsymbol{e}}_3|\tilde{\boldsymbol{n}}_{\mathcal{S}}(x)) < 0$ for all $x \in \mathcal{E}$. Note that the square root in $(\star)$ is well defined by the expression for s given by Lemma 3.1.2 and the fact that $u \in C^2(\mathcal{E})$. Thus, we must have

$$\sqrt{\frac{e(x_1)}{s(x_1)}} = \sqrt{\frac{e(x_2)}{s(x_2)}}.$$

This holds for any coordinate system, hence also for Cartesian coordinates for which we have $\sqrt{e(x_1)} = \sqrt{e(x_2)}$. This implies that in Cartesian coordinates we must have $\sqrt{s(x_1)} = \sqrt{s(x_2)}$ and moreover we have $s(x_1) = \|\boldsymbol{n}_{\mathcal{S}}(x_1)\|^2 = \|\boldsymbol{n}_{\mathcal{S}}(x_2)\|^2 = s(x_2)$. We have established that $\boldsymbol{n}_{\mathcal{S}}(x_1) = \boldsymbol{n}_{\mathcal{S}}(x_2)$ in Cartesian coordinates. If $\boldsymbol{n}_{\mathcal{S}}(x_1) = \boldsymbol{n}_{\mathcal{S}}(x_2)$ is true in Cartesian coordinates then it must be true for every coordinate system because it is an equation of two vectors and vectors are coordinate independent objects. Thus we find that $\nabla u(x_1) - \tilde{\boldsymbol{e}}_3 = \nabla u(x_2) - \tilde{\boldsymbol{e}}_3$ and hence $\nabla u(x_1) = \nabla u(x_2)$.

We will now use the strict convexity of u and follow a reasoning from [5, p.93]. Due to the strict convexity u lies above its tangent planes, i.e. $u(x_1) > u(x_2) + (\nabla u(x_2)|\boldsymbol{r}_{21})$, where $\boldsymbol{r}_{21}$ is

the vector pointing from x_2 to x_1. Similarly we have $u(x_2) > u(x_1) + (\nabla u(x_1)|\boldsymbol{r}_{12})$. Adding these two inequalities and subtracting $u(x_1) + u(x_2)$ from both sides we obtain

$$0 > (\nabla u(x_2) - \nabla u(x_1)|\boldsymbol{r}_{12}).$$

However, from our initial assumption it followed that $\nabla u(x_2) - \nabla u(x_1) = 0$. This shows that when one assumes $\nabla u(x_2) - \nabla u(x_1) = 0$ with $x_1 \neq x_2$ one runs into a contradiction, hence the map is injective. □

From $(\tilde{\boldsymbol{e}}_3|\tilde{\boldsymbol{n}}_{\mathcal{S}}) = -\sqrt{e/s}$, we find

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{e}}_3 + 2\frac{\sqrt{e}}{s}\boldsymbol{n}_{\mathcal{S}}. \tag{3.6}$$

The determinant of the metric e is continuously differentiable and the function $u \in C^2(\mathcal{E})$, therefore we see from Equation (3.3) that the mapping $x \mapsto \boldsymbol{n}_{\mathcal{S}}(x)$ is continuously differentiable with non-degenerate Jacobian determinant. The fact that $u \in C^2(\mathcal{E})$ also implies that s is continuously differentiable. This together with the expression above for $\tilde{\boldsymbol{y}}$ implies that the mapping $x \mapsto \tilde{\boldsymbol{y}}(x)$ is also continuously differentiable.

By the bijection $x \mapsto \tilde{\boldsymbol{y}}(x)$ the coordinate system x^1, x^2 on $\mathcal{E}$ induces a coordinate system y^1, y^2 on $\mathcal{G}$. This follows from the fact that the map $x \to \tilde{\boldsymbol{y}}(x)$ is a continuously differentiable bijection as can be seen in Definition 2.1.5. Let $\gamma^i : (\varepsilon, \varepsilon) \to \mathcal{E}$, $\varepsilon > 0$, such that $\gamma^i(t = 0) = x \in \mathcal{E}$, be the part of the coordinate line of the coordinate x^i that passes through x. The map $t \mapsto \tilde{\boldsymbol{y}}(\gamma^i(t))$ then gives part of the coordinate line through the point $\tilde{\boldsymbol{y}}(x) \in \mathcal{G}$ for the coordinate y^i. To obtain the coordinate basis on $\mathcal{G}$ corresponding to the coordinates y^1 and y^2, we need to differentiate $\tilde{\boldsymbol{y}}$ with respect to x^1 and x^2, respectively. However, if we want to find the general basis on $\mathcal{G}$ that corresponds to the, possibly anholonomic, basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$ then we need to take the directional derivatives of $\tilde{\boldsymbol{y}}$ in the direction of these basis vectors, therefore we define the basis of $\mathcal{G}$ by

$$\begin{aligned}\boldsymbol{g}_1 &:= \nabla_{\boldsymbol{e}_1}\tilde{\boldsymbol{y}},\\ \boldsymbol{g}_2 &:= \nabla_{\boldsymbol{e}_2}\tilde{\boldsymbol{y}}.\end{aligned}$$

In this context $\nabla : (\boldsymbol{e}_i, \tilde{\boldsymbol{y}}) \mapsto \nabla_{\boldsymbol{e}_i}\tilde{\boldsymbol{y}}$ should be seen as taking two elements $\boldsymbol{e}_i, \tilde{\boldsymbol{y}} \in T_{\tilde{\boldsymbol{y}}}\mathbb{R}^3 \cong \mathbb{R}^3$ and mapping them to a third element $\nabla_{\boldsymbol{e}_i}\tilde{\boldsymbol{y}} \in \tilde{\boldsymbol{y}} \in T_{\tilde{\boldsymbol{y}}}\mathbb{R}^3$. Furthermore, notice that we do have

$$\boldsymbol{g}_1 = \frac{\partial\tilde{\boldsymbol{y}}}{\partial x^1} \quad \text{and} \quad \boldsymbol{g}_2 = \frac{\partial\tilde{\boldsymbol{y}}}{\partial x^2},$$

if the basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$ on $\mathcal{E}$ is the coordinate basis.

The vectors $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ form a basis for the space $T_{y(x)}\mathcal{G}$, where $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$ is the basis for $T_x\mathcal{E}$. This basis introduces the metric $g_{ij} = (\boldsymbol{g}_1 \mid \boldsymbol{g}_2)$. The inner product here is the inner product on $T_{y(x)}\mathcal{G} \subset \mathbb{R}^3$ induced by the inner product on $\mathbb{R}^3$. The normal on $\mathcal{G}$ is given by $\boldsymbol{n}_{\mathcal{G}} := \boldsymbol{g}_1 \times \boldsymbol{g}_2$ and will be parallel to $\tilde{\boldsymbol{y}}$ but possibly of different length.

In Table 3.1 facts about $\mathcal{E}$, $\mathcal{S}$ and $\mathcal{G}$ derived sofar are summarized. For the surface $\mathcal{G}$, which is a subset of the unit-sphere, we have not yet obtained expressions in terms of the coordinates on $\mathcal{E}$. To derive the Monge-Ampère equation some of these gaps will be filled in the next section.

	$\mathcal{E}$	$\mathcal{S}$	$\mathcal{G}$
local basis	$\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$	$\{\boldsymbol{s}_1 = \boldsymbol{e}_1 + (\nabla_{\boldsymbol{e}_1} u)\tilde{\boldsymbol{e}}_3, \boldsymbol{s}_2 = \boldsymbol{e}_1 + (\nabla_{\boldsymbol{e}_2} u)\tilde{\boldsymbol{e}}_3\}$	$\{\boldsymbol{g}_1, \boldsymbol{g}_2\}$
metric	e_{ij}	$s_{ij} = e_{ij} + (\nabla_{\boldsymbol{e}_i} u)(\nabla_{\boldsymbol{e}_j} u)$	g_{ij}
determinant of metric	e	$s = e(\|\nabla u\|^2 + 1)$	g
normal	$\boldsymbol{e}_1 \times \boldsymbol{e}_2$	$\boldsymbol{n}_{\mathcal{S}} = \sqrt{e}(\nabla u - \tilde{\boldsymbol{e}}_3)$	$\boldsymbol{n}_{\mathcal{G}}$

Table 3.1: Summary of the expressions obtained sofar for the source $\mathcal{E}$, the reflector surface $\mathcal{S}$ and the target $\mathcal{G}$. For $\mathcal{G}$ we do not yet have expressions in the local coordinates on $\mathcal{E}$.

3.3 Energy conservation

A surface element of $\mathcal{E}$ gets mapped onto a surface element of $\mathcal{G}$ by the map $x \mapsto \tilde{\boldsymbol{y}}(x)$. A surface element on $\mathcal{E}$ is given by $\sqrt{e}\ \mathrm{d}x^1\mathrm{d}x^2$ and this surface element gets mapped to a surface element on $\mathcal{G}$ given by $\sqrt{g}\ \mathrm{d}x^1\mathrm{d}x^2$.* Suppose the luminous intensity [lm/sr] on $\mathcal{G}$ is given by the strictly positive function $G : \mathcal{G} \to \mathbb{R}_{>0}$. The principle of conservation of energy implies that the energy flux through $\sqrt{e}\ \mathrm{d}x^1\mathrm{d}x^2$ equals the energy flux through $\sqrt{g}\ \mathrm{d}x^1\mathrm{d}x^2$ and therefore we have

$$E(x)\sqrt{e} = G(y(x))\sqrt{g}, \tag{3.7}$$

where we have used the alternative $x \mapsto y(x)$ for the map $x \mapsto \tilde{\boldsymbol{y}}(x)$. We will often use this notation if we want to consider $y(x)$ as a point on $\mathcal{G}$ instead of as a vector in $\mathbb{R}^3$.

The rest of this section is devoted to writing Equation (3.7) in the local coordinates x^1, x^2 on $\mathcal{E}$ and in a form from which it can be seen that it is clearly independent of the coordinate system used. In order to do this we first focus on $\sqrt{g}$. The following lemma expresses $\sqrt{g}$ in terms of e, s and derivatives of u.

Lemma 3.3.1. *The square root of the determinant of the metric g_{ij} on $\mathcal{G}$ is given by*

$$\sqrt{g} = \frac{4e^{\frac{3}{2}}|(\nabla_{\boldsymbol{e}_1}(\nabla u) \times \nabla_{\boldsymbol{e}_2}(\nabla u)|\boldsymbol{n}_{\mathcal{S}})|}{s^2}, \tag{3.8}$$

where $|\cdot|$ denotes the absolute value function.

Proof. From Lemma 3.1.1 we conclude that $\sqrt{g} = \|\boldsymbol{n}_{\mathcal{G}}\|$, hence we have $\sqrt{g} = \|\boldsymbol{g}_1 \times \boldsymbol{g}_2\|$. Instead of directly evaluating this cross product it helps to rewrite it in the following way,

$$\sqrt{g} = \frac{\|\boldsymbol{g}_1 \times \boldsymbol{g}_2\|}{\|\tilde{\boldsymbol{y}}\|} = \frac{|(\boldsymbol{g}_1 \times \boldsymbol{g}_2|\tilde{\boldsymbol{n}}_{\mathcal{S}})|}{|(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{n}}_{\mathcal{S}})|}. \tag{$\star$}$$

Here we first divided by a vector with norm 1 and then looked at the ratio of the projection of both the numerator and denominator on the vector $\tilde{\boldsymbol{n}}_{\mathcal{S}}$. This second step is justified because the vectors $\boldsymbol{g}_1 \times \boldsymbol{g}_2$ and $\tilde{\boldsymbol{y}}$ are parallel.

The motivation for this step will become clear now that we work out the cross product. By definition of the basis vectors on $\mathcal{G}$ it holds that

$$\boldsymbol{g}_i = \nabla_{\boldsymbol{e}_i}\tilde{\boldsymbol{y}} = \nabla_{\boldsymbol{e}_i}\left(\tilde{\boldsymbol{e}}_3 + 2\frac{\sqrt{e}}{s}\boldsymbol{n}_{\mathcal{S}}\right) = 2\nabla_{\boldsymbol{e}_i}\left(\frac{\sqrt{e}}{s}\right)\boldsymbol{n}_{\mathcal{S}} + 2\left(\frac{\sqrt{e}}{s}\right)\nabla_{\boldsymbol{e}_i}(\boldsymbol{n}_{\mathcal{S}}).$$

This implies that

$$\begin{aligned}\boldsymbol{g}_1 \times \boldsymbol{g}_2 = {} & 4\left(\frac{e}{s^2}\right)\nabla_{\boldsymbol{e}_1}(\boldsymbol{n}_{\mathcal{S}}) \times \nabla_{\boldsymbol{e}_2}(\boldsymbol{n}_{\mathcal{S}}) + 4\left(\frac{\sqrt{e}}{s}\right)\nabla_{\boldsymbol{e}_1}\left(\frac{\sqrt{e}}{s}\right)\boldsymbol{n}_{\mathcal{S}} \times \nabla_{\boldsymbol{e}_2}(\boldsymbol{n}_{\mathcal{S}}) \\ & + 4\left(\frac{\sqrt{e}}{s}\right)\nabla_{\boldsymbol{e}_2}\left(\frac{\sqrt{e}}{s}\right)\nabla_{\boldsymbol{e}_1}(\boldsymbol{n}_{\mathcal{S}}) \times \boldsymbol{n}_{\mathcal{S}}.\end{aligned}$$

*The notion of a surface element is introduced here in a not particularly neat way. The surface element is the special two dimensional variant of a volume form on a manifold. Volume forms are treated for example in Section 2.9 of [1].

The last two terms are perpendicular to $\boldsymbol{n}_{\mathcal{S}}$ and they vanish when taking the inner product with $\tilde{\boldsymbol{n}}_{\mathcal{S}}$, hence we obtain

$$|(\boldsymbol{g}_1 \times \boldsymbol{g}_2|\tilde{\boldsymbol{n}}_{\mathcal{S}})| = \frac{4e(\nabla_{\boldsymbol{e}_1}(\boldsymbol{n}_{\mathcal{S}}) \times \nabla_{\boldsymbol{e}_2}(\boldsymbol{n}_{\mathcal{S}})|\tilde{\boldsymbol{n}}_{\mathcal{S}})}{s^2}.$$

For the covariant derivatives of $\boldsymbol{n}_{\mathcal{S}}$ we find

$$\nabla_{\boldsymbol{e}_i}\boldsymbol{n}_{\mathcal{S}} = \nabla_{\boldsymbol{e}_i}(\sqrt{e})(\nabla u - \tilde{\boldsymbol{e}}_3) + \sqrt{e}\nabla_{\boldsymbol{e}_i}(\nabla u).$$

From this we see that taking the cross product between $\nabla_{\boldsymbol{e}_1}\boldsymbol{n}_{\mathcal{S}}$ and $\nabla_{\boldsymbol{e}_2}\boldsymbol{n}_{\mathcal{S}}$ gives four terms, two of which are perpendicular to $\nabla u - \tilde{\boldsymbol{e}}_3$ and hence perpendicular to $\tilde{\boldsymbol{n}}_{\mathcal{S}}$ and we get one cross product of parallel vectors which equals zero. We obtain

$$(\nabla_{\boldsymbol{e}_1}(\boldsymbol{n}_{\mathcal{S}}) \times \nabla_{\boldsymbol{e}_2}(\boldsymbol{n}_{\mathcal{S}})|\tilde{\boldsymbol{n}}_{\mathcal{S}}) = \frac{e}{\sqrt{s}}(\nabla_{\boldsymbol{e}_1}(\nabla u) \times \nabla_{\boldsymbol{e}_2}(\nabla u)|\boldsymbol{n}_{\mathcal{S}}).$$

Evaluating the denominator in equation $(\star)$ gives

$$|(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{n}}_{\mathcal{S}})| = |(\tilde{\boldsymbol{e}}_3 + 2\frac{\sqrt{e}}{s}\boldsymbol{n}_{\mathcal{S}}|\tilde{\boldsymbol{n}}_{\mathcal{S}})| = \left| -\sqrt{\frac{e}{s}} + 2\sqrt{\frac{e}{s}}\right| = \sqrt{\frac{e}{s}}.$$

Combining the results for the numerator and denominator in equation $(\star)$ we find (3.8). □

It is possible to relate the inner product in (3.8) to the determinant of the Hessian matrix. However, in order to show this we will need the property $\nabla_{\boldsymbol{e}_i}(\flat \mathrm{d}u) = \flat\nabla_{\boldsymbol{e}_i}(\mathrm{d}u)$. Thus, we shall first need to prove that the covariant derivative and the musical isomorphisms $\flat$ and $\sharp$, from Definition 2.3.9, commute.

Lemma 3.3.2. *The covariant derivative and the musical isomorphisms commute, i.e. for a submanifold of Euclidean space M with metric $\boldsymbol{g}$, $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in TM$ and $\hat{\boldsymbol{v}} \in T^*M$ we have*

$$\nabla_{\boldsymbol{u}}(\sharp\boldsymbol{v})(\boldsymbol{w}) = \sharp\nabla_{\boldsymbol{u}}(\boldsymbol{v})(\boldsymbol{w}) \quad \textit{and} \quad \nabla_{\boldsymbol{u}}(\flat\hat{\boldsymbol{v}})(\boldsymbol{w}) = \flat\nabla_{\boldsymbol{u}}(\hat{\boldsymbol{v}})(\boldsymbol{w}).$$

Proof. By definition of the Levi-Civita connection it holds that $\nabla_{\boldsymbol{u}}(\boldsymbol{v}) = u^i\nabla_{\boldsymbol{e}_i}(\boldsymbol{v})$, hence it will be sufficient to show that the musical isomorphisms commute with $\nabla_{\boldsymbol{e}_i}$. By definition of the $\sharp$-operator, the fact that the covariant derivative and the Kronecker tensor commute* and the fourth property in Definition 2.4.6 of the Levi-Civita connection we have

$$\begin{aligned}
\nabla_{\boldsymbol{e}_i}(\sharp\boldsymbol{v})(\boldsymbol{w}) &= \langle\nabla_{\boldsymbol{e}_i}(\sharp\boldsymbol{v}), \boldsymbol{w}\rangle \\
&= \nabla_{\boldsymbol{e}_i}(\langle\sharp\boldsymbol{v}, \boldsymbol{w}\rangle) - \langle\sharp\boldsymbol{v}, \nabla_{\boldsymbol{e}_i}(\boldsymbol{w})\rangle \\
&= \nabla_{\boldsymbol{e}_i}(\sharp\boldsymbol{v}(\boldsymbol{w})) - (\sharp\boldsymbol{v})(\nabla_{\boldsymbol{e}_i}(\boldsymbol{w})) \\
&= \nabla_{\boldsymbol{e}_i}(\boldsymbol{g}(\boldsymbol{v}, \boldsymbol{w})) - (\sharp\boldsymbol{v})(\nabla_{\boldsymbol{e}_i}(\boldsymbol{w})) \\
&= \boldsymbol{g}(\nabla_{\boldsymbol{e}_i}\boldsymbol{v}, \boldsymbol{w}) + \boldsymbol{g}(\boldsymbol{v}, \nabla_{\boldsymbol{e}_i}(\boldsymbol{w})) - (\sharp\boldsymbol{v})(\nabla_{\boldsymbol{e}_i}(\boldsymbol{w})) \\
&= \sharp(\nabla_{\boldsymbol{e}_i}\boldsymbol{v})(\boldsymbol{w}) + (\sharp\boldsymbol{v})(\nabla_{\boldsymbol{e}_i}(\boldsymbol{w})) - (\sharp\boldsymbol{v})(\nabla_{\boldsymbol{e}_i}(\boldsymbol{w})) \\
&= \sharp(\nabla_{\boldsymbol{e}_i}\boldsymbol{v})(\boldsymbol{w}).
\end{aligned}$$

Thus we see that $\nabla_{\boldsymbol{e}_i}(\sharp\boldsymbol{v}) = \sharp(\nabla_{\boldsymbol{e}_i}\boldsymbol{v})$. From this the same property for the $\flat$-operator follows quickly by noticing that each covector $\hat{\boldsymbol{v}}$ can be written as $\sharp\boldsymbol{v}$ for some vector $\boldsymbol{v}$, because the $\sharp$-operator is an isomorphism. Using the fact that the $\sharp$-operator and the covariant derivative commute we find

$$\nabla_{\boldsymbol{e}_k}(\flat\hat{\boldsymbol{v}}) = \nabla_{\boldsymbol{e}_k}(\flat(\sharp\boldsymbol{v})) = \nabla_{\boldsymbol{e}_k}(\boldsymbol{v}) = \flat(\sharp(\nabla_{\boldsymbol{e}_k}(\boldsymbol{v}))) = \flat(\nabla_{\boldsymbol{e}_k}(\sharp\boldsymbol{v})) = \flat(\nabla_{\boldsymbol{e}_k}(\hat{\boldsymbol{v}})).$$

□

*The components of the covariant derivative of the Kronecker tensor are given by $D_k(\delta^i_j) = \nabla_{\boldsymbol{e}_k}\delta^i_j + \Gamma^i_{lk}\delta^l_j - \Gamma^l_{jk}\delta^i_l = \Gamma^i_{jk} - \Gamma^i_{jk} = 0$, therefore we have $\nabla_{\boldsymbol{e}_i}\langle\hat{\boldsymbol{v}}, \boldsymbol{w}\rangle = \langle\nabla_{\boldsymbol{e}_i}(\hat{\boldsymbol{v}}), \boldsymbol{w}\rangle + \langle\hat{\boldsymbol{v}}, \nabla_{\boldsymbol{e}_i}(\boldsymbol{w})\rangle$.

Now that it is clear that we can interchange the order of applying a musical isomorphism with covariant differentiation we turn our attention to the metric on $\mathcal{G}$ again. In the next lemma we show that the cross product in the inner product in equation (3.8) can be related to the determinant of the Hessian tensor.

Lemma 3.3.3. *The cross product $\nabla_{\boldsymbol{e}_1}(\nabla u) \times \nabla_{\boldsymbol{e}_2}(\nabla u)$ in (3.8) is parallel to $\tilde{\boldsymbol{e}}_3$ and has length $\det(H_{ij})/\sqrt{e}$, i.e.*

$$\nabla_{\boldsymbol{e}_1}(\nabla u) \times \nabla_{\boldsymbol{e}_2}(\nabla u) = \frac{\det(H_{ij})}{\sqrt{e}}\tilde{\boldsymbol{e}}_3,$$

where H_{ij} are the components of the Hessian tensor $\boldsymbol{H}(u)$ which is given by[*]

$$\boldsymbol{H} = \big(\nabla_{\boldsymbol{e}_j}(\nabla_{\boldsymbol{e}_i} u) - \Gamma^k_{ij}\nabla_{\boldsymbol{e}_k}(u)\big)\hat{\boldsymbol{e}}^i \otimes \hat{\boldsymbol{e}}^j.$$

Proof. By Lemma 3.3.2 it follows that

$$\begin{aligned}
\nabla_{\boldsymbol{e}_k}(\nabla u) &= \nabla_{\boldsymbol{e}_k}(\flat \mathrm{d}u) \\
&= \flat(\nabla_{\boldsymbol{e}_k}(\mathrm{d}u)) \\
&= \flat([\nabla_{\boldsymbol{e}_k}(\nabla_{\boldsymbol{e}_i} u) - \nabla_{\boldsymbol{e}_l}(u)\Gamma^l_{ik}]\hat{\boldsymbol{e}}^i) \\
&= e^{ij}(\nabla_{\boldsymbol{e}_k}(\nabla_{\boldsymbol{e}_i} u) - \nabla_{\boldsymbol{e}_l}(u)\Gamma^l_{ik})\boldsymbol{e}_j.
\end{aligned}$$

From this and the fact that the metric has the form as shown in equation (3.1) it follows that

$$\begin{aligned}
&\nabla_{\boldsymbol{e}_1}(\nabla u) \times \nabla_{\boldsymbol{e}_2}(\nabla u) \\
&\quad= \epsilon_{\alpha\beta\gamma} e^{\alpha\tau} e^{\beta\sigma} \left[(\nabla_{\boldsymbol{e}_1}(\nabla_{\boldsymbol{e}_\tau} u) - \nabla_{\boldsymbol{e}_\lambda}(u)\Gamma^\lambda_{\tau 1})\right]\left[(\nabla_{\boldsymbol{e}_2}(\nabla_{\boldsymbol{e}_\sigma} u) - \nabla_{\boldsymbol{e}_\rho}(u)\Gamma^\rho_{\sigma 2})\right] e^{\gamma\kappa}\boldsymbol{e}_\kappa \\
&\quad= \epsilon_{\alpha\beta 3} e^{\alpha\tau} e^{\beta\sigma} \left[(\nabla_{\boldsymbol{e}_1}(\nabla_{\boldsymbol{e}_\tau} u) - \nabla_{\boldsymbol{e}_\lambda}(u)\Gamma^\lambda_{\tau 1})\right]\left[(\nabla_{\boldsymbol{e}_2}(\nabla_{\boldsymbol{e}_\sigma} u) - \nabla_{\boldsymbol{e}_\rho}(u)\Gamma^\rho_{\sigma 2})\right] \tilde{\boldsymbol{e}}_3 \\
&\quad= \sqrt{e}\; e^{11} e^{22} \Bigg[\left[(\nabla_{\boldsymbol{e}_1}(\nabla_{\boldsymbol{e}_1} u) - \nabla_{\boldsymbol{e}_\lambda}(u)\Gamma^\lambda_{11})\right]\left[(\nabla_{\boldsymbol{e}_2}(\nabla_{\boldsymbol{e}_2} u) - \nabla_{\boldsymbol{e}_\rho}(u)\Gamma^\rho_{22})\right] \\
&\qquad - \left[(\nabla_{\boldsymbol{e}_1}(\nabla_{\boldsymbol{e}_2} u) - \nabla_{\boldsymbol{e}_\lambda}(u)\Gamma^\lambda_{21})\right]\left[(\nabla_{\boldsymbol{e}_2}(\nabla_{\boldsymbol{e}_1} u) - \nabla_{\boldsymbol{e}_\rho}(u)\Gamma^\rho_{12})\right] \Bigg] \tilde{\boldsymbol{e}}_3 \\
&\quad= \frac{\det(H_{ij})}{\sqrt{e}}\tilde{\boldsymbol{e}}_3,
\end{aligned}$$

□

The last lemma enables us to express the determinant of the metric g solely in terms of the local coordinates on the source $\mathcal{E}$. This allows us to also rewrite the energy conservation equation (3.7) in terms of the local coordinates on $\mathcal{E}$.

Theorem 3.3.4. *The conservation of energy under the mapping $x \mapsto \tilde{\boldsymbol{y}}(x)$ is expressed by the equation*

$$\frac{E(x)}{G(y(x))} = \frac{4|\det(H_{ij})|}{e(\|\nabla u\|^2 + 1)^2}, \tag{3.9}$$

where we write $y(x)$ for the element $\tilde{\boldsymbol{y}}(x)$ of $\mathcal{G}$.

Proof. From Lemma 3.3.3 it follows that

$$\begin{aligned}
|(\nabla_{\boldsymbol{e}_1}(\nabla u) \times \nabla_{\boldsymbol{e}_2}(\nabla u)|\boldsymbol{n}_{\mathcal{S}})| &= \frac{|\det(H_{ij})|}{\sqrt{e}}|(\tilde{\boldsymbol{e}}_3|\boldsymbol{n}_{\mathcal{S}})| \\
&= \frac{|\det(H_{ij})|}{\sqrt{e}}|(\tilde{\boldsymbol{e}}_3|\sqrt{e}(\nabla u - \tilde{\boldsymbol{e}}_3)| \\
&= |\det(H_{ij})|.
\end{aligned}$$

[*]We use Latin indices, because we mean the Hessian tensor with respect to the basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$, not the basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \tilde{\boldsymbol{e}}_3\}$.

Substituting this in (3.8) we obtain

$$\sqrt{g} = \frac{4e^{\frac{3}{2}}|\det(H_{ij})|}{s^2}.$$

In this expression we again substitute the expression for s in Lemma 3.1.2 and finally find

$$\sqrt{g} = \frac{4|\det(H_{ij})|}{\sqrt{e}(\|\nabla u\|^2+1)^2}.$$

Substituting $\sqrt{g}$ in equation (3.7) we obtain (3.9). □

Equation (3.9) is called the *Monge-Ampère equation*. For this equation to be really coordinate independent, $|\det(H_{ij})|/e$ needs to be an invariant. To establish this, let us consider a change of coordinate system which results in a change of basis given by a matrix A, i.e. $\boldsymbol{f}_i = A_i^j \boldsymbol{e}_j$. We use the overline to indicate that tensors are given in the new coordinate system. By the tensor transformation law it follows that for the components of the Hessian tensor hold that $\overline{H}_{ij} = A_i^k A_j^l H_{kl}$ and similarly for the components of the metric that $\overline{e}_{ij} = A_i^k A_j^l e_{kl}$. This implies

$$\det(\overline{H}_{ij}) = (\det(A))^2 \det(H_{ij})$$

and similarly for the metric that

$$\overline{e} = (\det(A))^2 e.$$

This shows that

$$\frac{|\det(\overline{H}_{ij})|}{\overline{e}} = \frac{|(\det(A))^2 \det(H_{ij})|}{(\det(A))^2 e} = \frac{|\det(H_{ij})|}{e}$$

and hence the quotient $|\det(H_{ij})|/e$ is indeed an invariant. Thus we have shown that the Monge-Ampère equation (3.9) is independent of the choice of orthogonal coordinate system.

3.4 Coordinate specific expressions for the Monge-Ampère equation

We will consider three specific coordinate systems: Cartesian coordinates, polar coordinates with the coordinate basis and polar coordinates with an orthonormal anholonomic basis. In Cartesian coordinates we hope to find the Monge-Ampèren equation as it was given in [5]. Furthermore, we will see that in polar coordinates for both the holonomic and anholonomic basis, we find the same equation, while the road leading to it is different. Nonetheless, it is not surprising that we find the same equation as the Monge-Ampère equation is a scalar equation and hence only dependent on the coordinates used and not on the basis chosen. Let us start out with the Cartesian coordinates.

Example 3.4.1. The Cartesian coordinate system is really the simplest case. In Cartesian coordinates with orthonormal basis the matrix representation of the metric is just the identity and all the Christoffel symbols vanish. For the gradient we find

$$\nabla_{\boldsymbol{e}_i}(u) e^{ij} \boldsymbol{e}_j = \frac{\partial u}{\partial x^i} e^{ij} \boldsymbol{e}_i = \frac{\partial u}{\partial x}\boldsymbol{e}_1 + \frac{\partial u}{\partial y}\boldsymbol{e}_2.$$

Writing u_x for the partial derivative of u with respect to x and similarly for y we find that $\|\nabla u\|^2 = u_x^2 + u_y^2$. The Hessian tensor is given by

$$\boldsymbol{H}(u) = \big(\partial_j(\partial_i u) - \Gamma_{ij}^k \partial_k u\big)\hat{\boldsymbol{e}}^i \otimes \hat{\boldsymbol{e}}^j = \big(\partial_j(\partial_i u)\big)\hat{\boldsymbol{e}}^i \otimes \hat{\boldsymbol{e}}^j.$$

This implies that the determinant of the matrix representation of the Hessian tensor equals $\det(H_{ij}) = u_{xx}u_{yy} - u_{xy}^2$. Collecting the results we find that the Monge-Ampère equation in Cartesian coordinates is given by

$$\frac{E(x)}{G(y(x))} = \frac{4|u_{xx}u_{yy} - u_{xy}^2|}{(u_x^2 + u_y^2 + 1)^2}.$$

This is the same equation as given in [5], where the Monge-Ampère equation was derived in Cartesian coordinates.

Let us now proceed with polar coordinates. First we consider polar coordinates with a holonomic basis.

Example 3.4.2. In Example 2.4.8 we considered polar coordinates with the coordinate basis for the Euclidean plane. There we found that the metric is given by

$$(e_{ij}) = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}$$

and hence $e = r^2$. The nonzero Christoffel symbols are $\Gamma^r_{\theta\theta} = -r$, $\Gamma^\theta_{r\theta} = 1/r$ and $\Gamma^\theta_{\theta r} = 1/r$. Moreover, the matrix representation of the Hessian tensor is given in (2.24). This results in the determinant

$$\det(H_{ij}) = u_{rr}u_{\theta\theta} + ru_ru_{rr} - u_{r\theta}^2 - \left(\frac{u_\theta}{r}\right)^2 + \frac{2u_\theta u_{r\theta}}{r}.$$

For the gradient of u we find

$$\nabla u = \partial_i(u)e^{ij}\boldsymbol{e}_j = u_r\boldsymbol{e}_r + \frac{u_\theta}{r^2}\boldsymbol{e}_\theta,$$

hence the norm of the gradient is given by $\|\nabla u\| = u_r^2 + (u_\theta/r^2)^2$. Combining the different results we find that in polar coordinates the Monge-Ampère equation is given by

$$\frac{E(x)}{G(y(x))} = \frac{4|u_{rr}u_{\theta\theta} + ru_ru_{rr} - u_{r\theta}^2 - (u_\theta/r)^2 + (2u_\theta u_{r\theta})/r|}{(r + ru_r^2 + u_\theta^2/r)^2}. \tag{3.10}$$

Now we end with the last example, that is that for polar coordinates with an anholonomic basis.

Example 3.4.3. Recall from Example 2.4.9 that when we rescale the $\boldsymbol{e}_\theta$ basis vector to unit length we obtain $\overline{\boldsymbol{e}}_r = \boldsymbol{e}_r$ and $\overline{\boldsymbol{e}}_\theta = r^{-1}\boldsymbol{e}_\theta$. This implies that $\overline{e} = 1$ The directional derivatives for the basis vectors no longer commute and we have nonzero commutation symbols, moreover the symmetry in the lower indices of the Christoffel symbols is lost. From the example just referred to we know that there are two nonzero Christoffel symbols, namely $\overline{\Gamma}^r_{\theta\theta} = -r^{-1}$ and $\overline{\Gamma}^\theta_{r\theta} = r^{-1}$. The matrix representation of the Hessian tensor is given by (2.25). The gradient is now given by

$$\nabla u = \nabla_{\overline{\boldsymbol{e}}_i}(u)e^{ij}\overline{\boldsymbol{e}}_j = u_r\overline{\boldsymbol{e}}_r + \frac{u_\theta}{r}\overline{\boldsymbol{e}}_\theta. \tag{3.11}$$

These facts imply that we again end up with Equation (3.10).

3.5 Boundary value problem

In Section 3.3 we derived the Monge-Ampère equation (3.9). This equation tells us how the source emittance E is related to the luminous intensity G when light from the source gets reflected by a reflecting surface. We are concerned with the following problem. Given a source $\mathcal{E} \subset \mathbb{R}^2$ with an emittance function E and a set of output directions $\mathcal{G} \subset \mathcal{S}^2$ with an output intensity function G,

what is the reflector shape given by a function $u : \mathcal{E} \to \mathbb{R}$ such that u satisfies the Monge-Ampère equation (3.9) and moreover $y(\mathcal{E}) = \mathcal{G}$, where $y : \mathcal{E} \to \mathcal{S}^2 : x \mapsto \tilde{\boldsymbol{y}}(x)$ is the map given in Lemma 3.2.1. The fact that u has to satisfy the Monge-Ampère equation corresponds to local conservation of energy. Moreover, $y(\mathcal{E}) = \mathcal{G}$ ensures that we have global conservation of energy, i.e. all the energy emitted by the source eventually gets reflected in the directions specified by $\mathcal{G}$.

It is not clear that for every combination of $\mathcal{E}$, $\mathcal{G}$, E and G a function u exists that solves this problem. The energy emitted by the source and the energy output specified by G do at least need to match up. We need to have

$$\int_{\mathcal{E}} E(x^1, x^2)\sqrt{e(x^1, x^2)}\mathrm{d}x^1\mathrm{d}x^2 = \int_{\mathcal{G}} G(y^1, y^2)\sqrt{g(y^1, y^2)}\mathrm{d}y^1\mathrm{d}y^2, \tag{3.12}$$

where x^1, x^2 are local coordinates on $\mathcal{E}$ with corresponding determinant of the metric on $\sqrt{e(x^1, x^2)}$ and similarly y^1, y^2 are local coordinates on $\mathcal{G}$ with corresponding determinant of the metric $\sqrt{g(y^1, y^2)}$.

Let us now clearly formulate the problem with this additional constraint on the functions E and G.

Problem 3.5.1. Let $\mathcal{E} \subset \mathbb{R}^2$ be convex, closed and bounded, and let $\mathcal{G} \subset \mathcal{S}^2$ be closed. Furthermore, let $E : \mathcal{E} \to \mathbb{R}_{>0}$ and $G : \mathcal{G} \to \mathbb{R}_{>0}$ be such that they satisfy (3.12). Find $u \in C^2(\mathcal{E})$ such that u satisfies the Monge-Ampère equation (3.9) and moreover $y(\mathcal{E}) = \mathcal{G}$ holds, where y is the map from $\mathcal{E}$ to $\mathcal{S}^2$ given by

$$x \mapsto \tilde{\boldsymbol{e}}_3 + \frac{2(\nabla u - \tilde{\boldsymbol{e}}_3)}{\|\nabla u - \tilde{\boldsymbol{e}}_3\|^2}.$$

We demand that $\mathcal{E}$ is convex, closed and bounded, and that $\mathcal{G}$ is closed because then we can reformulate this problem to another problem that is simpler and more convenient. To do this we will show that the mapping y can be seen as the composition of two maps, namely a composition of the map $\nabla u : \mathcal{E} \to \mathbb{R}^2$, given by

$$x \mapsto \nabla u(x),$$

and a second map that we will denote by ψ, which is given by

$$\boldsymbol{v} \mapsto \tilde{\boldsymbol{e}}_3 + \frac{2(\boldsymbol{v} - \tilde{\boldsymbol{e}}_3)}{\|\boldsymbol{v} - \tilde{\boldsymbol{e}}_3\|^2}.$$

Formally speaking, the vector $\nabla u(x)$ is a vector in the space $T_x\mathcal{E}$, however, we saw that for two-dimensional Euclidean space we can identify this space $\mathbb{R}^2$. We will interpret ∇u as such, i.e. as a vector in $\mathbb{R}^2 \times \{0\} \subset \mathbb{R}^3$. The map ψ is the familiar stereographic projection, which is shown in Figure 3.2. This map is a continuously differentiable bijection between the plane and the unit sphere without the north-pole. This we will show in the following lemma.

Lemma 3.5.2. *The map ψ defined by*

$$\psi(\boldsymbol{v}) := \tilde{\boldsymbol{e}}_3 + \frac{2(\boldsymbol{v} - \tilde{\boldsymbol{e}}_3)}{\|\boldsymbol{v} - \tilde{\boldsymbol{e}}_3\|^2}, \tag{3.13}$$

where $\boldsymbol{v} \in \mathbb{R}^2 = \{(x, y, 0) \in \mathbb{R}^3 \mid x, y \in \mathbb{R}\}$ and the vector $\tilde{\boldsymbol{e}}_3 = (0, 0, 1)$ is a unit vector perpendicular to the plane of $\boldsymbol{v}$, is a bijection from $\mathbb{R}^2$ to $\mathcal{S}^2\backslash(0, 0, 1)$, i.e. from $\mathbb{R}^2$ to $\mathcal{S}^2$ without the north pole.

Proof. We will first proof the injectivity of the mapping. Suppose we have two distinct $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^2$ such that $\psi(\boldsymbol{v}_1) = \psi(\boldsymbol{v}_2)$. This implies that

$$\frac{2(\boldsymbol{v}_1 - \tilde{\boldsymbol{e}}_3)}{\|\boldsymbol{v}_1 - \tilde{\boldsymbol{e}}_3\|^2} = \frac{2(\boldsymbol{v}_2 - \tilde{\boldsymbol{e}}_3)}{\|\boldsymbol{v}_2 - \tilde{\boldsymbol{e}}_3\|^2}.$$

Now, because $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are by definition both orthogonal to $\tilde{\boldsymbol{e}_3}$, this implies that we have

$$\begin{cases} \dfrac{2\boldsymbol{v}_1}{\|\boldsymbol{v}_1 - \tilde{\boldsymbol{e}}_3\|^2} &= \dfrac{2\boldsymbol{v}_2}{\|\boldsymbol{v}_2 - \tilde{\boldsymbol{e}}_3\|^2}, \\ \dfrac{2\tilde{\boldsymbol{e}}_3}{\|\boldsymbol{v}_1 - \tilde{\boldsymbol{e}}_3\|^2} &= \dfrac{2\tilde{\boldsymbol{e}}_3}{\|\boldsymbol{v}_2 - \tilde{\boldsymbol{e}}_3\|^2}. \end{cases}$$

The second of these equations implies that $\|\boldsymbol{v}_1 - \tilde{\boldsymbol{e}}_3\|^2 = \|\boldsymbol{v}_2 - \tilde{\boldsymbol{e}}_3\|^2$ and this fact together with the first equation implies that $\boldsymbol{v}_1 = \boldsymbol{v}_2$. Thus, ψ is injective.

Let us now proof the surjectivity. Suppose we have $\tilde{\boldsymbol{s}} \in \mathcal{S}\backslash(0,0,1)$. Let us denote $\tilde{\boldsymbol{s}}$ component-wise as $\tilde{\boldsymbol{s}} = (s^x, s^y, s^z)$, where $(s^x)^2 + (s^y)^2 + (s^z)^2 = 1$ and $\tilde{\boldsymbol{s}} \neq (0,0,1)$. We will now show that there exists a vector $\boldsymbol{v} \in \mathbb{R}^2$ such that $\psi(\boldsymbol{v}) = \tilde{\boldsymbol{s}}$. Let us denote $\boldsymbol{v}$ component-wise as $\boldsymbol{v} = (v^x, v^y, 0)$. A straightforward calculation shows that

$$\psi(\boldsymbol{v}) = \left(\frac{2v^x}{(v^x)^2 + (v^y)^2 + 1}, \frac{2v^y}{(v^x)^2 + (v^y)^2 + 1}, \frac{(v^x)^2 + (v^y)^2 - 1}{(v^x)^2 + (v^y)^2 + 1} \right).$$

If we switch to polar coordinates, i.e. define $v^r = \sqrt{(v^x)^2 + (v^y)^2}$ and $v^\theta = \tan^{-1}(v^y/v^x)$ with $\tan^{-1}$ as defined in Example 2.1.6, we see that

$$\psi(\boldsymbol{v}) = \left(\frac{2v^r \cos(v^\theta)}{(v^r)^2 + 1}, \frac{2v^r \sin(v^\theta)}{(v^r)^2 + 1}, \frac{(v^r)^2 - 1}{(v^r)^2 + 1} \right).$$

Equation this vector with $\tilde{\boldsymbol{s}}$ and dividing the second components by the first components of both vectors we obtain $\tan(v^\theta) = s^y/s^x$ which implies that $v^\theta = \tan^{-1}(s^y/s^x)$. From equation the third component of $\psi(\boldsymbol{v})$ and $\tilde{\boldsymbol{s}}$ we find that $(v^r)^2 - 1 = ((v^r)^2 + 1)s_z$ and this implies that $v^r = \sqrt{(1+s^z)/(1-s^z)}$. Thus, we see that if $\boldsymbol{v}$ is given by $v^r = \sqrt{(1+s^z)/(1-s^z)}$ and $v^\theta = \tan^{-1}(s^y/s^x)$, then $\psi(\boldsymbol{v}) = \tilde{\boldsymbol{s}}$ and hence ψ is surjective. □

We will use the bijection ψ to transfer the subset $\mathcal{G} \subset \mathcal{S}^2$ and function $G : \mathcal{G} \to \mathbb{R}_{>0}$ to a subset $\mathcal{F} \subset \mathbb{R}^2$ and a function $F : \mathcal{F} \to \mathbb{R}_{>0}$, respectively. We do this by defining $\mathcal{F}$ as the pre-image of $\mathcal{G}$ under ψ, i.e.

$$\mathcal{F} := \psi^{-1}(\mathcal{G}). \tag{3.14}$$

In order to be able to define the function F on $\mathcal{F}$ corresponding to the function G on $\mathcal{G}$ we need to relate the flux through a surface element of $\mathcal{G}$ to the flux through a surface element of $\mathcal{F}$. To do this, we determine in next lemma how the surface elements on $\mathcal{G}$ and $\mathcal{F}$ are related by the stereographic projection.

Lemma 3.5.3. *Let $\psi : \mathbb{R}^2 \times \{0\} \to \mathcal{S}^2\backslash(0,0,1)$ be the stereographic projection as defined earlier. Let us denote the surface element on the unit sphere by $\mathrm{d}A_{\mathcal{S}^2}$ and the surface element on the plane by $\mathrm{d}A_{\mathbb{R}^2}$. The two surface elements are related by*

$$\mathrm{d}A_{\mathcal{S}^2} = \frac{4\mathrm{d}A_{\mathbb{R}^2}}{(1 + \|\boldsymbol{v}\|^2)^2}, \tag{3.15}$$

where $\boldsymbol{v}$ is the vector in the plane that points from the origin to the point under consideration.

If x^1, x^2 are local coordinates on $\mathbb{R}^2 \times \{0\}$ and y^1, y^2 are local coordinates on $\mathcal{S}^2\backslash(0,0,1)$, then we can express this equivalently as

$$\sqrt{g(y^1, y^2)}\ \mathrm{d}y^1 \mathrm{d}y^2 = \frac{4\sqrt{f(x^1, x^2)}\ \mathrm{d}x^1 \mathrm{d}x^2}{(1 + \|(x^1, x^2)\|^2)^2},$$

where g is the determinant of the metric on the unit sphere, f is the determinant of the metric on the plane and $\|(x^1, x^2)\|$ is the distance from the point specified by x^1 and x^2 to the origin.

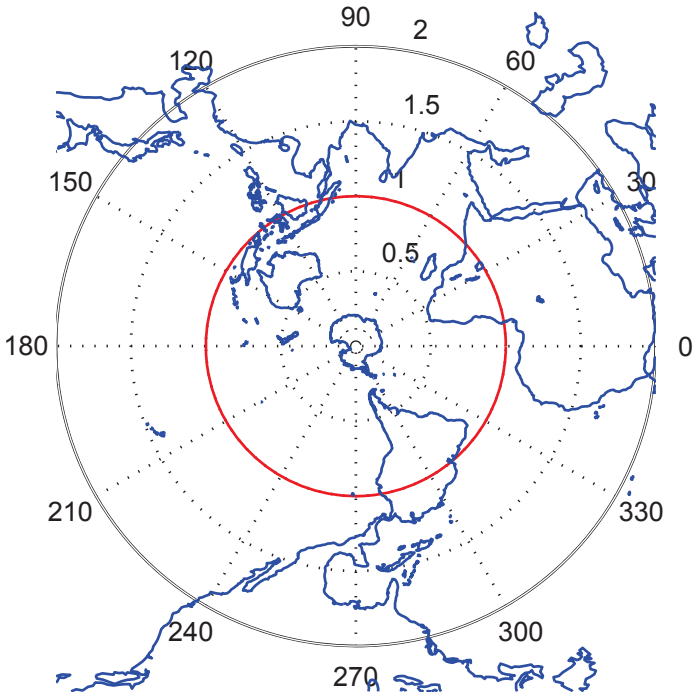


Figure 3.1: In this figure a world chart resulting from the stereographic projection is depicted. The map ψ^{-1} is a map from $\mathcal{S}^2 \backslash (0,0,1)$ to the plane $\mathbb{R}^2 \times \{0\}$ and can therefore be used to construct a 2D world chart. It can be seen that the lower hemisphere gets mapped to the unit disk. The equator is depicted in red. The northern hemisphere is projected around the unit disk and is severely distorted. The south pole is projected onto the origin. At approximately 30 degrees we find Africa, at approximately 135 degrees we find Australia and at approximately 300 degrees we find South America.

Proof. Suppose we have a spherical coordinate system on $\mathcal{S}^2\backslash(0,0,1)$ and a Cartesian coordinate system on $\mathbb{R}^2 \times \{0\}$. Suppose we have defined the Cartesian coordinate system on $\mathbb{R}^2 \times \{0\}$ with basis vectors $\boldsymbol{e}_x$ and $\boldsymbol{e}_y$. We define the spherical coordinates (θ, ϕ) with respect to these basis vectors on $\mathbb{R}^2 \times \{0\}$. Let θ be the azimuth angle with respect to $\boldsymbol{e}_x$ and let ϕ be the angle with respect to $\tilde{\boldsymbol{e}}_3$. The relation between these two coordinate systems is given by

$$\theta(x,y) = \tan^{-1}\left(\frac{x}{y}\right), \tag{3.16}$$

$$\phi(x,y) = \arccos\left(\frac{(x)^2+(y)^2-1}{(x)^2+(y)^2+1}\right). \tag{3.17}$$

In the spherical coordinate system the surface element on the unit sphere is given by

$$\mathrm{d}A_{\mathcal{S}^2} = \sin(\phi)\ \mathrm{d}\theta\mathrm{d}\phi.$$

The surface element on the plane is in Cartesian coordinates given by

$$\mathrm{d}A_{\mathbb{R}^2} = \mathrm{d}x\mathrm{d}y.$$

From the change of variables formula we know that

$$\mathrm{d}\theta\mathrm{d}\phi = \left|\det\left(\frac{\partial(\theta,\phi)}{\partial(x,y)}\right)\right| \mathrm{d}x\mathrm{d}y,$$

where

$$\frac{\partial(\theta,\phi)}{\partial(x,y)} = \begin{pmatrix} \dfrac{\partial\theta}{\partial x} & \dfrac{\partial\theta}{\partial y} \\ \dfrac{\partial\phi}{\partial x} & \dfrac{\partial\phi}{\partial y} \end{pmatrix}.$$

The derivative of $\arccos(x)$ with respect to x is given by $-(1-x^2)^{-1/2}$. Using this fact we find

$$\begin{aligned}\frac{\partial\phi}{\partial x} &= -\left(1-\left(\frac{x^2+y^2-1}{x^2+y^2+1}\right)^2\right)^{-1/2} \frac{2x(x^2+y^2+1)-2x(x^2+y^2-1)}{(x^2+y^2+1)^{-2}} \\ &= -\left(1-\left(\frac{x^2+y^2-1}{x^2+y^2+1}\right)^2\right)^{-1/2} \frac{4x}{(x^2+y^2+1)^2}.\end{aligned}$$

By the symmetry of the expression for ϕ when interchanging x and y we find that

$$\frac{\partial\phi}{\partial y} = -\left(1-\left(\frac{x^2+y^2-1}{x^2+y^2+1}\right)^2\right)^{-1/2} \frac{4y}{(x^2+y^2+1)^2}.$$

Furthermore, we have

$$\frac{\partial\theta}{\partial x} = \frac{-y}{x^2+y^2} \quad \text{and} \quad \frac{\partial\theta}{\partial y} = \frac{x}{x^2+y^2}.$$

From this we find that

$$\left|\det\left(\frac{\partial(\theta,\phi)}{\partial(x,y)}\right)\right| = \left|\frac{\partial\theta}{\partial x}\frac{\partial\phi}{\partial y} - \frac{\partial\theta}{\partial y}\frac{\partial\phi}{\partial x}\right| = 4\left(1-\left(\frac{x^2+y^2-1}{x^2+y^2+1}\right)^2\right)^{-1/2}\left(x^2+y^2+1\right)^{-2}.$$

Thus we have

$$\mathrm{d}\theta\mathrm{d}\phi = 4\left(1-\left(\frac{x^2+y^2-1}{x^2+y^2+1}\right)^2\right)^{-1/2}\left(x^2+y^2+1\right)^{-2}\mathrm{d}x\mathrm{d}y. \tag{$\star$}$$

Furthermore, we have

$$\begin{aligned}\sin(\phi) &= \sin\left(\arccos\left(\frac{(x)^2+(y)^2-1}{(x)^2+(y)^2+1}\right)\right)\\ &= \sqrt{1-\left(\frac{(x)^2+(y)^2-1}{(x)^2+(y)^2+1}\right)^2},\end{aligned}$$

where we used the fact that $\sin(\arccos(z)) = \sqrt{1-z^2}$, which can easily be verified by drawing a right triangle with a hypothenuse of length one and $\arccos(z)$ equal to one of the nonzero angles. Now by equation $(\star)$ it follows that

$$\begin{aligned}\mathrm{d}A_{\mathcal{S}^2} &= \sin(\phi)\mathrm{d}\theta\mathrm{d}\phi\\ &= 4\left(1-\left(\frac{x^2+y^2-1}{x^2+y^2+1}\right)^2\right)^{1/2}\left(1-\left(\frac{x^2+y^2-1}{x^2+y^2+1}\right)^2\right)^{-1/2}\left(x^2+y^2+1\right)^{-2}\mathrm{d}x\mathrm{d}y\\ &= \frac{4\mathrm{d}x\mathrm{d}y}{(x^2+y^2+1)}\\ &= \frac{4\mathrm{d}A_{\mathbb{R}^2}}{(x^2+y^2+1)}.\end{aligned}$$

For general coordinate systems y^1, y^2 and x^1, x^2 on the unit sphere and the plane, respectively, the surface elements are given by

$$\mathrm{d}A_{\mathcal{S}^2} = \sqrt{g(y^1,y^2)}\mathrm{d}y^1\mathrm{d}y^2 \quad \text{and} \quad \mathrm{d}A_{\mathbb{R}^2} = \sqrt{f(x^1,x^2)}\mathrm{d}x^1\mathrm{d}x^2,$$

where g is the determinant of the metric on the unit sphere and f is the determinant of the metric on the plane. Therefore we find for these general coordinate systems

$$\sqrt{g(y^1,y^2)}\ \mathrm{d}y^1\mathrm{d}y^2 = \frac{4\sqrt{f(x^1,x^2)}\ \mathrm{d}x^1\mathrm{d}x^2}{(1+\|(x^1,x^2)\|^2)^2},$$

where $\|(x^1,x^2)\|$ is the distance from the point specified by x^1 and x^2 to the origin. □

The results of Lemma 3.5.3 are visualized in Figure 3.2. We can use the result of the previous lemma to define the function F on $\mathcal{F}$. Suppose U is a subset of $\mathcal{F}$. By definition $\psi(U)$ is a subset of $\mathcal{G}$. By Lemma 3.5.3 we have

$$\int_{\psi(U)} G(y^1,y^2)\sqrt{g(y^1,y^2)}\ \mathrm{d}y^1\mathrm{d}y^2 = \int_U \frac{4G(y^1(x^1,x^2),y^2(x^1,x^2))}{(1+\|(x^1,x^2)\|^2)^2}\sqrt{f(x^1,x^2)}\ \mathrm{d}x^1\mathrm{d}x^2.$$

This implies that we should define the function F on $\mathcal{F}$ as

$$F(x^1,x^2) := \frac{4G(y^1(x^1,x^2),y^2(x^1,x^2))}{(1+\|(x^1,x^2)\|^2)^2}. \tag{3.18}$$

With this definition we have

$$\int_{\mathcal{G}} G(y^1,y^2)\sqrt{g(y^1,y^2)}\ \mathrm{d}y^1\mathrm{d}y^2 = \int_{\mathcal{F}} F(x^1,x^2)\sqrt{f(x^1,x^2)}\ \mathrm{d}x^1\mathrm{d}x^2. \tag{3.19}$$

Combining (3.19) with (3.12) we obtain

$$\int_{\mathcal{E}} E(x^1,x^2)\sqrt{e(x^1,x^2)}\mathrm{d}x^1\mathrm{d}x^2 = \int_{\mathcal{F}} F(x^1,x^2)\sqrt{f(y^1,y^2)}\mathrm{d}y^1\mathrm{d}y^2, \tag{3.20}$$

where now x^1, x^2 is a local coordinate system on $\mathcal{E}$ and y^1, y^2 is a local coordinate system on $\mathcal{F}$. Equation (3.20) is the new constraint on the functions $E : \mathcal{E} \to \mathbb{R}_{>0}$ and $F : \mathcal{F} \to \mathbb{R}_{>0}$. In

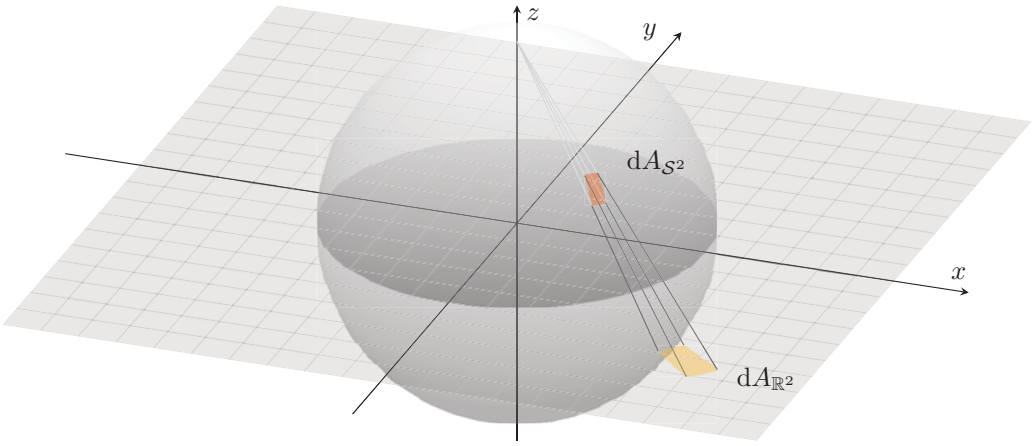


Figure 3.2: The stereographic projection is depicted. A line through the north pole of the sphere and a point $\boldsymbol{v}$ on the plane intersects the unit sphere in the point $\psi(\boldsymbol{v})$. In the picture it can also be seen how the two surface elements are related. For $\boldsymbol{v}$ near the unit circle on the plane, the two surface elements have the same size. If $\boldsymbol{v} = \mathbf{0}$, then the surface element $\mathrm{d}A_{\mathcal{S}^2}$ is four times the size of $\mathrm{d}A_{\mathbb{R}^2}$. As $\|\boldsymbol{v}\| \to \infty$ the ratio $\mathrm{d}A_{\mathcal{S}^2}/\mathrm{d}A_{\mathbb{R}^2}$

Figure 3.3 the relation between the Monge-Ampère equation, the maps $\nabla u, \psi, y$ and integration by substitution is represented in a diagram. We can now reformulate Problem 3.5.1 in terms of $\mathcal{F}$, F and a new Monge-Ampère equation

$$\frac{E(x)}{F(\nabla u(x))} = \frac{|\det(H_{ij})|}{e}.$$

An important theorem by Brenier [11, p.66] states that to this reformulated problem exists a unique convex solution u. We know that for a convex function the Hessian matrix (H_{ij}) is positive semi-definite. For a 2×2 matrix (H_{ij}) this implies that its determinant must be positive, because it equals the product of its eigenvalues. We will try to find Brenier's convex solution u and because the fraction $eE(x)/F(\nabla u(x)$ is positive by definition we will omit the absolute value bars around $\det(H_{ij})$. To be more precise, we will reformulate Problem 3.5.1 as the following problem.

Problem 3.5.4. Let $\mathcal{E} \subset \mathbb{R}^2$ be convex, closed and bounded, and let $\mathcal{F} \subset \mathbb{R}^2$ be closed. Furthermore, let $E : \mathcal{E} \to \mathbb{R}_{>0}$ and $F : \mathcal{F} \to \mathbb{R}_{>0}$ be such that they satisfy (3.20). Find $u \in C^2(\mathcal{E})$ such that u satisfies the Monge-Ampère equation

$$\frac{E(x)}{F(\nabla u(x))} = \frac{\det(H_{ij})}{e} \tag{3.21}$$

and, moreover, $\nabla u(\mathcal{E}) = \mathcal{F}$.

To this problem, the theorem by Brenier asserts, exists a unique convex solution. The functions E and F are strict positive functions and therefore $\det(H_{ij}) > 0$ on $\mathcal{E}$. Suppose for a moment that we work in Cartesian coordinates, then, if u is a convex function, the matrix (H_{ij}) is positive semi-definite. A 2×2 matrix is positive semi-definite if and only if the trace of (H_{ij}) is greater than or equal to zero, i.e. $\mathrm{Tr}(H_{ij}) \geq 0$, and $\det(H_{ij}) \geq 0$. Similarly, a 2×2 matrix is positive definite if and only if $\mathrm{Tr}(H_{ij}) > 0$ and $\det(H_{ij}) > 0$. Now, if (H_{ij}) is positive semi-definite and $\det(H_{ij}) > 0$, then also $\mathrm{Tr}(H_{ij}) > 0$. This follows immediate when we diagonalise (H_{ij}). Let u be the convex solution to Problem 3.5.4. The solution u is a convex function on a convex domain and such a function it holds that (H_{ij}) is positive semi-definite. Now, because, $\det(H_{ij}) > 0$ it follows that (H_{ij}) is even positive definite and therefore u is strictly convex. Thus the convex solution u to (3.5.4) is also strictly convex.

Before we proceed to the next chapter, where we will present a numerical method to find the convex solution u, we will first argue that we can replace the implicit boundary condition $\nabla u(\mathcal{E}) = \mathcal{F}$ by the explicit boundary condition $\nabla u(\partial\mathcal{E}) = \partial\mathcal{F}$. In Lemma 3.2.1 we showed that the map $y = \psi \circ \nabla u$ is a bijection and in Lemma 3.5.2 we showed that the map ψ is a bijection. From this if follows that the map $\nabla u : \mathcal{E} \to \mathcal{F}$ is also a bijection. We will now show that for the strictly convex solution u to Problem 3.5.4, the map ∇u is a *homeomorphism* from $\mathcal{E}$ to $\mathcal{F}$. A homeomorphism is a continuous bijection with continuous inverse. A map that maps open sets to open sets is called *open*. If an open continuous map is a bijection then it is a homeomorphism. This implies that of ∇u is open it is also a homeomorphism.

Lemma 3.5.5. *Suppose that $u \in C^2(\mathcal{E})$ is the strictly convex solution to Problem 3.5.4. Then the map ∇u is also open, i.e. for each open subset V of $\mathcal{E}$ the image $\nabla u(V)$ is an open subset of $\mathcal{F}$.*

Proof. As we just argued, for the strictly convex solution $u \in C^2(\mathcal{E})$, the map ∇u is a bijection. Moreover, because u is twice continuously differentiable, the mapping ∇u is a continuously differentiable mapping. In Cartesian coordinates, the matrix (H_{ij}) is also the Jacobian matrix of ∇u. The fact that $\det(H_{ij}) > 0$ therefore implies that the Jacobian of ∇u is always strictly positive. Thus, the conditions for the inverse function theorem [15] are satisfied. The inverse function theorem states (among other things) that for every open set E of $\mathcal{E}$ and $x_0 \in E$, there exists an open set U in E containing x_0, and an open set V in $\mathcal{F}$ containing $\nabla u(x_0)$ such that ∇u is a bijection from U to V and the inverse $(\nabla u)^{-1}$ is continuously differentiable on V.

From this it follows that ∇u is open. To see this suppose E is some open set in $\mathcal{E}$. By the inverse function theorem there exists for every $x \in \mathcal{E}$ open sets U_x and V_x such that $x \in U_x$, $\nabla u(x) \in V_x$ and $U_x \subset E$. $\nabla u(U_x) = V_x$ is open for every $x \in E$. Notice that $\cup_{x\in E} U_x = E$ and that $\nabla u(E) = \cup_{x\in E} \nabla u(U_x) = \cup_{x\in E} V_x$. Thus, we see that $\nabla u(E)$ is a union of open sets and hence open. We have established that for every open subset E of $\mathcal{E}$ the image $\nabla u(E)$ is an open subset of $\mathcal{F}$, i.e. ∇u is an open map. □

Thus, we see that ∇u is a homeomorphism from $\mathcal{E}$ to $\mathcal{F}$. We will use this convenient property of ∇u in the following lemma.

Lemma 3.5.6. *Let u be the strictly convex solution to Problem 3.5.4, then $\nabla u(\partial\mathcal{E}) = \partial\mathcal{F}$.*

Proof. The map ∇u is a homeomorphism and therefore it links every open map in $\mathcal{E}$ with an open map in $\mathcal{F}$. Let us by U° denote the interior of a set U. Suppose $A \subset \mathcal{E}$. It is obvious that $\nabla u(A^\circ) \subset \nabla u(A)$. However, because ∇u is an open map $\nabla u(A^\circ)$ is also open. The largest open subset of $\nabla u(A)$ is the interior $\nabla u(A)^\circ$, therefore we have $\nabla u(A^\circ) \subset \nabla u(A)^\circ$. If $\nabla u : \mathcal{E} \to \mathcal{F}$ is a homeomorphism, then $(\nabla u)^{-1} : \mathcal{F} \to \mathcal{E}$ is a homeomorphism also. This implies that we also have $(\nabla u)^{-1}(B^\circ) = (\nabla u)^{-1}(B)^\circ$ for all $B \subset \mathcal{F}$.

From this it follows that we have both $\nabla u(\mathcal{E}^\circ) \subset \nabla u(\mathcal{E})^\circ = \mathcal{F}^\circ$ and $(\nabla u)^{-1}(\mathcal{F}^\circ) \subset (\nabla u)^{-1}(\mathcal{F})^\circ = \mathcal{E}^\circ$. Using this we see that

$$\mathcal{F}^\circ = \nabla u\left((\nabla u)^{-1}(\mathcal{F}^\circ)\right) \subset \nabla u\left(\mathcal{E}^\circ\right) \subset \mathcal{F}^\circ.$$

Thus, we see that $\nabla u\left(\mathcal{E}^\circ\right) = \mathcal{F}^\circ$. Now, because ∇u is a bijection this implies that we must have $\nabla u(\partial\mathcal{E}) = \partial\mathcal{F}$. □

Thus the strictly convex solution of Problem 3.5.4 is also a solution to the same problem but with the implicit boundary condition $\nabla u(\mathcal{E}) = \mathcal{F}$ replaced by the explicit boundary condition $\nabla u(\partial\mathcal{E}) = \partial\mathcal{F}$, i.e. the following problem.

Problem 3.5.7. Let $\mathcal{E} \subset \mathbb{R}^2$ be convex, closed and bounded, and let $\mathcal{F} \subset \mathbb{R}^2$ be closed. Furthermore, let $E : \mathcal{E} \to \mathbb{R}_{>0}$ and $F : \mathcal{F} \to \mathbb{R}_{>0}$ be such that they satisfy (3.20). Find $u \in C^2(\mathcal{E})$ such that u satisfies the Monge-Ampère equation

$$\frac{E(x)}{F(\nabla u(x))} = \frac{\det(H_{ij})}{e}$$

and, moreover, $\nabla u(\partial\mathcal{E}) = \partial\mathcal{F}$.

Thus, a strictly convex solution of Problem 3.5.4 is also a solution to Problem 3.5.7. Now the following lemma states the converse.

Lemma 3.5.8. *Let u be a strictly convex solution to Problem 3.5.7. Then $\nabla u(\mathcal{E}) = \mathcal{F}$.*

Proof. The map ∇u is a homeomorphism from $\mathcal{E}$ to $\nabla u(\mathcal{E}) \subset \mathbb{R}^2$. The set $\mathcal{E}$ is convex and hence simply connected. The set $\partial\mathcal{E}$ is a simple and closed curve, i.e. a Jordan curve. The map ∇u is continuous and injective and hence $\nabla u(\partial\mathcal{E}) = \partial\mathcal{F}$ is a Jordan curve also. Now the Jordan curve theorem states that the complement $\mathbb{R}^2\backslash\partial\mathcal{F}$ has two connected components one of which is bounded and one of which is not, namely the interior and the exterior of the curve, and the boundary of both these sets is $\partial\mathcal{F}$. The set $\mathcal{E}$ is simply connected and simply connectedness is a topological property, therefore $\nabla u(\mathcal{E})$ is simply connected also. The interior and exterior to the curve $\nabla u(\mathcal{E}) = \partial\mathcal{F}$ are the only two subsets of $\mathbb{R}^2$ with $\partial\mathcal{F}$ as boundary. The map ∇u is a homeomorphism and therefore we have $\nabla u(\partial\mathcal{E}) = \partial(\nabla u(\mathcal{E}))$. This follows from the fact that $\nabla u(\mathcal{E}^\circ) = (\nabla u(\mathcal{E}))^\circ$ what we showed in the proof of Lemma 3.5.6. The fact that $\partial\mathcal{F} = \nabla u(\partial\mathcal{E}) = \partial(\nabla u(\mathcal{E}))$ implies that $\nabla u(\mathcal{E})^\circ$ is one of two sets of the Jordan curve theorem. The exterior set is clearly not simply connected, while $\nabla u(\mathcal{E})$ is, therefore $\nabla u(\mathcal{E})^\circ$ is the interior set in the Jordan curve theorem. Equation (3.20), the fact that $\mathcal{E}$ is bounded and the functions E and F are strictly positive imply that the set $\mathcal{F}$ is bounded also. This implies that $\mathcal{F}^\circ$ needs to be the interior set also and hence we find that $\nabla u(\mathcal{E}) = \mathcal{F}$. □

We have established that u is a strictly convex solution of Problem 3.5.4 if and only if u is a strictly convex solution of Problem 3.5.7. Thus to find the unique strictly convex solution to Problem 3.5.4 we can just as well try to find the strictly convex solution of Problem 3.5.7 and this is what we will do in next Chapter.

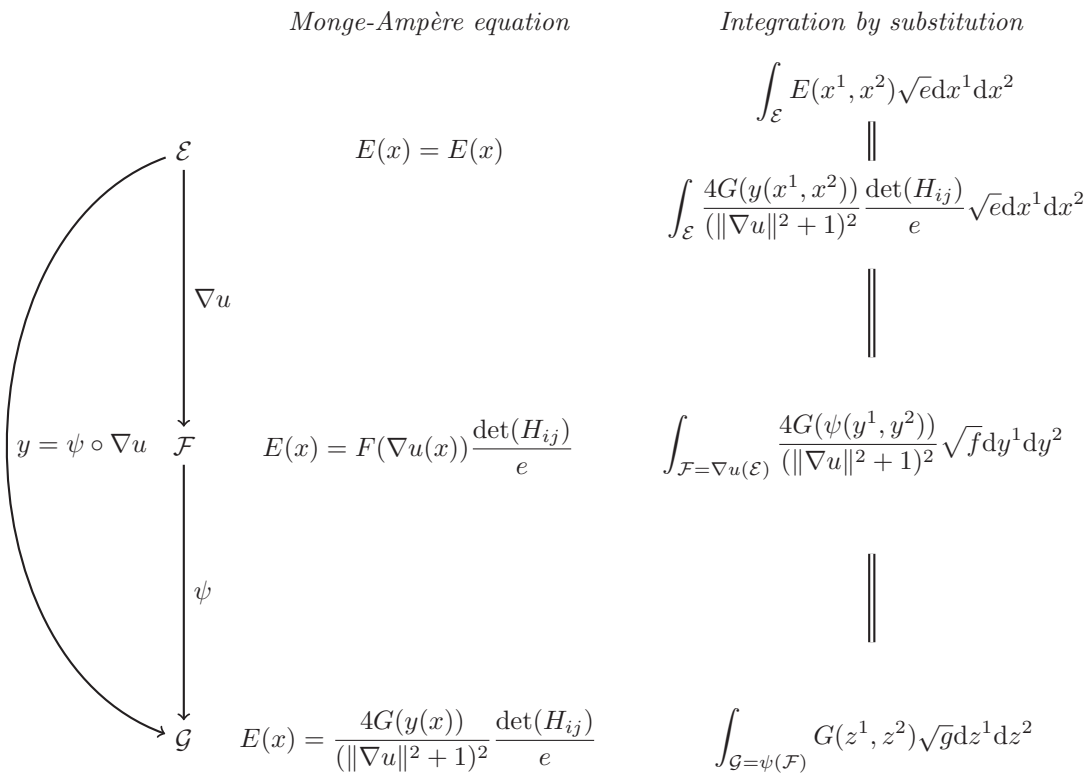


Figure 3.3: This is a graphical representation of the formation of the Monge-Ampère equation by successive integration by substitution. One starts out with the integral of G over $\mathcal{G}$ and applies integration by substitution to end up with an integral over $\mathcal{F}$. The integrand contains the original functions times a Jacobian for the map ψ. Subsequently, one applies integration by substitution again and one ends up with an integral over $\mathcal{E}$, while the integrand is multiplied by the second Jacobian for the map ∇u. This integral over $\mathcal{E}$ has to equal the emittance of $\mathcal{E}$ by conservation of energy.

3.6 The output intensity

In practice the output intensity of the reflector system will most often be given in spherical coordinates, i.e. set $\mathcal{G} \subset \mathcal{S}^2$ and the intensity function G will be given in spherical coordinates. However it could also be that some desired intensity pattern on a projection screen at a distance d from the reflector is specified. We will denote the intensity function on the plane describing this pattern by H and the subset of this plane for which $H > 0$ by $\mathcal{H}$. In this section we will discuss how to determine the pair $(\mathcal{F}, F)$ in Problem 3.5.7 from the pair $(\mathcal{G}, G)$ or the pair $(\mathcal{H}, H)$.

3.6.1 Output intensity specified in spherical coordinates

Suppose we have a output intensity on $\mathcal{S}^2$ specified by $(\mathcal{G}, G)$ given in a spherical coordinate system. In order to define the spherical coordinate system we need two perpendicular directions. We use the direction $\tilde{\boldsymbol{e}}_3$ as the zenith direction from which we measure the polar angle. Furthermore we need one direction in the plane from which we measure the azimuth angle. This direction we denote by $\tilde{\boldsymbol{a}}$. Once a coordinate system is chosen on $\mathcal{F}$ we will relate $\tilde{\boldsymbol{a}}$ to a basis vector in this coordinate system. So, for example, when a Cartesian coordinate system is used on $\mathcal{F}$ with basis vectors $\boldsymbol{e}_x$ and $\boldsymbol{e}_y$, one often takes $\tilde{\boldsymbol{a}} = \boldsymbol{e}_x$.

Given a certain vector $\boldsymbol{v} \in \mathcal{S}^2$ the corresponding spherical coordinates are then given by

$$\theta = \tan^{-1}\left(\frac{\sqrt{1-(\boldsymbol{v} \mid \tilde{\boldsymbol{e}}_3)^2-(\boldsymbol{v} \mid \tilde{\boldsymbol{a}})^2}}{(\boldsymbol{v} \mid \tilde{\boldsymbol{a}})}\right) \quad \text{and} \quad \phi = \arccos((\boldsymbol{v} \mid \tilde{\boldsymbol{e}}_3)), \tag{3.22}$$

where $\tan^{-1}$ is the function defined in Example 2.1.6. Substituting $\tilde{\boldsymbol{y}} = \psi(\boldsymbol{v})$ given by (3.13) in (3.22) we obtain ϕ and θ as a function of the vector $\boldsymbol{v} \in \mathcal{F}$:

$$\theta(\boldsymbol{v}) = \tan^{-1}\left(\frac{\sqrt{4\|\boldsymbol{v}\|^2-[(\|\boldsymbol{v}\|^2+1)(\tilde{\boldsymbol{y}} \mid \tilde{\boldsymbol{a}})]^2}}{(\|\boldsymbol{v}\|^2+1)(\tilde{\boldsymbol{y}} \mid \tilde{\boldsymbol{a}})}\right). \tag{3.23}$$

$$\phi(\boldsymbol{v}) = \arccos((\tilde{\boldsymbol{y}} \mid \tilde{\boldsymbol{e}}_3)) = \arccos\left(\frac{\|\boldsymbol{v}\|^2-1}{\|\boldsymbol{v}\|^2+1}\right), \tag{3.24}$$

If we use a Cartesian coordinate system on $\mathcal{F}$ and align $\tilde{\boldsymbol{a}}$ with $\boldsymbol{e}_x$ we find for $\boldsymbol{v} = v^x\boldsymbol{e}_x + v^y\boldsymbol{e}_y$, (3.16) and (3.17). Thus, when we use the Cartesian coordinate system on $\mathcal{F}$, equation (3.18) implies that F is given by

$$F(v^x, v^y) = \frac{4G(\theta(v^x, v^y), \phi(v^x, v^y))}{((v^x)^2+(v^y)^2+1)^2}, \tag{3.25}$$

with $\theta(v^x, v^y)$ and $\phi(v^x, v^y)$ as given in equations (3.16) and (3.17), respectively. However to be able to determine $\partial\mathcal{F}$ from $\partial\mathcal{G}$ we also need to invert relations (3.16) and (3.17). This is done in [5, p.77]. We will just state that result here:

$$v^x(\theta, \phi) = \frac{\sin(\phi)\cos(\theta)}{1-\cos(\phi)} \quad \text{and} \quad v^y(\theta, \phi) = \frac{\sin(\phi)\sin(\theta)}{1-\cos(\phi)}.$$

If we use polar coordinates on $\mathcal{F}$ defined by their usual relations with the Cartesian coordinate system we find

$$\theta(v^r, v^\theta) = v^\theta, \tag{3.26}$$

$$\phi(v^r, v^\theta) = \arccos\left(\frac{(v^r)^2-1}{(v^r)^2+1}\right). \tag{3.27}$$

From this we see that the function F is given by

$$F(v^r, v^\theta) = \frac{4G(\theta(v^r, v^\theta), \phi(v^r, v^\theta))}{((v^r)^2+1)^2}, \tag{3.28}$$

with $\theta(v^r, v^\theta)$ and $\phi(v^r, v^\theta)$ as given in equations (3.26) and (3.27), respectively. The relations (3.26) and (3.27) are easily inverted. Doing this we find that

$$v^r(\theta,\phi) = \sqrt{\frac{1+\cos(\phi)}{1-\cos(\phi)}} \quad \text{and} \quad v^\theta(\theta,\phi) = \theta.$$

These relations allow us to determine $\mathcal{F}$ once $\mathcal{G} \subset \mathcal{S}^2$ is given.

3.6.2 Target illuminance specified on a target plane

Suppose now that we want a certain illuminance [lm/m^2] on a plane at a distant d of the reflector in the direction of $\tilde{\boldsymbol{a}}$. This plane is also perpendicular to the direction of $\tilde{\boldsymbol{a}}$. Let us denote the illuminance by the function $H : \mathcal{H} \to \mathbb{R}_{>0}$, where $\mathcal{H}$ is some subset of the plane. For simplicity we will assume that we have a Cartesian coordinate system on $\mathcal{H}$. Let the basis vectors be given by the vectors $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ which correspond to coordinates h^1 and h^2. Let the vector $\boldsymbol{h}_2 = \tilde{\boldsymbol{e}}_3$ and let $\boldsymbol{h}_1 = -\tilde{\boldsymbol{e}}_3 \times \tilde{\boldsymbol{a}}$. We will now calculate the coordinates (h^1, h^2) on the plane at which a light ray in the direction specified by (θ, ϕ) intersects the plane. This gives us a map from the directions of $\mathcal{S}^2$ to the points on $\mathcal{H}$. We find that this map is given by

$$h^1(\theta,\phi) = -d\tan(\theta) \quad \text{and} \quad h^2(\theta,\phi) = \frac{\cos(\theta)\tan(\pi/2-\phi)}{d}. \tag{3.29}$$

The Jacobian matrix of this map is given by

$$J(h^1(\theta,\phi), h^2(\theta,\phi)) = \begin{pmatrix} \dfrac{\partial h^1}{\partial\theta} & \dfrac{\partial h^1}{\partial\phi} \\ \dfrac{\partial h^2}{\partial\theta} & \dfrac{\partial h^2}{\partial\phi} \end{pmatrix} = \begin{pmatrix} \dfrac{-d}{\cos^2(\theta)} & 0 \\ \dfrac{d\tan(\pi/2-\phi)\tan(\theta)}{\cos(\theta)} & \dfrac{-d}{\sin^2(\phi)\cos(\theta)} \end{pmatrix}.$$

We see that as long as $\theta \in (-\pi, \pi)$ and $\phi \in (0, \pi)$ the Jacobian is invertible and its determinant is given by

$$\det(J(h^1(\theta,\phi), h^2(\theta,\phi))) = \frac{d^2}{\cos^3(\theta)\sin^2(\phi)}. \tag{3.30}$$

Furthermore the map $(\theta, \phi) \mapsto (h^1(\theta,\phi), h^2(\theta,\phi)$ is continuously differentiable for $\theta \in (-\pi, \pi)$ and $\phi \in (0, \pi)$, therefore the inverse function theorem applies. The inverse function theorem states that the inverse map $(h^1, h^2) \mapsto (\theta(h^1, h^2), \phi(h^1, h^2))$ is also continuously differentiable and its jacobian matrix is given by

$$J(\theta(h^1,h^2), \phi(h^1,h^2)) = [J(h^1(\theta,\phi), h^2(\theta,\phi))]^{-1}.$$

By the change of variables formula it follows therefore that

$$H(h^1,h^2)\sqrt{h(h^1,h^2)}\ \mathrm{d}h^1\mathrm{d}h^2 = H(h^1(\theta,\phi), h^2(\theta,\phi))\ |\det(J(\theta(h^1,h^2),\phi(h^1,h^2)))|\sqrt{g(\theta,\phi)}\ \mathrm{d}\theta\mathrm{d}\phi.$$

The square root of the Cartesian metric is one and the square root of the metric of the spherical coordinate basis is equal to $\sin(\phi)$, therefore, using (3.30), we find

$$H(h^1,h^2)\ \mathrm{d}h^1\mathrm{d}h^2 = H(h^1(\theta,\phi), h^2(\theta,\phi))\left(\frac{\cos^3(\theta)\sin^2(\phi)}{d^2}\right)\sin(\phi)\ \mathrm{d}\theta\mathrm{d}\phi,$$

because the determinant of the inverse of a matrix is equal to the inverse of the determinant of the matrix. Now we define the function $G : \mathcal{G} \to \mathbb{R}_{>0}$ by

$$G(\theta,\phi) := H(h^1(\theta,\phi), h^2(\theta,\phi))\left(\frac{\cos^3(\theta)\sin^2(\phi)}{d^2}\right),$$

where $\mathcal{G} := \{(\theta(h^1, h^2), \phi(h^1, h^2)) \in \mathcal{S}^2 \mid (h^1, h^2) \in \mathcal{H}\}$.

In order to calculate the set $\mathcal{G}$ from $\mathcal{H}$ we need to have a formula for $(h^1, h^2) \mapsto (\theta(h^1, h^2), \phi(h^1, h^2))$. Note that $\cos(\theta) = d^2/\sqrt{(h^1)^2 + d^2}$. From this and equation (3.29) it follows that

$$\theta(h^1, h^2) = \tan^{-1}\left(\frac{-h^1}{d}\right) \quad \text{and} \quad \phi(h^1, h^2) = \frac{\pi}{2} - \tan^{-1}\left(h^2\sqrt{(h^1)^2 + d^2}\right).$$

Now that we have determined $(\mathcal{G}, G)$ from $(\mathcal{H}, H)$ we can also determine $(\mathcal{F}, F)$ from $(\mathcal{H}, H)$ if we use the results from last subsection to get from $(\mathcal{G}, G)$ to $(\mathcal{F}, F)$.

Chapter 4

Least-Squares Method for Arbitrary Coordinate Systems

In this chapter we will introduce a numerical method to solve the boundary value problem derived in last chapter. This method is called the *least-squares method* and was proposed in [5]. We will present the least-squares method here for an arbitrary coordinate system on the source $\mathcal{E}$. We will first give an outline of the numerical method before presenting the three main steps of the method in more detail in three subsequent sections. One of the three main steps of the method as represented in [5] contains a minor flaw. We will indicate this and present an improved version of this step.

4.1 Outline of the least-squares method

We will present the numerical method for an arbitrary coordinate system on $\mathcal{E}$, with coordinates x^1, x^2, local basis vectors $\boldsymbol{e}_1, \boldsymbol{e}_2$ and a metric $e_{ij} = (\boldsymbol{e}_i | \boldsymbol{e}_j)$. We will not try to solve Problem (3.5.7) directly for u. Instead we will look for a mapping $\boldsymbol{m} = \nabla u : \mathcal{E} \to \mathcal{F}$ that

(i) solves the following boundary value problem

$$\begin{aligned} \frac{\det(\nabla \hat{\boldsymbol{m}}(x))}{e} &= \frac{E(x)}{F(\boldsymbol{m}(x))} && x \in \mathcal{E}, \\ \boldsymbol{m}(\partial\mathcal{E}) &= \partial\mathcal{F}, \end{aligned}$$

where $\hat{\boldsymbol{m}} = m_i \boldsymbol{e}^i = e_{ij} m^i \boldsymbol{e}^j$,

(ii) $\boldsymbol{m}$ should be such that there exists a strictly convex $u \in C^2(\mathcal{E})$ such that $\boldsymbol{m} = \nabla u$.

From this mapping we will eventually find u.

We need to make (ii) more precise. In order to say more about this, we will first show that the Hessian tensor is always symmetric for the Levi-Civita connection.

Lemma 4.1.1. *Let M be a twice continuously differentiable submanifold of Euclidean space, endowed with a Levi-Civita connection and let $u \in C^2(M)$. Then the Hessian tensor $\boldsymbol{H}(u)$ is symmetric, i.e.*

$$H_{ij} = H_{ji}.$$

Proof. On page 28 we have shown that

$$\left(\nabla_{\boldsymbol{e}_j}(\nabla_{\boldsymbol{e}_i} u) - \Gamma^k_{ij} \nabla_{\boldsymbol{e}_k}\right) - \left(\nabla_{\boldsymbol{e}_i}(\nabla_{\boldsymbol{e}_j} u) - \Gamma^k_{ji} \nabla_{\boldsymbol{e}_k}\right) = T^k_{ij} \nabla_{\boldsymbol{e}_k} u,$$

where T_{ij}^k are the components of the torsion tensor. By Definition 2.4.6, the Levi-Civita connection is torsion-free, i.e. all the components $T_{ij}^k = 0$. Thus, we can conclude that the Hessian tensor is indeed symmetric. Note that being symmetric for a tensor is a property independent of the coordinate system. This we have shown on page 28 also. □

We see that in order for $\boldsymbol{m}$ to satisfy (ii), the tensor

$$\nabla \hat{\boldsymbol{m}} = \nabla_{\boldsymbol{e}_j}(\hat{\boldsymbol{m}}) \otimes \hat{\boldsymbol{e}}^j = (\nabla_{\boldsymbol{e}_j} m_i - \Gamma_{ij}^k m_k)\hat{\boldsymbol{e}}^i \otimes \hat{\boldsymbol{e}}^j$$

needs to be symmetric. This condition is actually enough to ensure that $\boldsymbol{m}$ equals the gradient of some function. The symmetry of $\nabla \hat{\boldsymbol{m}}$ implies that the curl of $\boldsymbol{m}$ is zero. Let us interpret $\boldsymbol{m} = m^i \boldsymbol{e}_i = m_i e^{ij} \boldsymbol{e}_j$, where summation runs over $i, j = 1, 2$ (Latin indices), as a vector in $\mathbb{R}^3$ and calculate its curl. In an arbitrary coordinate system the curl of a vector field $\boldsymbol{v}$ is given by

$$\nabla \times \boldsymbol{v} := \epsilon_{\alpha\beta\gamma} e^{\beta\delta} D_\delta(v^\alpha) e^{\gamma\sigma} \boldsymbol{e}_\sigma,$$

where $D_\delta(v^\beta)$ are the components of the covariant derivative of $\boldsymbol{v}$.

For the basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \tilde{\boldsymbol{e}}_3\}$, the third basis vector $\tilde{\boldsymbol{e}}_3$ does not depend on position as it is by definition always normal to the plane spanned by $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ and of unit length. Furthermore $\boldsymbol{m}$ is a vector in the plane of $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ and therefore $m^3 = 0$, hence we find

$$\begin{aligned}
\nabla \times \boldsymbol{m} &= \epsilon_{\alpha\beta\gamma} e^{\beta\delta} e^{\alpha\sigma} D_\delta(m_\sigma) e^{\gamma\rho} \boldsymbol{e}_\rho \\
&= \epsilon_{123} e^{11} e^{22} e^{33} (\nabla_{\boldsymbol{e}_2} m_1 - \Gamma_{12}^i m_i) + \epsilon_{213} e^{11} e^{22} e^{33} (\nabla_{\boldsymbol{e}_1} m_2 - \Gamma_{21}^i m_i) \\
&= \frac{(\nabla_{\boldsymbol{e}_2} m_1 - \Gamma_{12}^i m_i) - (\nabla_{\boldsymbol{e}_1} m_2 - \Gamma_{21}^i m_i)}{\sqrt{e}}.
\end{aligned}$$

From this we see that $\nabla \times \boldsymbol{m}$ vanishes if and only if $\nabla \hat{\boldsymbol{m}}$ is symmetric. A vector field with zero curl is called a *conservative field*. Conservative fields on a simply connected domain always equal the gradient of some function, see for example [13, p.551]. Thus we can conclude that $\boldsymbol{m} \in T\mathcal{E}_{C^1}$ equals the gradient of some function $u \in C^2(\mathcal{E})$ if and only if $\nabla \hat{\boldsymbol{m}}$ is symmetric. (Recall from page 21 that $T\mathcal{E}_{C^1}$ is the space of continuously differentiable vector fields on $\mathcal{E}$.)

However, this condition alone will not suffice for our goals, because we also need u to be strictly convex. The function $u \in C^2(\mathcal{E})$ is convex if and only if $\mathcal{E}$ is convex and the Hessian tensor $\boldsymbol{H}(u)$ is *positive semi-definite*, see for example [14, p.71]. The Hessian tensor is positive semi-definite if and only if for every $\boldsymbol{x} = x^i \boldsymbol{e}_i$ we have $H(u)(\boldsymbol{x}, \boldsymbol{x}) \geq 0$, where

$$\begin{aligned}
\boldsymbol{H}(u)(\boldsymbol{x}, \boldsymbol{x}) &= \big(\nabla_{\boldsymbol{e}_j}(\nabla_{\boldsymbol{e}_i} u) - \Gamma_{ij}^k \nabla_{\boldsymbol{e}_k} u\big) \langle \hat{\boldsymbol{e}}^i, \boldsymbol{x} \rangle \langle \hat{\boldsymbol{e}}^j, \boldsymbol{x} \rangle \\
&= H_{ij} x^i x^j \\
&= x_k e^{ki} H_{ij} x^j \\
&= \boldsymbol{x}^T (e^{ki} H_{ij}) \boldsymbol{x}.
\end{aligned}$$

From this we see that $\boldsymbol{H}(u)$ is positive semi-definite if and only if the matrix $(e^{ki} H_{ij})$ is positive semi-definite. For our orthogonal basis the metric is diagonal and therefore

$$(e^{ki} H_{ij}) = \begin{pmatrix} e^{11} H_{11} & e^{11} H_{12} \\ e^{22} H_{21} & e^{22} H_{22} \end{pmatrix}. \tag{4.1}$$

Unfortunately, we can not demand positive definiteness, because, although every $u \in C^2(\mathcal{E})$ with positive definite Hessian tensor is strictly convex, not every strictly convex $u \in C^2(\mathcal{E})$ has a positive definite Hessian tensor.* Thus we cannot ask more than for $\nabla \hat{\boldsymbol{m}}$ to be positive semi-definite, because this would be too restrictive. The numerical method that we will soon start introducing will solve the following boundary value problem.

*Consider for example the strictly convex function $f(x) = x^4$ on the real line. Although f is strictly convex, the Hessian tensor, i.e. f'', is zero for $x = 0$ and hence not positive definite.

Problem 4.1.2. Find $\boldsymbol{m} \in T\mathcal{E}_{C^1}$ that satisfies

$$\frac{\det(\nabla \hat{\boldsymbol{m}}(x))}{e} = \frac{E(x)}{F(\boldsymbol{m}(x))}, \qquad x \in \mathcal{E}, \tag{4.2a}$$

$$\boldsymbol{m}(\partial\mathcal{E}) = \partial\mathcal{F}, \tag{4.2b}$$

and for which $\nabla\hat{\boldsymbol{m}}$ is a symmetric positive semi-definite tensor. In this problem the functions E and F are strictly positive functions such that

$$\int_{\mathcal{E}} E(x^1, x^2)\sqrt{e}\ \mathrm{d}x^1\mathrm{d}x^2 = \int_{\mathcal{F}} F(y^1, y^2)\sqrt{f}\ \mathrm{d}y^1\mathrm{d}y^2,$$

where x^1, x^2 are local coordinates on $\mathcal{E}$ with corresponding metric e_{ij} and y^1, y^2 are local coordinates on $\mathcal{F}$ with corresponding metric f_{ij}.

It is clear that if u is a solution to Problem 3.5.7, then $\boldsymbol{m} = \nabla u$ will be a solution to Problem 4.1.2. The reverse statement is not true because a solution $\boldsymbol{m}$ of Problem 4.1.2 may be such that the u in $\nabla u = \boldsymbol{m}$ is convex but not strictly convex. Problem 4.1.2 allows also for convex solutions. For convex but not strictly convex $u \in C^2(\mathcal{E})$ the map $x \mapsto \tilde{\boldsymbol{y}}(x)$ in Lemma 3.2.1 is no longer a bijection. However, in practice, due to numerical inaccuracies, solutions u that are convex but not strictly convex do not occur and hence this is not a serious problem.

We will numerically solve Problem 4.1.2 by starting with an initial guess $\boldsymbol{m}^0$ and improving this initial guess in an iterative manner. We will try to find a solution $\boldsymbol{m}$ satisfying equation (4.2a) by minimizing the functional

$$J_I(\boldsymbol{m}, \boldsymbol{P}) := \frac{1}{2}\iint_{\mathcal{E}} \|\nabla\hat{\boldsymbol{m}} - \boldsymbol{P}\|^2\sqrt{e}\mathrm{d}x^1\mathrm{d}x^2 \tag{4.3}$$

over the set

$$\mathcal{P}(\boldsymbol{m}) := \left\{\boldsymbol{P} \in \boldsymbol{T}_0^2(T\mathcal{E})_{C^1} \mid [\det(P_{ij}(x)) = eE(x)/F(\boldsymbol{m}(x)),\ \boldsymbol{P}(x) \text{ is spsd}]\right\},$$

where "spsd" stands for symmetric positive semi-definite. It seems as if we demand more smoothness than necessary because for Problem 4.1.2 we only need $\boldsymbol{P}$ to be continuous for continuous E and F. However, in one of the minimization procedures we need $\nabla\hat{\boldsymbol{m}}$ to be continuously differentiable and therefore we also need $\boldsymbol{P}$ to be continuously differentiable.

The norm in equation (4.3) is defined in the following way. Let $\boldsymbol{A}, \boldsymbol{B} \in \boldsymbol{T}_0^2(T_x\mathcal{E})$, i.e., the tangent space of T_0^2-tensors in the point $x \in \mathcal{E}$, then $\boldsymbol{A} : \boldsymbol{B} := e^{ik}e^{jl}A_{ij}B_{kl} = A_{ij}B^{ij}$ defines an inner product on $\boldsymbol{T}_0^2(T_x\mathcal{E})$. This inner product is the inner product on $\boldsymbol{T}_0^2(T_x\mathcal{E})$ induced by the metric. The fact that this is indeed an inner product follows by the symmmetry, linearity and positivity of the metric $\boldsymbol{e}$. Let $\|\cdot\|$ be the norm associated with this inner product.* It is clear that if $J_I = 0$, $\boldsymbol{m}$ will satisfy equation (4.2a) and $\nabla\hat{\boldsymbol{m}}$ will be symmetric positive semi-definite.

To satisfy the boundary condition (4.2b) we will minimize another functional simultaneously. This functional is given by

$$J_B(\boldsymbol{m}, \boldsymbol{b}) := \frac{1}{2}\oint_{\partial\mathcal{E}} \|\boldsymbol{m} - \boldsymbol{b}\|^2\mathrm{d}s. \tag{4.4}$$

We will minimize this functional over the space

$$\mathcal{B} := \{\boldsymbol{b} \in T\mathcal{E}_C \mid \boldsymbol{b}(x) \in \partial\mathcal{F}\}. \tag{4.5}$$

Analogously to the functional J_I we notice that if $J_B = 0$, $\boldsymbol{m}$ satisfies equation (4.2b).

*We use the same notation as for the vector norm, but this is not very likely to cause confusion because it will be clear from the argument which norm we mean.

Our goal is to minimize J_I and J_B simultaneously. In order to do that we define a third functional:

$$J(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b}) := \alpha J_I(\boldsymbol{m}, \boldsymbol{P}) + (1-\alpha) J_B(\boldsymbol{m}, \boldsymbol{b}) \tag{4.6}$$

with $\alpha \in (0,1)$. This functional we will minimize for $\boldsymbol{m}$ over

$$\mathcal{V} := T\mathcal{E}_{C^2}. \tag{4.7}$$

One iteration of the numerical method consists of three steps. Assume that $\boldsymbol{m}^n$ is given. In order to determine $\boldsymbol{m}^{n+1}$ we subsequently perform three steps:

$$\boldsymbol{b}^{n+1} = \underset{\boldsymbol{b}\in\mathcal{B}}{\operatorname{argmin}} J_B(\boldsymbol{m}^n, \boldsymbol{b}), \tag{4.8a}$$

$$\boldsymbol{P}^{n+1} = \underset{\boldsymbol{P}\in\mathcal{P}(\boldsymbol{m}^n)}{\operatorname{argmin}} J_I(\boldsymbol{m}^n, \boldsymbol{P}), \tag{4.8b}$$

$$\boldsymbol{m}^{n+1} = \underset{\boldsymbol{m}\in\mathcal{V}}{\operatorname{argmin}} J(\boldsymbol{m}, \boldsymbol{P}^{n+1}, \boldsymbol{b}^{n+1}). \tag{4.8c}$$

We will in the next three sections focus on each of these minimization steps. We will treat (4.8b) quite extensively, because this minimization procedure contains an improvement with respect to the method as presented in [5]. In [5] the tensor $\boldsymbol{P}$ was not required to be positive semi-definite, while this is necessary to ensure the convexity of the reflector surface. This imperfection in the numerical method as presented there gives rise to some convergence issues. We will show that, just as in [5], it is still possible to solve the minimization problem (4.8b) algebraically, despite the extra condition on $\boldsymbol{P}$. We will also cover minimization problem (4.8c) in a lot of detail, because this minimization problem becomes somewhat more involved for arbitrary coordinate systems. We will now start with minimization problem (4.8a).

4.2 Minimization of J_B

In the first step we minimize J_B for fixed $\boldsymbol{m}$ over the space $\mathcal{B}$ of possible continuous vector fields $\boldsymbol{b}$ on $\partial\mathcal{E}$ that map to $\partial\mathcal{F}$. In order to solve this minimization problem we will make a linear approximation of the boundary $\partial\mathcal{F}$ by N_b straight line segments. We use N_b grid points $\boldsymbol{y}_i \in \partial\mathcal{F}$ and connect these by straight lines. We number the $\boldsymbol{y}_i$ in increasing anti-clockwise direction such that $\boldsymbol{y}_{N_b+1} = \boldsymbol{y}_1$. We will then for each boundary grid point $x \in \partial\mathcal{E}$ minimize

$$\|\boldsymbol{m}(x) - \boldsymbol{b}(x)\|^2.$$

Let us denote the line segment between $\boldsymbol{y}_i$ and $\boldsymbol{y}_{i+1}$ by $(\boldsymbol{y}_i, \boldsymbol{y}_{i+1})$. We consider an arbitrary grid point on $\partial\mathcal{E}$ and determine the nearest point to $\boldsymbol{m}(x)$ on the line segment $(\boldsymbol{y}_i, \boldsymbol{y}_{i+1})$ by calculating the projection $\boldsymbol{m}^P(x)$ of $\boldsymbol{m}(x)$ on the line through $\boldsymbol{y}_i$ and $\boldsymbol{y}_{i+1}$. This projection is given by

$$\boldsymbol{m}^P(x) = \boldsymbol{y}_i + t_i(\boldsymbol{y}_{i+1} - \boldsymbol{y}_i),$$

where the parameter t_i is given by

$$t_i = \frac{(\boldsymbol{m}(x) - \boldsymbol{y}_i | \boldsymbol{y}_{i+1} - \boldsymbol{y}_i)}{\|\boldsymbol{y}_{i+1} - \boldsymbol{y}_i\|^2}. \tag{4.9}$$

If the parameter $t_i \in [0,1]$, then $\boldsymbol{m}^P(x)$ lies on the line segment $(\boldsymbol{y}_i, \boldsymbol{y}_{i+1})$. If, however, $t_i > 1$, then we take $\boldsymbol{y}_{i+1}$ to be point on $(\boldsymbol{y}_i, \boldsymbol{y}_{i+1})$ closest to $\boldsymbol{m}(x)$ and similarly if $t_i < 0$ then we take $\boldsymbol{y}_i$ to be the point on $(\boldsymbol{y}_i, \boldsymbol{y}_{i+1})$ closest to $\boldsymbol{m}(x)$. Thus the closest point to $\boldsymbol{m}(x)$ on the line segment $(\boldsymbol{y}_i, \boldsymbol{y}_{i+1})$ is given by

$$\boldsymbol{b}_i(x) = \boldsymbol{y}_i + \min(1, \max(0, t_i))(\boldsymbol{y}_{i+1} - \boldsymbol{y}_i),$$

with t_i given by Equation (4.9). The distance from $\boldsymbol{m}(x)$ to line segment $(\boldsymbol{y}_i, \boldsymbol{y}_{i+1})$ is given by $\|\boldsymbol{m}(x) - \boldsymbol{b}_i(x)\|$. Let now $k \in [1, N_b]$ be the index for which this distance is minimal, i.e.

$$k = \operatorname*{argmin}_{1 \le i \le N_b} \|\boldsymbol{m}(x) - \boldsymbol{b}_i(x)\|.$$

It follows that the point $\boldsymbol{b}(x)$ closest to $\boldsymbol{m}(x)$ on the approximation of $\partial\mathcal{F}$ is given by $\boldsymbol{b} = \boldsymbol{b}_k$ and lies on the line segment $(\boldsymbol{y}_k, \boldsymbol{y}_{k+1})$.

In this way we can for each grid point $x \in \partial\mathcal{E}$ calculate $\boldsymbol{b}(x)$, the nearest point on the approximation of $\partial\mathcal{F}$, analytically. This give us $\boldsymbol{b}$ that minimizes J_B in a discretized sense. We will now turn our attention to the minimization problem (4.8b).

4.3 Minimization of J_I

We will now show how to minimize $J_I(\boldsymbol{m}, \boldsymbol{P})$ for $\boldsymbol{P} \in \mathcal{P}(\boldsymbol{m})$ for fixed $\boldsymbol{m}$. The integrand of J_I does not contain derivatives of $\boldsymbol{P}$, therefore we can carry out the minimization for each grid point $x \in \mathcal{E}$ individually. For each grid point $x \in \mathcal{E}$ we want to minimize $\|\nabla\hat{\boldsymbol{m}}(x) - \boldsymbol{P}(x)\|^2/2$. Let us denote by $\delta_{\boldsymbol{e}_i} m_j$ the central difference approximation of $\nabla_{\boldsymbol{e}_i} m_j$. The tensor $\nabla\hat{\boldsymbol{m}}$ will then be approximated by $d_{ij}\hat{\boldsymbol{e}}^i \otimes \hat{\boldsymbol{e}}^j$, where $d_{ij} := (\delta_{\boldsymbol{e}_j} m_i - \Gamma^k_{ij} m_k)$. Assuming this approximation of $\nabla\hat{\boldsymbol{m}}$, we will minimize

$$\begin{aligned}\frac{1}{2}\|(d_{ij}) - (P_{ij})\|^2 &= \frac{1}{2}\big(d_{ij} - P_{ij}\big)\big(d_{kl} - P_{kl}\big)e^{ik}e^{jl} \\ &= \frac{1}{2e}\left[e^{11}e_{22}(d_{11} - P_{11})^2 + (d_{12} - P_{12})^2 + (d_{21} - P_{21})^2 + e^{22}e_{11}(d_{22} - P_{22})^2\right],\end{aligned}$$

where we used the fact that the basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$ is orthogonal and hence (e_{ij}) is diagonal. The tensor $\boldsymbol{P}(x)$ is positive semi-definite if and only if the matrix $(e^{ij}P_{jk})$ is positive semi-definite. Recall that symmetric 2×2 matrices are positive semi-definite if and only if their trace and determinant are both positive. However, the matrix is not symmetric, because

$$(e^{ij}P_{jk}) = \begin{pmatrix} e^{11}P_{11} & e^{11}P_{12} \\ e^{22}P_{12} & e^{22}P_{22} \end{pmatrix}, \tag{4.10}$$

where we used that $P_{21} = P_{12}$. It is a familiar result that a matrix is positive semi-definite if and only if its eigenvalues are nonnegative. A quick calculation shows that the eigenvalues of the matrix $(e^{ij}P_{jk})$ are given by

$$\lambda_\pm = \frac{1}{2}\left(-(e^{11}P_{11} + e^{22}P_{22}) \pm \sqrt{(e^{11}P_{11} - e^{22}P_{22})^2 + 4e^{11}e^{22}P_{12}^2}\right). \tag{4.11}$$

A similar calculation shows that the eigenvalues of the matrix

$$\begin{pmatrix} e^{11}P_{11} & P_{12}/\sqrt{e} \\ P_{12}/\sqrt{e} & e^{22}P_{22} \end{pmatrix} \tag{4.12}$$

are also given by (4.11). This implies that $(e^{ij}P_{jk})$ is positive semi-definite if and only if the matrix in (4.12) is positive semi-definite. The matrix in (4.12) is symmetric, hence we can conclude that $\boldsymbol{P}(x)$ is positive semi-definite if and only if the trace and determinant of the matrix in (4.12) are nonnegative, i.e. if and only if

$$e^{11}P_{11} + e^{22}P_{22} \ge 0 \text{ and } (P_{11}P_{22} - P_{12}^2)/e \ge 0.$$

The metric e_{ij} is derived from an ordinary Pythagorean inner product hence we have $e > 0$ and we can simplify the last requirement to $\det(P_{ij}) \ge 0$.

The determinant of (P_{ij}) needs to equal eE/F. This quotient is positive by definition and hence $\det(P_{ij}) > 0$ is always satisfied. We can formulate the minimization problem as follows.

Problem 4.3.1. Given (d_{ij}), find $P_{11}, P_{12}, P_{22} \in \mathbb{R}$ that minimize the function

$$H(P_{11}, P_{12}, P_{22}) := \frac{1}{2e}\left[e^{11}e_{22}(d_{11} - P_{11})^2 + (d_{12} - P_{12})^2 + (d_{21} - P_{12})^2 + e^{22}e_{11}(d_{22} - P_{22})^2\right]. \tag{4.13}$$

under the constraints $P_{11}P_{22} - P_{12}^2 = eE/F > 0$ and $e^{11}P_{11} + e^{22}P_{22} \geq 0$.

We can rewrite Problem 4.3.1 in a more convenient form by introducing the new variables

$$\begin{aligned} \overline{P}_{11} &= e^{11}P_{11}, \\ \overline{P}_{12} &= P_{12}/\sqrt{e}, \\ \overline{P}_{22} &= e^{22}P_{22}. \end{aligned} \tag{4.14}$$

With these new variables we can rewrite the minimization function of Problem 4.3.1 as

$$\begin{aligned} H(P_{11}, P_{12}, P_{22}) &= \frac{1}{2e}\left[e^{11}e_{22}(d_{11} - e_{11}\overline{P}_{11})^2 + (d_{12} - \sqrt{e}\,\overline{P}_{12})^2\right. \\ &\qquad \left. +(d_{21} - \sqrt{e}\,\overline{P}_{12})^2 + e^{22}e_{11}(d_{22} - e_{22}\overline{P}_{22})^2\right] \\ &= \frac{1}{2}\left[(e^{11}d_{11} - \overline{P}_{11})^2 + (d_{12}/\sqrt{e} - \overline{P}_{12})^2 + (d_{21}/\sqrt{e} - \overline{P}_{12})^2 + (e^{22}d_{22} - \overline{P}_{22})^2\right]. \end{aligned}$$

This implies that we can equally well solve the following problem.

Problem 4.3.2. Given (d_{ij}), find $\overline{P}_{11}, \overline{P}_{12}, \overline{P}_{22} \in \mathbb{R}$ that minimize the function H under the constraints $\overline{P}_{11}\overline{P}_{22} - \overline{P}_{12}^2 = E/F$ and $\overline{P}_{11} + \overline{P}_{22} \geq 0$, where H is given by

$$H(\overline{P}_{11}, \overline{P}_{12}, \overline{P}_{22}) = \frac{1}{2}\left[(\overline{d}_{11} - \overline{P}_{11})^2 + (\overline{d}_{12} - \overline{P}_{12})^2 + (\overline{d}_{21} - \overline{P}_{12})^2 + (\overline{d}_{22} - \overline{P}_{22})^2\right],$$

where $\overline{d}_{11} = e^{11}d_{11}$, $\overline{d}_{12} = d_{12}/\sqrt{e}$, $\overline{d}_{21} = d_{21}/\sqrt{e}$ and $\overline{d}_{22} = e^{22}d_{22}$.

The minimizers (P_{11}, P_{12}, P_{22}) of Problem 4.3.1 are related to the minimizers $(\overline{P}_{11}, \overline{P}_{12}, \overline{P}_{22})$ of Problem 4.3.2 by equations (4.14).

We will algebraically solve Problem 4.3.2 by using the method of Lagrange multipliers. Besides this we will give a pictorial geometric representation of this problem. This serves to get more intuition for the problem and also provides a convenient way to verify the algebraically found solutions.

4.3.1 Lagrange minimizers and their geometric representation

We will find the minimizers of Problem 4.3.2 with the help of the *Lagrange function*

$$\Lambda(\overline{P}_{11}, \overline{P}_{12}, \overline{P}_{22}, \lambda) = H(\overline{P}_{11}, \overline{P}_{12}, \overline{P}_{22}) - \lambda\left(\overline{P}_{11}\overline{P}_{22} - \overline{P}_{12}^2 - \frac{E}{F}\right). \tag{4.15}$$

In a local minimum of this function all the partial derivatives have to equal zero, hence we find the following set of equations,

$$\overline{P}_{11} + \lambda\overline{P}_{22} = \overline{d}_{11}, \tag{4.16a}$$

$$(1 - \lambda)\overline{P}_{12} = \widetilde{d}_{12} := \frac{1}{2}(\overline{d}_{12} + \overline{d}_{21}), \tag{4.16b}$$

$$\lambda\overline{P}_{11} + \overline{P}_{22} = \overline{d}_{22}, \tag{4.16c}$$

$$\overline{P}_{11}\overline{P}_{22} - \overline{P}_{12}^2 = \frac{E}{F}. \tag{4.16d}$$

In the Lagrange function (4.15) the condition $\overline{P}_{11} + \overline{P}_{22} \geq 0$ has not been taken into account, hence a solution of (4.16a)-(4.16d) might have $\overline{P}_{11} + \overline{P}_{22} < 0$. In what follows, we will show that there always exists a solution to (4.16a)-(4.16d) such that $\overline{P}_{11} + \overline{P}_{22} \geq 0$.

We will now give a geometric interpretation to the Lagrange minimizers. We will show that they correspond to a joint tangent plane of a hyperboloid and an ellipsoid. Note that we can rewrite the function H as

$$H(\overline{P}_{11}, \overline{P}_{12}, \overline{P}_{22}) = \frac{1}{2}\big[(\overline{d}_{11} - \overline{P}_{11})^2 + 2(\widetilde{d}_{12} - \overline{P}_{12})^2 + (\overline{d}_{22} - \overline{P}_{22})^2\big] + \frac{1}{4}(\overline{d}_{12} - \overline{d}_{21})^2, \quad (4.17)$$

where $\widetilde{d}_{12}$ is as defined in equation (4.16b). Let us introduce $C := H(\overline{P}_{11}, \overline{P}_{12}, \overline{P}_{22}) - \frac{1}{4}(\overline{d}_{12} - \overline{d}_{21})^2$. From (4.17) we see that $H(\overline{P}_{11}, \overline{P}_{12}, \overline{P}_{22}) \geq \frac{1}{4}(\overline{d}_{12} - \overline{d}_{21})^2$ and hence is $C \geq 0$ for every value of H. Every value of C corresponds to an iso-surface of the function H. Using C we can rewrite (4.17) as

$$\left(\frac{\overline{P}_{11} - \overline{d}_{11}}{\sqrt{2C}}\right)^2 + \left(\frac{\overline{P}_{12} - \widetilde{d}_{12}}{\sqrt{C}}\right)^2 + \left(\frac{\overline{P}_{22} - \overline{d}_{22}}{\sqrt{2C}}\right)^2 = 1. \quad (4.18)$$

Equation (4.18) describes an ellipsoid in $\mathbb{R}^3$ with center $(\overline{d}_{11}, \widetilde{d}_{12}, \overline{d}_{22})$ and semi-axes $\sqrt{2C}$, $\sqrt{C}$ and $\sqrt{2C}$. Thus the iso-surfaces of H can be interpreted as ellipsoids in $\mathbb{R}^3$.

Let us now focus on the constraint

$$\overline{P}_{11}\overline{P}_{22} - \overline{P}_{12}^2 = E/F. \quad (4.19)$$

This constraint describes an hyperboloid in $\mathbb{R}^3$ with symmetry axis given by $\overline{P}_{11} = \overline{P}_{22}$ and $\overline{P}_{12} = 0$. To see this we will in a slightly different coordinate system. This coordinate transformation is given by

$$\begin{aligned} x &= \overline{P}_{11} - \overline{P}_{22}, \\ y &= \overline{P}_{12}, \\ z &= \overline{P}_{11} + \overline{P}_{22}. \end{aligned} \quad (4.20)$$

Using these coordinates, equation (4.19) turns into

$$\left(\frac{x}{2\sqrt{E/F}}\right)^2 + \left(\frac{y}{\sqrt{E/F}}\right)^2 - \left(\frac{z}{2\sqrt{E/F}}\right)^2 = -1. \quad (4.21)$$

This equation describes a hyperboloid of two separate sheets. One sheet is located in the half-space $z > 0$ and the other one is located in the half-space $z < 0$. The distance from the origin to the extremum of the upper sheet and the extremum of the lower sheet is $2\sqrt{E/F}$, respectively, $-2\sqrt{E/F}$.

If we substitute the new coordinates (4.20) into equation (4.18) we obtain

$$\left(\frac{x - (\overline{d}_{11} - \overline{d}_{22})}{2\sqrt{C}}\right)^2 + \left(\frac{y - \widetilde{d}_{12}}{\sqrt{C}}\right)^2 + \left(\frac{z - (\overline{d}_{11} + \overline{d}_{22})}{2\sqrt{C}}\right)^2 = 1.$$

We see (Figure 4.1) that the principal axes of both the ellipsoids and the hyperboloids are such that the x- and z-principal axis are equally long and twice the length of the y-principal axis. This fact will play a role in the minimization problem.

The local minimizers of the Lagrange function (4.15) are exactly the points where an iso-surface of H is tangent to the hyperboloid. This can be seen from the equations (4.16) in the following way. Equation (4.16d) implies that a local minimizer of the Lagrange function is a point of the hyperboloid. Furthermore, a minimizer of the Lagrange function Λ is a local minimum of H when

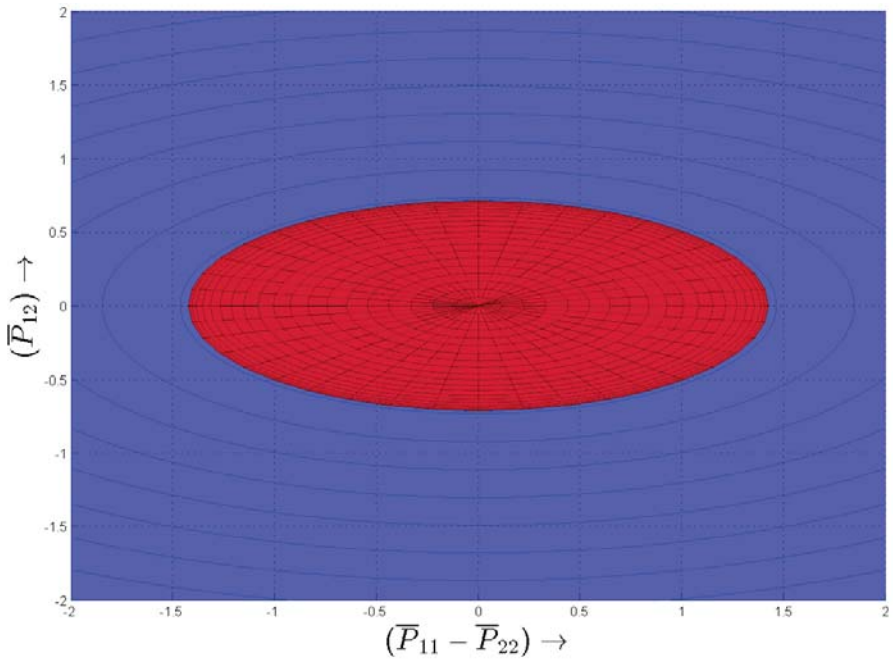


Figure 4.1: In this figure an example of an ellipsoidal iso-surface of H and a hyperboloid are shown from above the plane $z = \overline{P}_{11} + \overline{P}_{22} = 0$. The viewing direction is in the negative z-direction. The blue lines are the isolines of the hyperboloid. We see that the principal x- and y-axis have the same proportion for the hyperboloid and the ellipsoid.

confined to the hyperboloid. To see this suppose that a point p on the hyperboloid is a minimizer to λ. Now if p is not a local minimum of H restricted to the hyperboloid, then there would be a direction to go, while staying on the hyperboloid, in which the function H increases while the second term in (4.15) stays constant. This would then contradict the fact that the point p is a minimizer to Λ. Thus we can conclude that the minimizers of Λ are the points local minimima of H restricted to the hyperboloid. Now, a local minimum of H restricted to the surface of the hyperboloid is exactly a point where an iso-surface of H is tangent to the hyperboloid. The plane $z = \overline{P}_{11} + \overline{P}_{22} = 0$ lies precisely between the two sheets of the hyperboloid. Thus, only the points where an iso-surface of H is tangent to the upper sheet of the hyperboloid are actual minimizers of Problem 4.3.2. In Figure 4.2 an example of a hyperboloid with ellipsoid is shown. The global minimizer corresponds to the smallest ellipsoid that is tangent to the upper sheet of the hyperboloid. An example of this is shown in Figure 4.3.

In the remaining part of this section we will algebraically solve the system of equations (4.16). We will verify the algebraic solutions that we find by these geometric pictures. This allows us to get more intuition for the problem and visualizes symmetries that are not directly apparent from the equations (4.16a) - (4.16d).

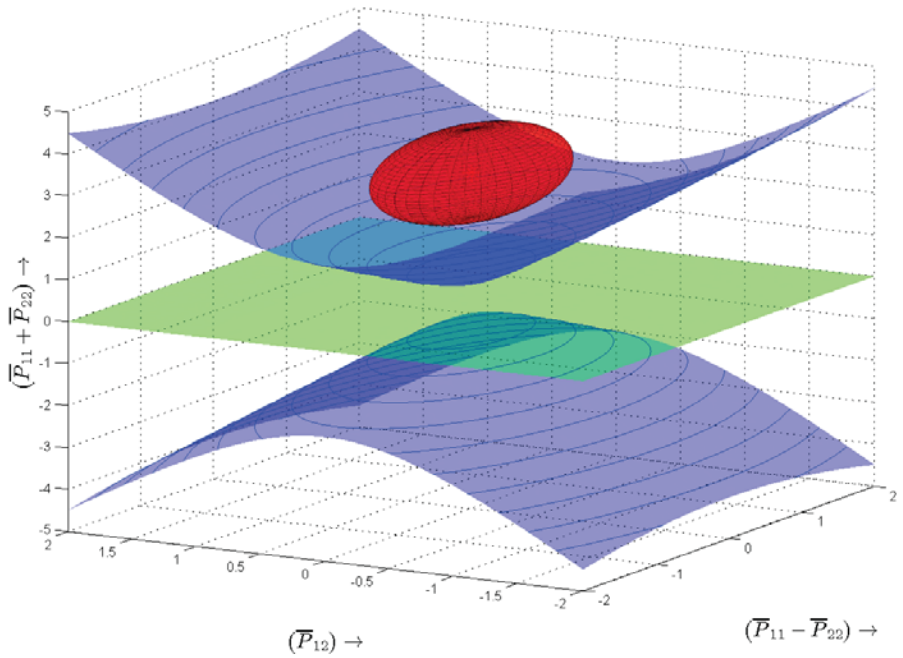


Figure 4.2: In this figure an example of an ellipsoidal iso-surface of H and a hyperboloid are shown together with the plane $z = \overline{P}_{11} + \overline{P}_{22} = 0$ in green. It can be seen that one of the two sheets of the hyperboloid lies above the green plane and one lies below the green plane. Only minimizers above the green plane satisfy $\overline{P}_{11} + \overline{P}_{22} > 0$.

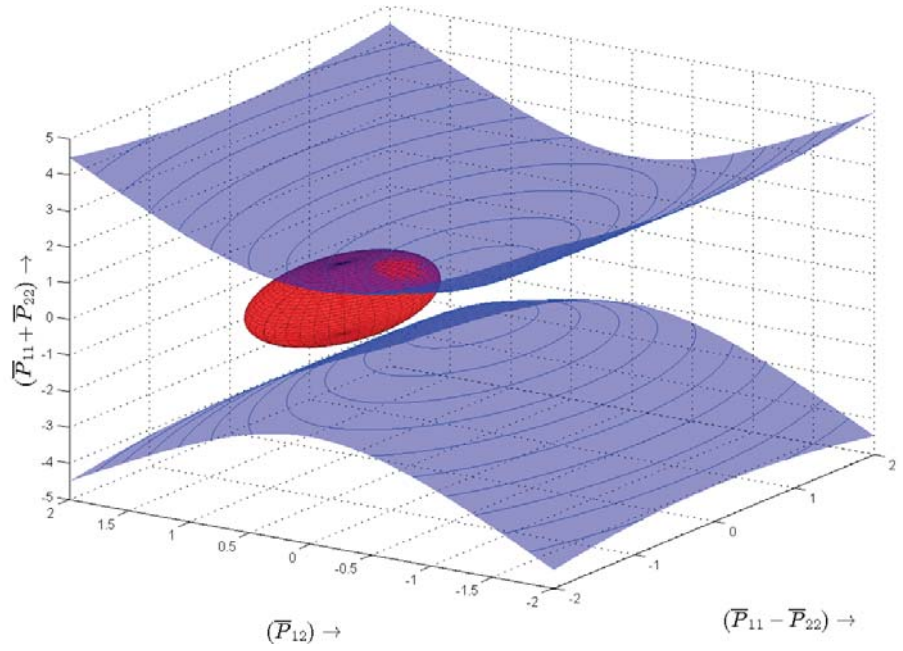


Figure 4.3: A little bit of red of the ellipsoid can be seen appearing through the upper sheet of the hyperboloid. This is the smallest ellipsoid around this point that is tangent to the upper sheet of the hyperboloid and therefore corresponds to a global minimizer.

4.3.2 Determining the minimizers

We will show that for each given $d_{ij}\hat{e}^i \otimes \hat{e}^j$ we can find $\overline{P}_{11}, \overline{P}_{12}, \overline{P}_{22} \in \mathbb{R}$ that are the solution to Problem 4.3.2. If $\lambda \neq \pm 1$, we can invert equations (4.16a) - (4.16c). Doing this we obtain

$$\overline{P}_{11} = \frac{\lambda \overline{d}_{22} - \overline{d}_{11}}{\lambda^2 - 1}, \tag{4.22a}$$

$$\overline{P}_{12} = \frac{\widetilde{d}_{12}}{1 - \lambda}, \tag{4.22b}$$

$$\overline{P}_{22} = \frac{\lambda \overline{d}_{11} - \overline{d}_{22}}{\lambda^2 - 1}. \tag{4.22c}$$

However, these equations only hold if $\lambda \neq \pm 1$. From equations (4.16a) - (4.16c) we have the following immediate logical implications:

$$\lambda = 1 \implies [\overline{d}_{11} = \overline{d}_{22}] \wedge [\widetilde{d}_{12} = 0], \tag{4.23a}$$

$$\lambda = -1 \implies [\overline{d}_{11} = -\overline{d}_{22}]. \tag{4.23b}$$

From these implications we see there are only two situations that have to be dealt with separately, namely the cases $[\overline{d}_{11} = \overline{d}_{22}] \wedge [\widetilde{d}_{12} = 0]$ and $[\overline{d}_{11} = -\overline{d}_{22}]$. When we are not in one of these two cases we can faultlessly write down equations (4.22a) - (4.22c). We will now treat the three different cases in turn, starting out with the general case.

Lemma 4.3.3. *When $\neg[[\overline{d}_{11} = \overline{d}_{22}] \wedge [\widetilde{d}_{12} = 0]] \wedge \neg[\overline{d}_{11} = -\overline{d}_{22}]$, the global minimizer to Problem 4.3.2 is given by equations (4.22a) - (4.22c). In these expressions λ is given by one of the following four expressions:*

$$\begin{aligned} \lambda_1 &= -\sqrt{\frac{y}{2}} + \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} + \frac{a_1}{2a_4\sqrt{2y}}}, \quad \lambda_2 = -\sqrt{\frac{y}{2}} - \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} + \frac{a_1}{2a_4\sqrt{2y}}}, \\ \lambda_3 &= \sqrt{\frac{y}{2}} + \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} + \frac{a_1}{2a_4\sqrt{2y}}}, \qquad \lambda_4 = \sqrt{\frac{y}{2}} - \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} + \frac{a_1}{2a_4\sqrt{2y}}}. \end{aligned} \tag{4.24}$$

In (4.24) y is given by the following two sets of equations:

$$\begin{aligned} y &= A + \frac{Q}{A} - \frac{b_2}{3}, & A &= -\operatorname{sgn}(R)\left(|A| + \sqrt{R^2 - Q^3}\right)^{1/3}, \\ R &= \frac{2b_2^3 - 9b_1b_2 + 27b_0}{54}, & Q &= \frac{b_2^3 - 3b_1}{9}, \end{aligned} \tag{4.25}$$

and

$$\begin{aligned} b_0 &= -\frac{1}{8}\left(\frac{a_1}{a_4}\right)^2, & b_1 &= \frac{1}{4}\left(\frac{a_2}{a_4}\right)^2 - \frac{a_0}{a_4}, \\ b_2 &= \frac{a_2}{a_4}, & & \\ a_0 &= \frac{E}{F} - \det(\widetilde{\boldsymbol{D}}), & a_1 &= \overline{d}_{11}^2 + \overline{d}_{22}^2 + 2\widetilde{d}_{12}^2, \\ a_2 &= -2E/F - \det(\widetilde{\boldsymbol{D}}), & a_4 &= E/F, \end{aligned} \tag{4.26}$$

where

$$\widetilde{\boldsymbol{D}} = \begin{pmatrix} \overline{d}_{11} & \widetilde{d}_{12} \\ \widetilde{d}_{12} & \overline{d}_{22} \end{pmatrix}.$$

At least one of the four choices for λ is such that the requirement $\overline{P}_{11} + \overline{P}_{22} > 0$ is satisfied by (4.22a) - (4.22c).

Proof. Substituting the expressions (4.22a) - (4.22c) in (4.16d) we obtain the following quartic polynomial for λ:

$$\Pi(\lambda) := a_4\lambda^4 + a_2\lambda^2 + a_1\lambda + a_0 = 0,$$

where the coefficients are as given (4.26). In [5, p.135] it is shown that this polynomial admits the four solutions (4.24). The leading term of Π is $a_4 = E/F$ which is by definition greater than zero, hence

$$\lim_{\lambda\to\pm\infty} \Pi(\lambda) = \infty.$$

Furthermore, we can rewrite $\Pi(\lambda)$ as

$$\Pi(\lambda) = a_4(\lambda^2-1)^2 - \det(\widetilde{\boldsymbol{D}})(\lambda^2+1) + (\overline{d}_{11}^2 + \overline{d}_{22}^2 + 2\widetilde{d}_{12})\lambda.$$

From this we see that

$$\begin{aligned}\Pi(\lambda=-1) &= -2\det(\widetilde{\boldsymbol{D}}) - (\overline{d}_{11}^2 + \overline{d}_{22}^2 + 2\widetilde{d}_{12}) \\ &= -2\overline{d}_{11}\overline{d}_{22} + 2\widetilde{d}_{12}^2 - \overline{d}_{11}^2 - \overline{d}_{22}^2 - 2\widetilde{d}_{12} \\ &= -(\overline{d}_{11} + \overline{d}_{22})^2.\end{aligned}$$

By assumption $\overline{d}_{11} \neq \overline{d}_{22}$, hence $\Pi(\lambda=-1) < 0$. This combined with the fact that $\Pi(\lambda) \to +\infty$ for $\lambda \to \pm\infty$ implies that Π must have at least two real roots.

From (4.22a) and (4.22c) it follows that

$$\overline{P}_{11} + \overline{P}_{22} = \frac{\overline{d}_{11} + \overline{d}_{22}}{1+\lambda}.$$

This shows that for one of the two real roots it holds that $\overline{P}_{11} + \overline{P}_{22} > 0$, while for the other real root it holds that $\overline{P}_{11} + \overline{P}_{22} < 0$.

We now have established the fact that one of the four λ in (4.24) is such that (4.22a) - (4.22c) is a minimum of the Lagrange function such that it adheres to $\overline{P}_{11} + \overline{P}_{22} > 0$, thereby it follows that a global minimizer exists. Moreover, the minimizer is given by (4.22a) - (4.22c), with λ given by one of the real roots of (4.24). □

Now that we have dealt with the general case we will turn our attention to the cases $[\overline{d}_{11} = \overline{d}_{22}] \wedge [\widetilde{d}_{12} = 0]$ and $[\overline{d}_{11} = -\overline{d}_{22}]$. We first handle $[\overline{d}_{11} = -\overline{d}_{22}]$.

Lemma 4.3.4. *When $\overline{d}_{11} = -\overline{d}_{22}$, the global minimizer to Problem 4.3.2 is given by*

$$\overline{P}_{11} = \frac{\overline{d}_{11} + \sqrt{\overline{d}_{11}^2 + 4E/F + \widetilde{d}_{12}^2}}{2}, \tag{4.27a}$$

$$\overline{P}_{12} = \frac{\widetilde{d}_{12}}{2}, \tag{4.27b}$$

$$\overline{P}_{22} = \frac{-\overline{d}_{11} + \sqrt{\overline{d}_{11}^2 + 4E/F + \widetilde{d}_{12}^2}}{2}. \tag{4.27c}$$

Proof. When $\overline{d}_{11} = -\overline{d}_{22}$, the Lagrange conditions (4.16a) and (4.16c) imply that $(\lambda+1)(\overline{P}_{11} + \overline{P}_{22}) = 0$. From this it follows that we have either $\lambda = -1$ or $\overline{P}_{11} = -\overline{P}_{22}$. When $\overline{P}_{11} = -\overline{P}_{22}$, it holds by (4.16d) that

$$-\overline{P}_{11}^2 - \overline{P}_{12}^2 = \frac{E}{F}.$$

However, this situation cannot occur because $E/F > 0$. We find that $\lambda = -1$ must hold.

The Lagrange conditions (4.16a) - (4.16d) now simplify to

$$\begin{aligned}\overline{P}_{11} - \overline{P}_{22} &= \overline{d}_{11},\\ 2\overline{P}_{12} &= \widetilde{d}_{12},\\ \overline{P}_{22} - \overline{P}_{11} &= \overline{d}_{22},\\ \overline{P}_{11}\overline{P}_{22} &= \frac{E}{F} + \frac{\widetilde{d}_{12}^2}{4}.\end{aligned}$$

Combining equations first and fourth of these equations gives us

$$\overline{P}_{11}^2 - \overline{P}_{11}\overline{d}_{11} = \frac{E}{F} + \frac{\widetilde{d}_{12}^2}{4},$$

which we can rewrite to

$$\overline{P}_{11}^2 - \overline{P}_{11}\overline{d}_{11} - \frac{E}{F} - \frac{\widetilde{d}_{12}^2}{4} = 0.$$

This polynomial has for any combination of $\overline{d}_{11}$, $\widetilde{d}_{12}$ and $\overline{d}_{22}$ always two real solutions, which are given by

$$\overline{P}_{11} = \frac{\overline{d}_{11} \pm \sqrt{\overline{d}_{11}^2 + 4E/F + \widetilde{d}_{12}^2}}{2}.$$

However, if the minus sign holds we see that

$$\overline{P}_{11} + \overline{P}_{22} = -\sqrt{\overline{d}_{11}^2 + 4E/F + \widetilde{d}_{12}^2} \leq 0.$$

Thus, when $\overline{d}_{11} = -\overline{d}_{22}$, the global minimizer to Problem 4.3.2 is given by (4.27a) - (4.27c). In Figure 4.4 these findings are illustrated. □

Now we only have to deal yet with the case $[\overline{d}_{11} = \overline{d}_{22}] \wedge [\widetilde{d}_{12} = 0]$.

Lemma 4.3.5. *Suppose $\overline{d}_{11} = \overline{d}_{22}$ and $\widetilde{d}_{12} = 0$. The solution to Problem 4.3.2 is the global minimum given by*

$$\overline{P}_{11} = \sqrt{\frac{E}{F}}, \tag{4.28a}$$

$$\overline{P}_{12} = 0, \tag{4.28b}$$

$$\overline{P}_{22} = \sqrt{\frac{E}{F}}, \tag{4.28c}$$

if $\overline{d}_{11} < 2\sqrt{E/F}$, and a continuum of global minimizers given by

$$\overline{P}_{11} \in \left[\frac{\overline{d}_{11} - \sqrt{\overline{d}_{11}^2 - 4E/F}}{2}, \frac{\overline{d}_{11} + \sqrt{\overline{d}_{11}^2 - 4E/F}}{2}\right], \tag{4.29a}$$

$$\overline{P}_{12} = \pm\sqrt{\overline{d}_{11}\overline{P}_{11} - \overline{P}_{11}^2 - E/F}, \tag{4.29b}$$

$$\overline{P}_{22} = \overline{d}_{11} - \overline{P}_{11}, \tag{4.29c}$$

if $\overline{d}_{11} \geq 2\sqrt{E/F}$.

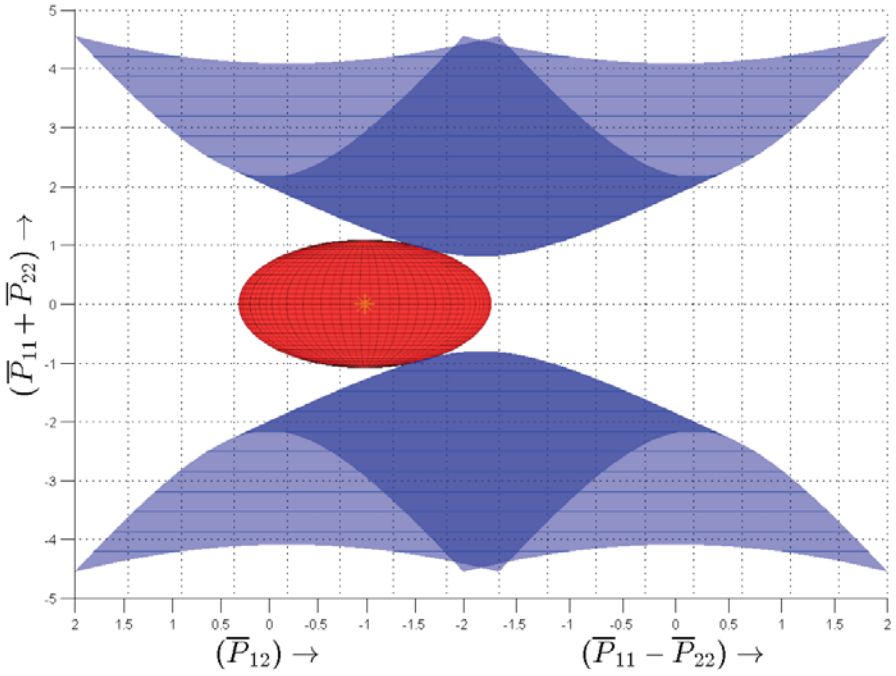


Figure 4.4: This plot corresponds to Lemma 4.3.4. The ellipsoid is centered around $(x = \overline{d}_{11} - \overline{d}_{22} = 2\overline{d}_{11}, y = \overline{d}_{12}, z = \overline{d}_{11} + \overline{d}_{22} = 0)$. This results in two local minima with the same function value for H. One of the minima is on the upper sheet and the other is on the lower sheet of the hyperboloid. These are the two minima that have been found in the proof of Lemma 4.3.4, the minimum on the lower sheet was discarded as it did not satisfy $\overline{P}_{11} + \overline{P}_{22} > 0$.

Proof. In the case that $\overline{d}_{11} = \overline{d}_{22}$ and $\widetilde{d}_{12} = 0$, Lagrange conditions (4.16a) and (4.16c) imply that

$$(1-\lambda)(\overline{P}_{11} - \overline{P}_{22}) = 0.$$

From this it follows that we must either have $\lambda = 1$ or $\lambda \neq 1$ and then $\overline{P}_{11} = \overline{P}_{22}$. Let us first deal with the case $\lambda \neq 1$.

When $\lambda \neq 1$, the Lagrange conditions (4.16b) and (4.16d) read

$$\begin{aligned}(1-\lambda)\overline{P}_{12} &= \widetilde{d}_{12} = 0,\\ \overline{P}_{11}^2 - \overline{P}_{12}^2 &= \frac{E}{F}.\end{aligned}$$

As $\lambda \neq 1$, the first of these equations implies that $\overline{P}_{12} = 0$. This fact combined with the second equation implies that

$$\overline{P}_{11} = \overline{P}_{22} = \pm\sqrt{\frac{E}{F}}.$$

The condition $\overline{P}_{11}+\overline{P}_{22} > 0$ is only satisfied when the plus sign holds, hence we find one minimizer. This is the minimizer given by equations (4.28a) - (4.28c).

Now suppose that $\lambda = 1$. From Lagrange condition (4.16b) we find that

$$\overline{P}_{22} = \overline{d}_{11} - \overline{P}_{11} \tag{$\star$}$$

and from Lagrange condition (4.16d) we obtain

$$\overline{P}_{12} = \pm\sqrt{\overline{P}_{11}\overline{P}_{22} - E/F}. \tag{$\dagger$}$$

Substituting ($\star$) in ($\dagger$) gives us

$$\overline{P}_{12} = \pm\sqrt{\overline{d}_{11}\overline{P}_{11} - \overline{P}_{11}^2 - E/F},$$

which is only real if

$$\overline{P}_{11}^2 - \overline{d}_{11}\overline{P}_{11} + \frac{E}{F} \leq 0,$$

that is, when

$$\overline{P}_{11} \in \left[\frac{\overline{d}_{11} - \sqrt{\overline{d}_{11}^2 - 4E/F}}{2}, \frac{\overline{d}_{11} + \sqrt{\overline{d}_{11}^2 - 4E/F}}{2}\right]. \tag{$\ddagger$}$$

This gives us the continuum of minimizers (4.29a) - (4.29c). However, the interval in ($\ddagger$) only contains real values when $\overline{d}_{11} \notin (-2\sqrt{E/F}, 2\sqrt{E/F})$. Moreover, because $\overline{P}_{11} + \overline{P}_{22} = \overline{d}_{11}$, we see that $\overline{P}_{11} + \overline{P}_{22} > 0$ is only satisfied when $\overline{d}_{11} > 0$. From this we see that the continuum of minimizers can only be a solution to Problem 4.3.2 when $\overline{d}_{11} \geq 2\sqrt{E/F}$. Thus, when $\overline{d}_{11} < 2\sqrt{E/F}$, the global minimizer is given by (4.28a) - (4.28c). To decide for $\overline{d}_{11} \geq 2\sqrt{E/(eF)}$ if the global minimizer is given by (4.28a) - (4.28c) or by an element of the continuum (4.29a) - (4.29c), we must compare the value of the function being minimized, i.e. F, for the local minimizers.

First, we remark that $H(\overline{P}_{11}, \overline{P}_{12}, \overline{P}_{22})$ has the same value for every element of the continuum of minimizers, because otherwise not all the elements of continuum would have been local minima.

We denote the value of $H(\overline{P}_{11}, \overline{P}_{12}, \overline{P}_{22})$ in the continuum by H_{cont}. For H_{cont} we obtain

$$\begin{aligned} H_{\text{cont}} &= H\left(\frac{\overline{d}_{11}}{2}, \sqrt{\frac{\overline{d}_{11}^2}{4} - \frac{E}{F}}, \frac{\overline{d}_{11}}{2}\right) \\ &= \frac{1}{2}\left(\frac{\overline{d}_{11}^2}{4} + 2\left(\frac{\overline{d}_{11}^2}{4} - \frac{E}{F}\right) + \frac{\overline{d}_{11}^2}{4}\right) \\ &= \frac{\overline{d}_{11}^2}{2} - \frac{E}{F}. \end{aligned}$$

We denote the local minimizer given by (4.28a) - (4.28c) by H_{ext} and have

$$\begin{aligned} H_{\text{ext}} &= H\left(\sqrt{\frac{E}{F}}, 0, \sqrt{\frac{E}{F}}\right) \\ &= \frac{E}{F} + \overline{d}_{11}^2 - 2\overline{d}_{11}\sqrt{\frac{E}{F}}. \end{aligned}$$

In these calculations we used the fact that $\widetilde{d}_{12} = 0$ implies that $\overline{d}_{12} = -\overline{d}_{21}$. Subtracting H_{cont} from H_{ext} gives us

$$H_{\text{ext}} - H_{\text{cont}} = \frac{\overline{d}_{11}^2}{2} - 2\sqrt{\frac{E}{F}}\overline{d}_{11} + \frac{2E}{F}. \tag{4.30}$$

The polynomial (4.30) considered as a function of $\overline{d}_{11}$ attains a minimum for $\overline{d}_{11} = \sqrt{E/F}$ in which it equals 0 and in all other points it is positive. Thus, we find that $H_{\text{cont}} \leq H_{\text{ext}}$. This implies that if $\overline{d}_{11} \geq 2\sqrt{E/F}$, the solution to Problem 4.3.2 is given by the continuum of minimizers (4.29a) - (4.29c). □

In Figure 4.5 and Figure 4.6 examples of the results from Lemma 4.3.5 are geometrically shown. Recall that the extrema of the two sheets of the hyperboloid are located at

$$(\overline{P}_{11} - \overline{P}_{22}, \overline{P}_{12}, \overline{P}_{11} + \overline{P}_{22}) = \pm(0, 0, 2\sqrt{E/F}).$$

Thus Lemma 4.3.5 implies that the global minimizer is $(0, 0, 2\sqrt{E/F})$ if $\overline{d}_{11} < 2\sqrt{E/F}$, $\overline{d}_{11} = \overline{d}_{22}$ and $\tilde{d}_{12} = 0$, i.e. when the center of the ellipsoid is located in $(0, 0, \overline{P}_{11} + \overline{P}_{22})$, where $\overline{P}_{11} + \overline{P}_{22} = 2\overline{d}_{11} < 4\sqrt{E/F}$. Or to put it in words, in the case that $\overline{d}_{11} = \overline{d}_{22}$ and $\tilde{d}_{12} = 0$, if the distance from the center of the ellipsoid to the origin is less that two times the distance to the minimum of the upper sheet of the hyperboloid, or if the center of the ellipsoid is situated beneath the plane $\overline{P}_{11} + \overline{P}_{22} = 0$, then the global minimizer is given by the extremum of the upper sheet of the hyperboloid. If $\overline{d}_{11} = \overline{d}_{22}$, $\tilde{d}_{12} = 0$, the center of the ellipsoid is located above the plane $\overline{P}_{11} + \overline{P}_{22} = 0$ and its distance to the origin is more than twice the distance from the extremum to the origin, then we have the continuum of global minimizers. This case is depicted in Figure 4.5. In Figure 4.6, the center of the ellipsoid is farther away from the origin than the extremum of the upper sheet of the hyperboloid, but it is less far away than two times the distance between this extremum and the origin. This results in the extremum as single global minimizer, as can be seen in this figure.

Let us summarize this section by the following theorem.

Theorem 4.3.6. *The minimization problem, Problem 4.3.2, can be solved algebraically. In the general case, when* $\neg[[\overline{d}_{11} = \overline{d}_{22}] \wedge [\widetilde{d}_{12} = 0]] \wedge \neg[\overline{d}_{11} = -\overline{d}_{22}]$*, the solution to Problem 4.3.2 is given by (4.16a) - (4.16c), with* λ *given by one of the four possibilities in (4.24). At least two of the* λ*'s in (4.24) are real and one of these corresponds to the global minimizer. Explicit calculation of the function value* $H(\overline{P}_{11}, \overline{P}_{12}, \overline{P}_{22})$ *shows which of the two real* λ*'s gives the global minimizer.*

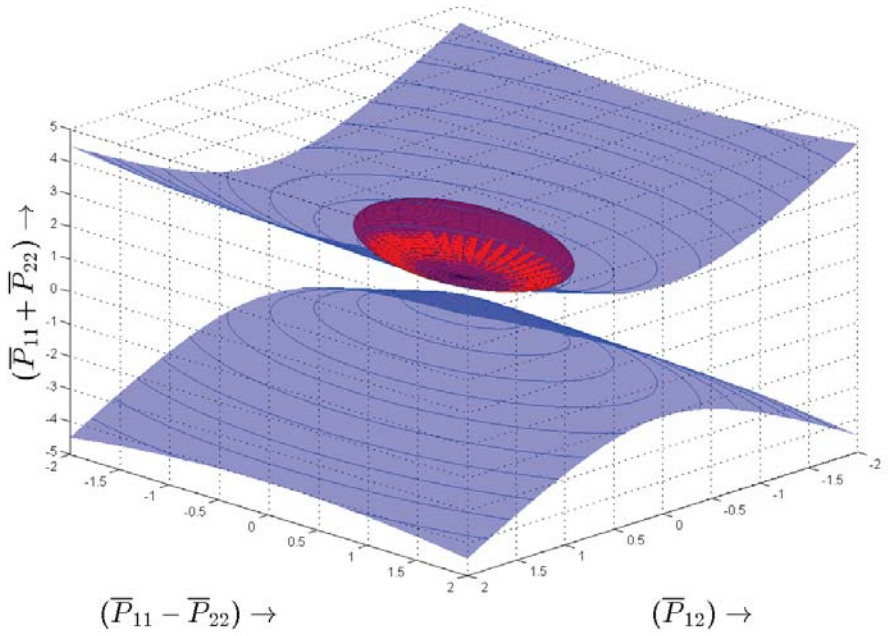


Figure 4.5: This plot corresponds to the continuum of minimizers found in Lemma 4.3.5. We see the upper sheet of the hyperboloid from below. The ellipsoid is centered around a point $(x = \overline{d}_{11} - \overline{d}_{22} = 0, y = \widetilde{d}_{12} = 0, z = \overline{d}_{11} + \overline{d}_{22} = 2\overline{d}_{11})$, where $\overline{d}_{11} \geq 2\sqrt{E/F}$. The continuum found referred to in the lemma turns out to be an ellipse in which the ellipsoidal iso-surface of H is tangent to the upper sheet of the hyperboloid.

In the case that $[\overline{d}_{11} = -\overline{d}_{22}]$*, there is a unique solution to Problem 4.3.2. This global minimizer is given by (4.27a) - (4.27c).*

Finally, in the case that $[[\overline{d}_{11} = \overline{d}_{22}] \wedge [\widetilde{d}_{12} = 0]]$*, there is unique solution to Problem 4.3.2 if* $\overline{d}_{11} < 2\sqrt{E/F}$ *and it is given by (4.28a) - (4.28c). If* $\overline{d}_{11} \geq 2\sqrt{E/F}$*, there is a whole continuum of solutions to Problem 4.3.2, which is given by (4.29a) - (4.29c).*

Proof. This theorem is just a summary of Lemma 4.3.3, Lemma 4.3.4 and Lemma 4.3.5 and hence directly follows from these. □

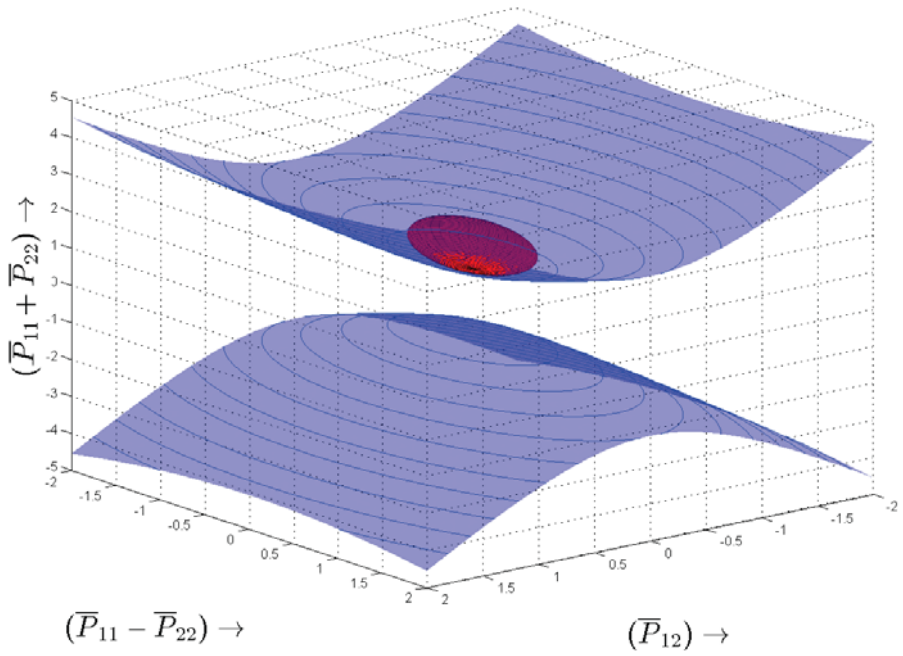


Figure 4.6: This plot corresponds to the isolated minimizer found in Lemma 4.3.5. We see the upper sheet of the hyperboloid from below. The ellipsoid is centered around a point $(x = \overline{d}_{11} - \overline{d}_{22} = 0, y = \widetilde{d}_{12} = 0, z = \overline{d}_{11} + \overline{d}_{22} = 2\overline{d}_{11})$, where $-2\sqrt{E/F} < \overline{d}_{11} < 2\sqrt{E/F}$. Thus the center of the ellipsoid is located between the two sheets of the hyperboloid. The ellipsoidal iso-surface of H is tangent to the upper sheet of the hyperboloid in the minimum of the upper sheet of the hyperboloid.

4.4 Minimization of J

In this section we will focus on the last step of the least-squares method, i.e. (4.8c). We will minimize the functional J, defined in equation (4.6). We will minimize J for $\boldsymbol{m} \in \mathcal{V}$, with $\mathcal{V}$ as defined in (4.7), while keeping $\boldsymbol{P}$ and $\boldsymbol{b}$ constant. Again we will do this for arbitrary coordinates x^1 and x^2 on $\mathcal{E}$, with metric e_{ij}. This will occupy the coming subsection and we will derive a boundary value problem for the mapping $\boldsymbol{m}$. In the subsequent subsection we will give coordinate dependent formulations of this boundary value problem for Cartesian and polar coordinate systems. We will see that in the Cartesian case we end up with the same boundary value problem for $\boldsymbol{m}$ as derived in [5, p.142-p.144].

4.4.1 Derivation of a boundary value problem for the mapping

We will use *Calculus of Variations* to determine the minimizer $\boldsymbol{m}$ for J. For a minimum to be attained the Fréchet derivative of the functional J must be zero, i.e.

$$\lim_{\varepsilon \to 0} \frac{J(\boldsymbol{m} + \varepsilon \boldsymbol{\eta}, \boldsymbol{P}, \boldsymbol{b}) - J(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b})}{\varepsilon} = 0. \tag{4.31}$$

This must hold for every $\boldsymbol{\eta}$ in $\mathcal{V}$, because if $\boldsymbol{\eta} \in \mathcal{V}$ then also $\boldsymbol{m} + \varepsilon\boldsymbol{\eta} \in \mathcal{V}$. The Fréchet derivative of J can be rewritten as a linear combination of the Fréchet derivatives of J_I and J_B:

$$\begin{aligned}\lim_{\varepsilon \to 0} \frac{J(\boldsymbol{m} + \varepsilon \boldsymbol{\eta}, \boldsymbol{P}, \boldsymbol{b}) - J(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b})}{\varepsilon} &= \alpha \lim_{\varepsilon \to 0} \frac{J_I(\boldsymbol{m} + \varepsilon \boldsymbol{\eta}, \boldsymbol{P}) - J_I(\boldsymbol{m}, \boldsymbol{P})}{\varepsilon} \\ &\quad + (1-\alpha) \lim_{\varepsilon \to 0} \frac{J_B(\boldsymbol{m} + \varepsilon \boldsymbol{\eta}, \boldsymbol{b}) - J_B(\boldsymbol{m}, \boldsymbol{b})}{\varepsilon}.\end{aligned}$$

Let us first determine the Fréchet derivative of J_I. By the linearity of the covariant derivative it follows that

$$\begin{aligned}\lim_{\varepsilon \to 0} \frac{J_I(\boldsymbol{m} + \varepsilon \boldsymbol{\eta}, \boldsymbol{P}) - J_I(\boldsymbol{m}, \boldsymbol{P})}{\varepsilon} &= \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \left[\iint_{\mathcal{E}} \|\nabla(\hat{\boldsymbol{m}} + \varepsilon \hat{\boldsymbol{\eta}}) - \boldsymbol{P}\|^2 - \|\nabla \hat{\boldsymbol{m}} - \boldsymbol{P}\|^2 \, \mathrm{d}A \right] \\ &= \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \left[\iint_{\mathcal{E}} \|\varepsilon \nabla \hat{\boldsymbol{\eta}} + \nabla \hat{\boldsymbol{m}} - \boldsymbol{P}\|^2 - \|\nabla \hat{\boldsymbol{m}} - \boldsymbol{P}\|^2 \, \mathrm{d}A \right].\end{aligned}$$

We will now need the following convenient property of inner product on $\boldsymbol{T}_0^2(T_x\mathcal{E})$ as defined on page 57. Let $\boldsymbol{A}, \boldsymbol{B} \in \boldsymbol{T}_0^2(T_x\mathcal{E})$, we have

$$\begin{aligned}\|\boldsymbol{A} + \boldsymbol{B}\|^2 &= (A_{ij} + B_{ij})(A^{ij} + B^{ij}) \\ &= A_{ij}A^{ij} + B_{ij}A^{ij} + A_{ij}B^{ij} + B_{ij}B^{ij} \\ &= \|\boldsymbol{A}\|^2 + 2\boldsymbol{A} : \boldsymbol{B} + \|\boldsymbol{B}\|^2.\end{aligned}$$

Using this property on $\|\varepsilon \nabla \hat{\boldsymbol{\eta}} + \nabla \hat{\boldsymbol{m}} - \boldsymbol{P}\|^2$, with $\boldsymbol{A} = \varepsilon \nabla \hat{\boldsymbol{\eta}}$ and $\boldsymbol{B} = \nabla \hat{\boldsymbol{m}} - \boldsymbol{P}$ gives us

$$\begin{aligned}\lim_{\varepsilon \to 0} \frac{J_I(\boldsymbol{m} + \varepsilon \boldsymbol{\eta}, \boldsymbol{P}) - J_I(\boldsymbol{m}, \boldsymbol{P})}{\varepsilon} &= \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \left[\iint_{\mathcal{E}} \varepsilon^2 \|\nabla \hat{\boldsymbol{\eta}}\|^2 + 2\varepsilon \nabla \hat{\boldsymbol{\eta}} : (\nabla \hat{\boldsymbol{m}} - \boldsymbol{P}) \, \mathrm{d}A \right] \\ &= \iint_{\mathcal{E}} \nabla \hat{\boldsymbol{\eta}} : (\nabla \hat{\boldsymbol{m}} - \boldsymbol{P}) \, \mathrm{d}A.\end{aligned}$$

Now we will determine the Fréchet derivative of J_B. Using the fact that

$$\|\boldsymbol{m} + \varepsilon \boldsymbol{\eta} - \boldsymbol{b}\|^2 = \varepsilon^2 \|\boldsymbol{\eta}\|^2 + 2\varepsilon(\boldsymbol{\eta} \mid \boldsymbol{m} - \boldsymbol{b}) + \|\boldsymbol{m} - \boldsymbol{b}\|^2, \tag{4.32}$$

we find

$$\begin{aligned}\frac{J_B(\boldsymbol{m}+\varepsilon\boldsymbol{\eta},\boldsymbol{b})-J_B(\boldsymbol{m},\boldsymbol{b})}{\varepsilon} &= \lim_{\varepsilon\to 0}\frac{1}{2\varepsilon}\left[\oint_{\partial\mathcal{E}}\|\boldsymbol{m}+\varepsilon\boldsymbol{\eta}-\boldsymbol{b}\|^2-\|\boldsymbol{m}-\boldsymbol{b}\|^2\,\mathrm{d}s\right]\\ &= \lim_{\varepsilon\to 0}\frac{1}{2\varepsilon}\left[\oint_{\partial\mathcal{E}}2\varepsilon(\boldsymbol{\eta}\mid(\boldsymbol{m}-\boldsymbol{b}))+\varepsilon^2\|\boldsymbol{\eta}\|^2\,\mathrm{d}s\right]\\ &= \oint_{\partial\mathcal{E}}(\boldsymbol{\eta}\mid(\boldsymbol{m}-\boldsymbol{b}))\,\mathrm{d}s.\end{aligned}$$

Combining the results for J_I and J_B we find that

$$\forall\boldsymbol{\eta}\in\mathcal{V}:\qquad \alpha\iint_{\mathcal{E}}\nabla\hat{\boldsymbol{\eta}}:(\nabla\hat{\boldsymbol{m}}-\boldsymbol{P})\,\mathrm{d}A+(1-\alpha)\oint_{\partial\mathcal{E}}(\boldsymbol{\eta}\mid(\boldsymbol{m}-\boldsymbol{b}))\,\mathrm{d}s=0. \tag{4.33}$$

In order to proceed we will rewrite the integrals in terms of the coordinate system on $\mathcal{E}$. For the first integral in (4.33) we have

$$\alpha\iint_{\mathcal{E}}\nabla\hat{\boldsymbol{\eta}}:(\nabla\hat{\boldsymbol{m}}-\boldsymbol{P})\,\mathrm{d}A=\alpha\iint_{\mathcal{E}}D_j\eta_i(D^jm^i-P^{ij})\sqrt{e}\,\mathrm{d}x^1\mathrm{d}x^2,$$

where $D_j\eta_i$ are the components of the covariant derivative of $\hat{\boldsymbol{\eta}}$. Note that $D^j=e^{ij}D_i$. By the product rule it follows that

$$D_j\eta_i(D^jm^i-P^{ij})=D_j(\eta_i(D^jm^i-P^{ij}))-\eta_iD_j(D^jm^i-P^{ij}),$$

hence we obtain

$$\alpha\iint_{\mathcal{E}}\nabla\hat{\boldsymbol{\eta}}:(\nabla\hat{\boldsymbol{m}}-\boldsymbol{P})\,\mathrm{d}A=\alpha\iint_{\mathcal{E}}\left[D_j(\eta_i(D^jm^i-P^{ij}))-\eta_iD_j(D^jm^i-P^{ij})\right]\sqrt{e}\,\mathrm{d}x^1\mathrm{d}x^2.$$

On the first term in this integral we can apply *Stokes' theorem*, [16, p.124]. This gives us

$$\iint_{\mathcal{E}}D_j(\eta_i(D^jm^i-P^{ij}))\sqrt{e}\,\mathrm{d}x^1\mathrm{d}x^2=\oint_{\partial\mathcal{E}}(D^jm^i-P^{ij})\eta_in_j\,\mathrm{d}s,$$

where n_j are the covariant components of the outward unit normal vector on the boundary $\partial\mathcal{E}$. From this we see that

$$\alpha\iint_{\mathcal{E}}\nabla\hat{\boldsymbol{\eta}}:(\nabla\hat{\boldsymbol{m}}-\boldsymbol{P})\,\mathrm{d}A=\alpha\oint_{\partial\mathcal{E}}(D^jm^i-P^{ij})\eta_in_j\,\mathrm{d}s-\alpha\iint_{\mathcal{E}}[D_j(D^jm^i-P^{ij})]\eta_i\sqrt{e}\,\mathrm{d}x^1\mathrm{d}x^2.$$

Combining this result with equation (4.33) we obtain

$$\begin{aligned}0=&-\alpha\iint_{\mathcal{E}}\left[D_j(D^jm^i-P^{ij})\right]\eta_i\sqrt{e}\,\mathrm{d}x^1\mathrm{d}x^2\\ &+\oint_{\partial\mathcal{E}}\left[\alpha(D^jm^i-P^{ij})n_j+(1-\alpha)(m^i-b^i)\right]\eta_i\,\mathrm{d}s,\end{aligned}$$

for all $\boldsymbol{\eta}\in\mathcal{V}$. Invoking the *Fundamental Lemma of Calculus of Variations* we find from this the boundary value problem

$$D_jD^jm^i=D_jP^{ij}\qquad\text{in }\mathcal{E}, \tag{4.34a}$$

$$\alpha(D^jm^i)n_j+(1-\alpha)m^i=\alpha P^{ij}n_j+(1-\alpha)b^i\qquad\text{on }\partial\mathcal{E}. \tag{4.34b}$$

The solution of boundary value problem (4.34) will minimize J for constant $\boldsymbol{P}$ and $\boldsymbol{b}$. Note that this is a vector equation. Equations (4.34a) and (4.34b) are really four equations, two for $i=1$ and two for $i=2$. The term $D_jD^jm^i$ is the so-called *vector Laplacian*. In Cartesian coordinates

$$D_jD^jm^i=\partial_j\partial^jm^i=\Delta m^i.$$

Thus, in Cartesian coordinates the Laplacian of a vector amounts to just taking the Laplacian component-wise. However, in different coordinate systems this is not true, because nonzero Christoffel symbols imply that $[D_j D^j m^i]_{i=1}$ depends on both m^1 and m^2, and similarly for $[D_j D^j m^i]_{i=2}$. This results for an arbitrary coordinate system in two coupled sets of equations, while for Cartesian coordinate systems these two sets decouple. This will become more clear when we derive from (4.34) the coordinate specific boundary value problem for Cartesian coordinates and polar coordinates in next subsection.

4.4.2 The boundary value problem in specific coordinate systems

In Cartesian coordinates the partial differential equations in (4.34) decouple. Let us define

$$\boldsymbol{p}_x = \begin{pmatrix} P^{xx} \\ P^{xy} \end{pmatrix} = \begin{pmatrix} P^{11} \\ P^{12} \end{pmatrix} \quad \text{and} \quad \boldsymbol{p}_y = \begin{pmatrix} P^{yx} \\ P^{yy} \end{pmatrix} = \begin{pmatrix} P^{21} \\ P^{22} \end{pmatrix}.$$

With the use of this definition we can rewrite $[D_j P^{ij}]_{i=x}$ as $\operatorname{div} \boldsymbol{p}_x$ and $[D_j P^{ij}]_{i=y}$ as $\operatorname{div} \boldsymbol{p}_y$. From this we see that in Cartesian coordinates (4.34) reduces to the decoupled set of equations

$$\begin{aligned} \Delta m^x &= \operatorname{div} \boldsymbol{p}_x && \text{in } \mathcal{E}, \\ \alpha(\nabla m^x \mid \boldsymbol{n}) + (1-\alpha) m^x &= \alpha(\boldsymbol{p}_x \mid \boldsymbol{n}) + (1-\alpha) b^x && \text{on } \partial\mathcal{E}, \end{aligned} \tag{4.35a}$$

$$\begin{aligned} \Delta m^y &= \operatorname{div} \boldsymbol{p}_y && \text{in } \mathcal{E}, \\ \alpha(\nabla m^y \mid \boldsymbol{n}) + (1-\alpha) m^y &= \alpha(\boldsymbol{p}_y \mid \boldsymbol{n}) + (1-\alpha) b^y && \text{on } \partial\mathcal{E}. \end{aligned} \tag{4.35b}$$

The boundary value problems (4.35a) and (4.35b) are exactly the boundary value problems for $\boldsymbol{m}$ derived in [5, p.143].

In polar coordinates the equations do not decouple as we shall see soon. Notice that the coordinate specific boundary value problem that we deduce from (4.34) does depend on the choice of basis for polar coordinates, because (4.34) is a vector equation. Thus, we shall find for polar coordinates with an anholonomic basis a boundary value problem different from the one that we shall find when using polar coordinates with its coordinate basis.

In order to derive the boundary value in polar coordinates, let us first write out the components of the covariant derivatives appearing in (4.34) in terms of Christoffel symbols and derivatives. We start out with the vector Laplacian. By the definition of D_j on pages 26 and 27 it follows that

$$\begin{aligned} D_j D^j m^i &= e^{jk} D_j D_k m^i \\ &= e^{jk}\big(\nabla_{\boldsymbol{e}_j}(D_k m^i) - \Gamma^l_{kj} D_l m^i + \Gamma^i_{lj} D_k m^l\big) \\ &= e^{jk}\big(\nabla_{\boldsymbol{e}_j}(\nabla_{\boldsymbol{e}_k} m^i + \Gamma^i_{lk} m^l) - \Gamma^l_{kj}(\nabla_{\boldsymbol{e}_l} m^i + \Gamma^i_{sl} m^s) + \Gamma^i_{lj}(\nabla_{\boldsymbol{e}_k} m^l + \Gamma^l_{sk} m^s)\big) \\ &= e^{jk}\big(\nabla_{\boldsymbol{e}_j}\nabla_{\boldsymbol{e}_k} m^i + \nabla_{\boldsymbol{e}_j}(\Gamma^i_{lk}) m^l + \Gamma^i_{lk}\nabla_{\boldsymbol{e}_j} m^l - \Gamma^l_{kj}\nabla_{\boldsymbol{e}_l} m^i \\ &\quad - \Gamma^l_{kj}\Gamma^i_{sl} m^s + \Gamma^i_{lj}\nabla_{\boldsymbol{e}_k} m^l + \Gamma^i_{lj}\Gamma^l_{sk} m^s\big). \end{aligned} \tag{4.36}$$

Doing the same thing for the divergence of $\boldsymbol{P}$ we obtain*

$$D_j P^{ij} = \delta^k_j D_k P^{ij} = \delta^k_j\big(\nabla_{\boldsymbol{e}_k} P^{ij} + \Gamma^i_{lk} P^{lj} + \Gamma^j_{lk} P^{il}\big) = \nabla_{\boldsymbol{e}_j} P^{ij} + \Gamma^i_{lj} P^{lj} + \Gamma^j_{lj} P^{il}. \tag{4.37}$$

Similarly we find for the normal derivative of $\boldsymbol{m}$ in equation (4.34b) that

$$(D^j m^i) n_j = e^{jk}\big(\nabla_{\boldsymbol{e}_k} m^i + \Gamma^i_{lk} m^l\big) n_j. \tag{4.38}$$

We use (4.36) - (4.38) to determine the boundary value problem (4.34) in polar coordinates. We first consider polar coordinates with its coordinate basis. In Example 2.4.8 we obtained that for

*Note, that due to the symmetry of $\boldsymbol{P}$ it is clear what mean when we speak of the divergence of $\boldsymbol{P}$. It does not matter if we contract D_k with the first or second index of P^{ij}, the result is the same.

polar coordinates with the coordinate basis the only nonzero Christoffel symbols are $\Gamma^r_{\theta\theta} = -r$ and $\Gamma^\theta_{r\theta} = \Gamma^\theta_{\theta r} = r^{-1}$. If we evaluate each term in final expression in (4.36) for $i = r$ we obtain

$$\begin{aligned}
e^{jk}\nabla_{\boldsymbol{e}_j}\nabla_{\boldsymbol{e}_k}m^r &= e^{rr}\frac{\partial^2 m^r}{\partial r^2} + e^{\theta\theta}\frac{\partial^2 m^r}{\partial\theta^2} = \frac{\partial^2 m^r}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 m^r}{\partial\theta^2},\\
e^{jk}\nabla_{\boldsymbol{e}_j}(\Gamma^r_{lk})m^l &= 0,\\
e^{jk}\Gamma^r_{lk}\nabla_{\boldsymbol{e}_j}m^l &= e^{\theta\theta}\Gamma^r_{\theta\theta}\frac{\partial m^\theta}{\partial\theta} = -\frac{1}{r}\frac{\partial m^\theta}{\partial\theta},\\
-e^{jk}\Gamma^l_{kj}\nabla_{\boldsymbol{e}_l}m^r &= -e^{\theta\theta}\Gamma^r_{\theta\theta}\frac{\partial m^r}{\partial r} = \frac{1}{r}\frac{\partial m^r}{\partial r},\\
-e^{jk}\Gamma^l_{kj}\Gamma^r_{sl}m^s &= 0,\\
e^{jk}\Gamma^r_{lj}\nabla_{\boldsymbol{e}_k}m^l &= e^{\theta\theta}\Gamma^r_{\theta\theta}\frac{\partial m^\theta}{\partial\theta} = -\frac{1}{r}\frac{\partial m^\theta}{\partial\theta},\\
e^{jk}\Gamma^r_{lj}\Gamma^l_{sk}m^s &= e^{\theta\theta}\Gamma^r_{\theta\theta}\Gamma^\theta_{r\theta}m^r = -\frac{m^r}{r^2}.
\end{aligned}$$

Adding these terms up we find

$$[D_jD^jm^i]_{i=r} = \frac{\partial^2 m^r}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 m^r}{\partial\theta^2} + \frac{1}{r}\frac{\partial m^r}{\partial r} - \frac{2}{r}\frac{\partial m^\theta}{\partial\theta} - \frac{m^r}{r^2}. \tag{4.39}$$

Doing the similar calculations for the $i = \theta$ component we find

$$[D_jD^jm^i]_{i=\theta} = \frac{\partial^2 m^\theta}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 m^\theta}{\partial\theta^2} + \frac{3}{r}\frac{\partial m^\theta}{\partial r} + \frac{2}{r^3}\frac{\partial m^r}{\partial\theta}. \tag{4.40}$$

In the same way we calculate the expressions for the divergence of $\boldsymbol{P}$. For the $i = r$ component we have

$$\begin{aligned}
[D_jP^{ij}]_{i=r} &= \nabla_{\boldsymbol{e}_j}P^{rj} + \Gamma^r_{lj}P^{lj} + \Gamma^j_{lj}P^{rl}\\
&= \partial_jP^{rj} + \Gamma^r_{lj}P^{lj} + \Gamma^j_{lj}P^{rl}\\
&= \frac{\partial P^{rr}}{\partial r} + \frac{\partial P^{r\theta}}{\partial\theta} - rP^{\theta\theta} + \frac{P^{rr}}{r},
\end{aligned}$$

and, similarly, we find for the $i = \theta$ component that

$$[D_jP^{ij}]_{i=\theta} = \frac{\partial P^{\theta r}}{\partial r} + \frac{\partial P^{\theta\theta}}{\partial\theta} + \frac{P^{r\theta} + 2P^{\theta r}}{r}.$$

Lastly we determine the expression for the boundary derivative of $\boldsymbol{m}$ from (4.38). We find

$$[(D^jm^i)n_j]_{i=r} = e^{jk}\big(\nabla_{\boldsymbol{e}_k}m^r + \Gamma^r_{lk}m^l\big)n_j = (\partial_km^r)n^k - rm^\theta n^\theta$$

and similarly

$$[(D^jm^i)n_j]_{i=\theta} = \frac{m^rn^\theta + m^\theta n^r}{r}.$$

Collecting all these results we find the following set of partial differential equations:

$$\frac{\partial^2 m^r}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 m^r}{\partial\theta^2} + \frac{1}{r}\frac{\partial m^r}{\partial r} - \frac{2}{r}\frac{\partial m^\theta}{\partial\theta} - \frac{m^r}{r^2} = \frac{\partial P^{rr}}{\partial r} + \frac{\partial P^{r\theta}}{\partial\theta} - rP^{\theta\theta} + \frac{P^{rr}}{r} \quad \text{in } \mathcal{E}, \tag{4.41a}$$

$$\alpha(\partial_km^r)n^k - \alpha rm^\theta n^\theta + (1-\alpha)m^r = \alpha P^{rj}n_j + (1-\alpha)b^r \quad \text{on } \partial\mathcal{E}, \tag{4.41b}$$

and

$$\frac{\partial^2 m^\theta}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 m^\theta}{\partial\theta^2} + \frac{3}{r}\frac{\partial m^\theta}{\partial r} + \frac{2}{r^3}\frac{\partial m^r}{\partial\theta} = \frac{\partial P^{\theta r}}{\partial r} + \frac{\partial P^{\theta\theta}}{\partial\theta} + \frac{P^{r\theta} + 2P^{\theta r}}{r} \quad \text{in } \mathcal{E}, \tag{4.42a}$$

$$\alpha(\partial_km^\theta)n^k + \alpha\left(\frac{m^rn^\theta + m^\theta n^r}{r}\right) + (1-\alpha)m^\theta = \alpha P^{\theta j}n_j + (1-\alpha)b^\theta \quad \text{on } \partial\mathcal{E}. \tag{4.42b}$$

The equations (4.41) and (4.42) are coupled, because in the m^θ appears in both (4.41a) and (4.41b), and m^r appears in both (4.42a) and (4.42b).

We will now derive the boundary value problem corresponding to polar coordinates with the orthonormal anholonomic basis of Example 2.4.9. We will again denote these basis vectors by $\overline{\boldsymbol{e}}_r$ and $\overline{\boldsymbol{e}}_\theta$. For this basis there are only two nonzero Christoffel symbols. The two nonzero Christoffel symbols are given by $\overline{\Gamma}^r_{\theta\theta} = -r^{-1}$ and $\overline{\Gamma}^\theta_{r\theta} = r^{-1}$. We substitute these Christoffel symbols in the expressions for $D_j D^j m^i$, $D_j P^{ij}$ and $(D^j m^i)n_j$ given in equations (4.36), (4.37) and (4.38). The calculations are similar to the ones for polar coordinates with the coordinate basis, therefore we will just state the results. For the $i = r$ component we find

$$\begin{aligned}
\big[D_j D^j m^i\big]_{i=r} &= \frac{\partial^2 m^r}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 m^r}{\partial \theta^2} + \frac{1}{r}\frac{\partial m^r}{\partial r} - \frac{2}{r^2}\frac{\partial m^\theta}{\partial \theta} - \frac{m^r}{r^2},\\
\big[D_j P^{ij}\big]_{i=r} &= \frac{\partial P^{rr}}{\partial r} + \frac{1}{r}\frac{\partial P^{r\theta}}{\partial \theta} + \frac{P^{rr} - P^{\theta\theta}}{r},\\
\big[(D^j m^i)n_j\big]_{i=r} &= n^r\left(\frac{\partial m^r}{\partial r}\right) + \frac{n^\theta}{r}\frac{\partial m^r}{\partial \theta} - \frac{m^\theta n^\theta}{r},
\end{aligned}$$

and for the $i = \theta$ component we find

$$\begin{aligned}
\big[D_j D^j m^i\big]_{i=\theta} &= \frac{\partial^2 m^\theta}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 m^\theta}{\partial \theta^2} + \frac{1}{r}\frac{\partial m^\theta}{\partial r} + \frac{2}{r^2}\frac{\partial m^r}{\partial \theta} - \frac{m^\theta}{r^2},\\
\big[D_j P^{ij}\big]_{i=\theta} &= \frac{\partial P^{\theta r}}{\partial r} + \frac{1}{r}\frac{\partial P^{\theta\theta}}{\partial \theta} + \frac{P^{r\theta} + P^{\theta r}}{r},\\
\big[(D^j m^i)n_j\big]_{i=\theta} &= n^r\left(\frac{\partial m^\theta}{\partial r}\right) + \frac{n^\theta}{r}\frac{\partial m^\theta}{\partial \theta} + \frac{m^r n^\theta}{r}.
\end{aligned}$$

Substituting these results in equation (4.34) we obtain the set of coupled partial differential equations in the case of polar coordinates with the orthonormal anholonomic basis. These partial differential equations we find are

$$\frac{\partial^2 m^r}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 m^r}{\partial \theta^2} + \frac{1}{r}\frac{\partial m^r}{\partial r} - \frac{2}{r^2}\frac{\partial m^\theta}{\partial \theta} - \frac{m^r}{r^2} = \frac{\partial P^{rr}}{\partial r} + \frac{1}{r}\frac{\partial P^{r\theta}}{\partial \theta} + \frac{P^{rr} - P^{\theta\theta}}{r} \quad \text{in } \mathcal{E}, \tag{4.43a}$$

$$\alpha\left(n^r\frac{\partial m^r}{\partial r} + \frac{n^\theta}{r}\frac{\partial m^r}{\partial \theta} - \frac{m^\theta n^\theta}{r}\right) + (1-\alpha)m^r = \alpha P^{rj}n_j + (1-\alpha)b^r \quad \text{on } \partial\mathcal{E}, \tag{4.43b}$$

and

$$\frac{\partial^2 m^\theta}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 m^\theta}{\partial \theta^2} + \frac{1}{r}\frac{\partial m^\theta}{\partial r} + \frac{2}{r^2}\frac{\partial m^r}{\partial \theta} - \frac{m^\theta}{r^2} = \frac{\partial P^{\theta r}}{\partial r} + \frac{1}{r}\frac{\partial P^{\theta\theta}}{\partial \theta} + \frac{P^{r\theta} + P^{\theta r}}{r} \quad \text{in } \mathcal{E}, \tag{4.44a}$$

$$\alpha\left(n^r\frac{\partial m^\theta}{\partial r} + \frac{n^\theta}{r}\frac{\partial m^\theta}{\partial \theta} + \frac{m^r n^\theta}{r}\right) + (1-\alpha)m^\theta = \alpha P^{\theta j}n_j + (1-\alpha)b^\theta \quad \text{on } \partial\mathcal{E}. \tag{4.44b}$$

This is the set of equations that we will actually implement in Chapter 5. However, before proceeding to the next chapter, we first have to clear up how to determine function $u : \mathcal{E} \to \mathbb{R}_{>0}$ once we have found a mapping $\boldsymbol{m} \in T\mathcal{E}_{C^2}$ that is a solution to Problem 4.1.2. This will be the topic of next section.

4.5 Determining the reflector surface

Suppose that we have found a mapping $\boldsymbol{m} \in T\mathcal{E}_{C^2}$ that is an approximate numerical solution to Problem 4.1.2. The reflector surface is now determined by obtaining u from $\boldsymbol{m}$. We will in this section briefly describe how to determine u from $\boldsymbol{m}$. This section is based on section 8.6 of [5]. We will again generalize the procedure set forth in [5] for Cartesian coordinates to arbitrary coordinates x^1, x^2 on $\mathcal{E}$, with basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$ and metric e_{ij}. We will derive a boundary

value problem for u and we will give an explicit expressions for this boundary value problem in Cartesian coordinates and polar coordinates.

If $\boldsymbol{m}$ was an exact solution of Problem 4.1.2, then $\nabla\hat{\boldsymbol{m}}$ would be symmetric and there would exist a function u such that $\boldsymbol{m} = \nabla u$. However, in practice $\boldsymbol{m}$ is a numerical approximation and hence $\nabla\hat{\boldsymbol{m}}$ will not be precisely symmetric. To overcome this problem, we will look for a function $u \in C^2(\mathcal{E})$ that minimizes the integral

$$I(u) = \frac{1}{2}\iint_{\mathcal{E}} \|\nabla u - \boldsymbol{m}\|^2 \,\mathrm{d}A. \tag{4.45}$$

We can derive from this minimization problem a boundary value problem by applying Calculus of Variations again. We will consider arbitrary variations $u + \varepsilon v \in C^2(\mathcal{E})$ on u and will set the Fréchet derivative of I equal to zero. The Fréchet derivative of I is given by

$$\begin{aligned}\lim_{\varepsilon\to 0}\frac{I(u+\varepsilon v) - I(u)}{\varepsilon} &= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left[\frac{1}{2}\iint_{\mathcal{E}}\|\nabla(u+\varepsilon v) - \boldsymbol{m}\|^2 \,\mathrm{d}A - \frac{1}{2}\iint_{\mathcal{E}}\|\nabla u - \boldsymbol{m}\|^2 \,\mathrm{d}A\right]\\ &= \lim_{\varepsilon\to 0}\frac{1}{2}\iint_{\mathcal{E}}\varepsilon\|\nabla v\|^2 + 2(\nabla u - \boldsymbol{m} \mid \nabla v)\,\mathrm{d}A\\ &= \iint_{\mathcal{E}}(\nabla u - \boldsymbol{m} \mid \nabla v)\,\mathrm{d}A,\end{aligned}$$

where we used an identity similar to (4.32). We can express the inner product in terms of the components of the vector and the metric as

$$\begin{aligned}(\nabla u - \boldsymbol{m} \mid \nabla v) &= (e^{ik}\nabla_{\boldsymbol{e}_k}u - m^i)(e^{jl}\nabla_{\boldsymbol{e}_l}v)e_{ij}\\ &= (e^{ik}\nabla_{\boldsymbol{e}_k}u - m^i)(\nabla_{\boldsymbol{e}_l}v)\delta^l_i\\ &= (e^{ik}D_k(u) - m^i)D_i(v)\\ &= (D^i(u) - m^i)D_i(v),\end{aligned}$$

where we use $D_i(v)$ to denote the components of the covariant derivative of v, i.e. the components of the gradient $\mathrm{d}v = \nabla_{\boldsymbol{e}_i}(v)\hat{\boldsymbol{e}}^i$.

Setting the Fréchet derivative equal to zero, while writing the integral in the local coordinates on $\mathcal{E}$, we obtain

$$\forall v \in C^2(\mathcal{E}): \qquad \iint_{\mathcal{E}}(D^i(u) - m^i)D_i(v)\sqrt{e}\,\mathrm{d}x^1\mathrm{d}x^2 = 0.$$

With use of the product rule we can rewrite the integrand as

$$(D^i(u) - m^i)D_i(v) = D_i\left[(D^i(u) - m^i)v\right] - D_i\left[D^i(u) - m^i\right]v. \tag{4.46}$$

We substitute this in the integral and apply Stokes' theorem ([16, p.124]) on the first term of the integral:

$$\begin{aligned}&\iint_{\mathcal{E}} D_i\left[(D^i(u) - m^i)v\right] - D_i\left[D^i(u) - m^i\right]D_i(v)\sqrt{e}\,\mathrm{d}x^1\mathrm{d}x^2\\ &\quad= \oint_{\partial\mathcal{E}}(D^i(u) - m^i)n_i v\,\mathrm{d}s - \iint_{\mathcal{E}} D_i\left[D^i(u) - m^i\right]v\sqrt{e}\,\mathrm{d}x^1\mathrm{d}x^2,\end{aligned}$$

where n_i are the covariant components of the outward unit normal to $\partial\mathcal{E}$. From this we see that we end up with

$$\forall v \in C^2(\mathcal{E}): \qquad \oint_{\partial\mathcal{E}}(D^i(u) - m^i)n_i v\,\mathrm{d}s - \iint_{\mathcal{E}} D_i\left[D^i(u) - m^i\right]v\sqrt{e}\,\mathrm{d}x^1\mathrm{d}x^2.$$

We invoke the Fundamental Lemma of Calculus of Variations and find

$$D_i(D^i(u)) = D_i(m^i) \qquad \text{in } \mathcal{E}, \tag{4.47a}$$
$$D^i(u)n_i = m^i n_i \qquad \text{on } \partial\mathcal{E}. \tag{4.47b}$$

Note that $D_i(D^i(u))$ and $D_i(m^i)$ are the Laplacian of u and the divergence of $\boldsymbol{m}$, respectively, in general coordinate systems. Equations (4.47b) states that on $\partial\mathcal{E}$ the boundary derivative of u is equal to the inner product $(\boldsymbol{m} \mid \boldsymbol{n})$. The boundary value problem (4.47) is the same boundary value problem as was derived in [5] in Cartesian coordinates. In Cartesian coordinates, we have the familiar expressions for the Laplacian and the divergence, hence we find that in Cartesian coordinates (4.47) is given by

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} &= \frac{\partial m^x}{\partial x} + \frac{\partial m^y}{\partial y} && \text{in } \mathcal{E}, \\ \frac{\partial u}{\partial x}n^x + \frac{\partial u}{\partial y}n^y &= m^x n^x + m^y n^y && \text{on } \partial\mathcal{E}. \end{aligned}$$

Lastly, let us determine the explicit form of (4.47) in polar coordinates. We start by calculating the Laplacian and the divergence in the polar coordinate system. For the Laplacian of u we find

$$\begin{aligned} D_i(D^i(u)) &= e^{ik} D_i(D_k(u)) \\ &= e^{ik}\big(\nabla_{\boldsymbol{e}_i}(D_k(u)) - \Gamma^j_{ki} D_j(u)\big) \\ &= e^{ik}\nabla_{\boldsymbol{e}_i}(\nabla_{\boldsymbol{e}_k}(u)) - e^{ik}\Gamma^j_{ki}\nabla_{\boldsymbol{e}_j}(u) \\ &= e^{ik}\partial_i(\partial_k(u)) - \frac{1}{r^2}\Gamma^r_{\theta\theta}\partial_r(u) \\ &= \frac{\partial^2 u}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} + \frac{1}{r}\frac{\partial u}{\partial r} \\ &= \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}. \end{aligned}$$

The divergence of $\boldsymbol{m}$ in the coordinate basis is given by

$$\begin{aligned} D_i(m^i) &= e^{ik} D_i(m_k) \\ &= e^{ik}\partial_i(m_k) - e^{ik}\Gamma^j_{ki} m_j \\ &= e^{ik}\partial_i(e_{kl}m^l) + e^{\theta\theta}\Gamma^r_{\theta\theta} m_r \\ &= e^{ik}\partial_i(e_{kl})m^l + \delta^i_l\partial_i(m^l) + \frac{m^r}{r} \\ &= \partial_i(m^i) + \frac{m^r}{r} \\ &= \frac{1}{r}\frac{\partial(rm^r)}{\partial r} + \frac{\partial m^\theta}{\partial \theta}. \end{aligned}$$

In the orthonormal anholomic basis the divergence of $\boldsymbol{m}$ is given by

$$\begin{aligned} D_i(m^i) &= e^{ik} D_i(m_k) \\ &= e^{ki}\big(\nabla_{\boldsymbol{e}_i}(m_k) - \Gamma^j_{ki} m_j\big) \\ &= \frac{\partial m^r}{\partial r} + \frac{1}{r}\frac{\partial m^\theta}{\partial \theta} - e^{\theta\theta}\Gamma^r_{\theta\theta} m_r \\ &= \frac{\partial m^r}{\partial r} + \frac{1}{r}\frac{\partial m^\theta}{\partial \theta} + \frac{m^r}{r} \\ &= \frac{1}{r}\frac{\partial(rm^r)}{\partial r} + \frac{1}{r}\frac{\partial m^\theta}{\partial \theta}. \end{aligned}$$

From these calculations we see that in polar coordinates with the coordinate basis boundary value problem (4.47) is given by

$$\begin{aligned} \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right)+\frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} &= \frac{1}{r}\frac{\partial (rm^r)}{\partial r}+\frac{\partial m^\theta}{\partial \theta} \quad \text{in } \mathcal{E}, \\ \frac{\partial u}{\partial r}n^r+\frac{\partial u}{\partial \theta}n^\theta &= m^r n^r + r^2(m^\theta n^\theta) \quad \text{on } \partial\mathcal{E}, \end{aligned} \tag{4.48}$$

while in the orthonormal anholonomic basis (4.47) is given by

$$\begin{aligned} \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right)+\frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} &= \frac{1}{r}\frac{\partial (rm^r)}{\partial r}+\frac{1}{r}\frac{\partial m^\theta}{\partial \theta} \quad \text{in } \mathcal{E}, \\ \frac{\partial u}{\partial r}n^r+\frac{\partial u}{\partial \theta}\frac{n^\theta}{r} &= m^r n^r + m^\theta n^\theta \qquad \text{on } \partial\mathcal{E}. \end{aligned} \tag{4.49}$$

In equations (4.48) the components of $\boldsymbol{m}$ are given with respect to the coordinate basis $\{\boldsymbol{e}_r, \boldsymbol{e}_\theta\}$, i.e. $\boldsymbol{m} = m^r\boldsymbol{e}_r + m^\theta\boldsymbol{e}_\theta$, and similarly for $\boldsymbol{n}$. In equations (4.49) the components of $\boldsymbol{m}$ are given with respect to the orthonormal basis $\{\overline{\boldsymbol{e}}_r = \boldsymbol{e}_r, \overline{\boldsymbol{e}}_\theta = \boldsymbol{e}_\theta/r\}$, i.e. $\boldsymbol{m} = m^r\overline{\boldsymbol{e}}_r + m^\theta\overline{\boldsymbol{e}}_\theta = m^r\boldsymbol{e}_r + (m^\theta/r)\boldsymbol{e}_\theta$, and again similarly for $\boldsymbol{n}$. From this we see that the two boundary value problem (4.48) and (4.49) are the same.

Let us now briefly summarize what we have done in this chapter. We have presented the Least-squares method in detail. The Least-squares method will find a mapping $\boldsymbol{m} \in T\mathcal{E}_{C^1}$ that solves Problem 4.1.2. The mapping $\boldsymbol{m}$ will be such that there exists a convex function $u : \mathcal{E} \to \mathbb{R}_{>0}$ for which holds that $\boldsymbol{m} = \nabla u$, moreover this u will satisfy Problem 3.5.7. The Least-squares method numerically determines such a mapping $\boldsymbol{m}$ by an iterative process in which the functional $J(\boldsymbol{m}, \boldsymbol{Pb})$ in (4.6) gets minimized. Each iteration consists of three steps in which $J(\boldsymbol{m}, \boldsymbol{Pb})$ gets minimized for one of its three arguments. In the first step the function J gets minimized for $\boldsymbol{b}$ while keeping $\boldsymbol{P}$ and $\boldsymbol{m}$ fixed. This minimization step can be performed point-wise for each grid point on the boundary. This is described in Section 4.2. In the second step J gets minimized for $\boldsymbol{P}$ while keeping $\boldsymbol{m}$ and $\boldsymbol{b}$ fixed. In Section 4.3 we showed how this minimization step can also be performed point-wise and algebraically. In the final third step J gets minimized for $\boldsymbol{m}$, while keeping $\boldsymbol{b}$ and $\boldsymbol{P}$ fixed. In Section 4.4 we showed that this is done by solving two boundary value problems. Moreover, we showed that in a general coordinate system these two boundary value problems are coupled, while they decouple in Cartesian coordinates. Now that we in this section have shown how to determine the reflector surface from $\boldsymbol{m}$, the solution to Problem 4.1.2, we are ready to proceed to the next chapter. In the next chapter we will focus on the implementation of the Least-squares method in polar coordinates. Furthermore, we will show that the Least-squares method actually works by presenting numerical results.

Chapter 5

Implementation and Numerical Results

This chapter will consist of two parts. In the first part we will discuss the numerical implementation of the least-squares method and in the second part we will focus on numerical results. In Chapter 1 we saw that the light source of interest to us has a disk-like, shape, i.e.,

$$\mathcal{E} = \mathcal{D}_R := \{x \in \mathbb{R}^2 \mid \|x - x_0\| < R\}, \tag{5.1}$$

where $R > 0$ is the radius of the disk and $x_0 \in \mathbb{R}^2$ is the center of the disk. From now on, when working in Cartesian or polar coordinates we will take x_0 to be the center of the coordinate system. In [5] the least-squares method was presented and implemented in Cartesian coordinates only. We will discuss in the first section, Section 5.1, how arbitrary shaped sources are treated in the Cartesian coordinate least-squares algorithm of [5]. We will in Section 5.1 discuss the consequences resulting from using a grid that does not nicely fit the geometry of the source $\mathcal{E}$. We will also discuss what the consequences of this are when we want to extrapolate the reflector surface by extending the least-squares method to an extension of $\mathcal{E}$. This extension of the least-squares method will be the subject of Chapter 6. The discussion in Section 5.1 will motivate the use of a polar coordinate grid that does fit the geometry of $\mathcal{E}$.

The implementation of the least-squares method in polar coordinates will be the topic of Section 5.2. We will in that section mostly focus on the implementation of the minimization problem of Section 4.4, because this is the part of the least-squares method that changes the most when switching to polar coordinates.

Lastly, we will in Section 5.3 compare the implementation in Cartesian coordinates with the implementation in polar coordinates. We will analyze the convergence of both implementations for the source $\mathcal{E}$ as specified in (5.1) and three target distributions $\mathcal{F}$ of increasing complexity. The first target we will consider will be a simple square, the second target will be non-convex and the third case will be a target intensity corresponding to a famous painting.

5.1 Implementation for Cartesian coordinates

General shapes of $\mathcal{E}$ are handled by taking the smallest bounding box $B = [x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}]$ such that $\mathcal{E} \subset B$. (See [5, p.131].) The source emittance function $E : \mathcal{E} \to \mathbb{R}_{>0}$ is then extend to B by setting $E(x) = 0$ for $x \in B \backslash \mathcal{E}$. This bounding box B is then covered with a Cartesian grid. The Least-squares method will then be performed with B as source. The only difference with the method as represented in Chapter 4, is that the source emittance E is allowed to equal zero on part of the source B.

The fact that the method works with B as source instead of $\mathcal{E}$ has as a consequence that it will try to satisfy the boundary condition $\boldsymbol{m}(\partial B) = \partial \mathcal{F}$ instead of the desired boundary condition

$\boldsymbol{m}(\mathcal{E}) = \partial\mathcal{F}$. Thus instead of Problem 4.1.2, the Least-squares method will try to solve the following problem.

Problem 5.1.1. Find $\boldsymbol{m} \in TB_{C^1}$ that satisfies

$$\frac{\det(\nabla\hat{\boldsymbol{m}}(x))}{e} = \frac{E(x)}{F(\boldsymbol{m}(x))}, \qquad \text{in } \mathcal{E}, \tag{5.2a}$$

$$\det(\nabla\hat{\boldsymbol{m}}(x)) = 0, \qquad \text{in } B\backslash\mathcal{E}, \tag{5.2b}$$

$$\boldsymbol{m}(\partial B) = \partial\mathcal{F}, \tag{5.2c}$$

and for which $\nabla\hat{\boldsymbol{m}}$ is a symmetric positive semi-definite tensor. The functions $E : \mathcal{E} \to \mathbb{R}_{>0}$ and $F : \mathcal{F} \to \mathbb{R}_{>0}$ are such that

$$\int_{\mathcal{E}} E(x^1, x^2)\sqrt{e}\ \mathrm{d}x^1\mathrm{d}x^2 = \int_{\mathcal{F}} F(y^1, y^2)\sqrt{f}\ \mathrm{d}y^1\mathrm{d}y^2,$$

where x^1, x^2 are local coordinates on $\mathcal{E}$ with corresponding metric e_{ij} and y^1, y^2 are local coordinates on $\mathcal{F}$ with corresponding metric f_{ij}.

Let us, to simplify the discussion, consider Problem 5.2 in Cartesian coordinates. Equation (5.2b) then becomes

$$\det\begin{pmatrix} \dfrac{\partial m^x}{\partial x} & \dfrac{\partial m^x}{\partial y} \\ \dfrac{\partial m^y}{\partial x} & \dfrac{\partial m^y}{\partial y} \end{pmatrix} = 0,$$

hence we see that on $B\backslash\mathcal{E}$ the Jacobian determinant of the mapping $\boldsymbol{m}$ equals zero. A useful theorem in this context is *Sard's theorem*, [16, p.72]:

Theorem 5.1.2. *Let $\boldsymbol{g} : U \to \mathbb{R}^n$ be continuously differentiable, where $U \subset \mathbb{R}^n$ is open, and let $A = \{x \in U \mid \det(\boldsymbol{J}(\boldsymbol{g})) = 0\}$, where $\boldsymbol{J}(\boldsymbol{g})$ is the Jacobian matrix of $\boldsymbol{g}$. Then $\boldsymbol{g}(A)$ has measure zero.*

The proof of this theorem is given on page 72 of [16]. Suppose $\boldsymbol{m}$ is a solution to Problem 5.2. In Cartesian coordinates $\boldsymbol{m}$ is a continuously differentiable map from B to $\mathcal{F}$. Now if we restrict $\boldsymbol{m}$ to the interior of B, i.e. $\mathrm{int}(B)$, then the map $\boldsymbol{m} : \mathrm{int}(B) \to \mathbb{R}^2$ satisfies the conditions of Theorem 5.1.2. Equation (5.2b) states that the Jacobian of $\boldsymbol{m}$ equals zero on the set $\mathrm{int}(B)\backslash\mathcal{E}$ and hence Theorem 5.1.2 implies that the set $\boldsymbol{m}(\mathrm{int}(B)\backslash\mathcal{E}) \subset \mathcal{F}$ has measure zero. The boundary of B, ∂B, has also measure zero. The map $\boldsymbol{m} : B \to \mathcal{F}$ is continuously differentiable and because B is clearly bounded, the derivatives of $\boldsymbol{m}$ are also bounded. Now, take arbitrary $\boldsymbol{x}, \boldsymbol{y} \in B$. Working in Cartesian coordinates we have

$$\begin{aligned} m^i(\boldsymbol{x}) - m^i(\boldsymbol{y}) &= \int_0^1 \left(\frac{\partial m^i}{\partial x^j}(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))(y^j - x^j)\right)\mathrm{d}t \\ &= \int_0^1 \left(\frac{\partial m^i}{\partial x^j}(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))\mathrm{d}t\right)(y^j - x^j). \end{aligned}$$

This implies that we have

$$\boldsymbol{m}(\boldsymbol{x}) - \boldsymbol{m}(\boldsymbol{y}) = \left(\int_0^1 \boldsymbol{J}(\boldsymbol{m})(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))\mathrm{d}t\right)(\boldsymbol{y} - \boldsymbol{x}),$$

where $\boldsymbol{J}(\boldsymbol{m})(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))$ is the Jacobian matrix of $\boldsymbol{m}$ at $\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})$. As the derivatives of $\boldsymbol{m}$ are bounded, also the Jacobian is bounded, i.e. there exists a $K \in \mathbb{R}_{>0}$ such that $\|\boldsymbol{J}(\boldsymbol{m})(\boldsymbol{x})\| \leq K$

for all $\boldsymbol{x} \in B$. From this we find that

$$\begin{aligned}
\|\boldsymbol{m}(\boldsymbol{y}) - \boldsymbol{m}(\boldsymbol{x})\| &= \| \left(\int_0^1 \boldsymbol{J}(\boldsymbol{m})(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))\mathrm{d}t \right) (\boldsymbol{y} - \boldsymbol{x}) \| \\
&= \| \int_0^1 (\boldsymbol{J}(\boldsymbol{m})(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x}))\, \mathrm{d}t \| \\
&\leq \int_0^1 \|\boldsymbol{J}(\boldsymbol{m})(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x})\| \mathrm{d}t \\
&\leq \int_0^1 K\|(\boldsymbol{y} - \boldsymbol{x})\| \mathrm{d}t \\
&= K\|(\boldsymbol{y} - \boldsymbol{x})\|,
\end{aligned}$$

where we have applied the *Cauchy-Schwarz inequality*. Thus we have shown that there exists a $K > 0$ such that $\|\boldsymbol{m}(\boldsymbol{y}) - \boldsymbol{m}(\boldsymbol{x})\| \leq K\|\boldsymbol{y} - \boldsymbol{x}\|$ for all $\boldsymbol{x}, \boldsymbol{y} \in B$, i.e. we have shown that $\boldsymbol{m} : B \to \mathcal{F}$ is *Lipschitz continuous*. Let $\mu[U]$ denote the (Lebesgue) measure of a set U. Now, the set ∂B has measure zero and therefore there exist for all $\varepsilon > 0$ a collection of balls U_i with radius r_i that together cover ∂B and are such that $\mu[\cup_i U_i] < \varepsilon$. It is clear that we have $\mu[\boldsymbol{m}(\partial B)] \leq \mu[\boldsymbol{m}(\cup_i U_i)]$. From the fact that $\boldsymbol{m}$ is Lipschitz continuous it follows that $\mu[\boldsymbol{m}(\cup_i U_i)] \leq K\mu[\cup_i U_i]$. This implies that $\mu[\boldsymbol{m}(\partial B)] < K\varepsilon$. However, $\varepsilon > 0$ can be taken arbitrarily small and hence we find $\mu[\boldsymbol{m}(\partial B)] = 0$. Thus at last we find that the set $\boldsymbol{m}(B \backslash \mathcal{E})$ has measure zero.

Thus we expect that the grid points in $B \backslash \mathcal{E}$ get mapped very close together. Moreover, the fact that not the boundary condition $\boldsymbol{m}(\partial\mathcal{E}) = \partial\mathcal{F}$ is imposed, but instead the condition $\boldsymbol{m}(\partial\mathcal{E}) = \partial\mathcal{F}$ could lead to convergence problems on the boundary of $\mathcal{E}$ and $\mathcal{F}$. This is indeed what we see in Figure 5.1. Strange bulges appear on the sides of $\partial\mathcal{F}$ and these do not shrink one the number of iterations gets increased. In Chapter 6 it will be seen that these bulges cause serious problems when one wants to extrapolate the reflector surface to a larger source that contains $\mathcal{E}$. An enlargement of parts of the plot in Figure 5.1 can be seen in Figure 5.2.

The convergence issues on the boundary do not only appear for the particular choice for $\mathcal{F}$ shown in Figure 5.1. In Figure 5.3 the mapping can be seen for $\mathcal{F}$ and F corresponding to the famous painting by Vermeer depicted in Figure 5.4. In the subsequent sections we will consider the Least-squares method in polar coordinates and we will see that these defects disappear when we use the polar coordinate grid that precisely fits $\mathcal{E}$.

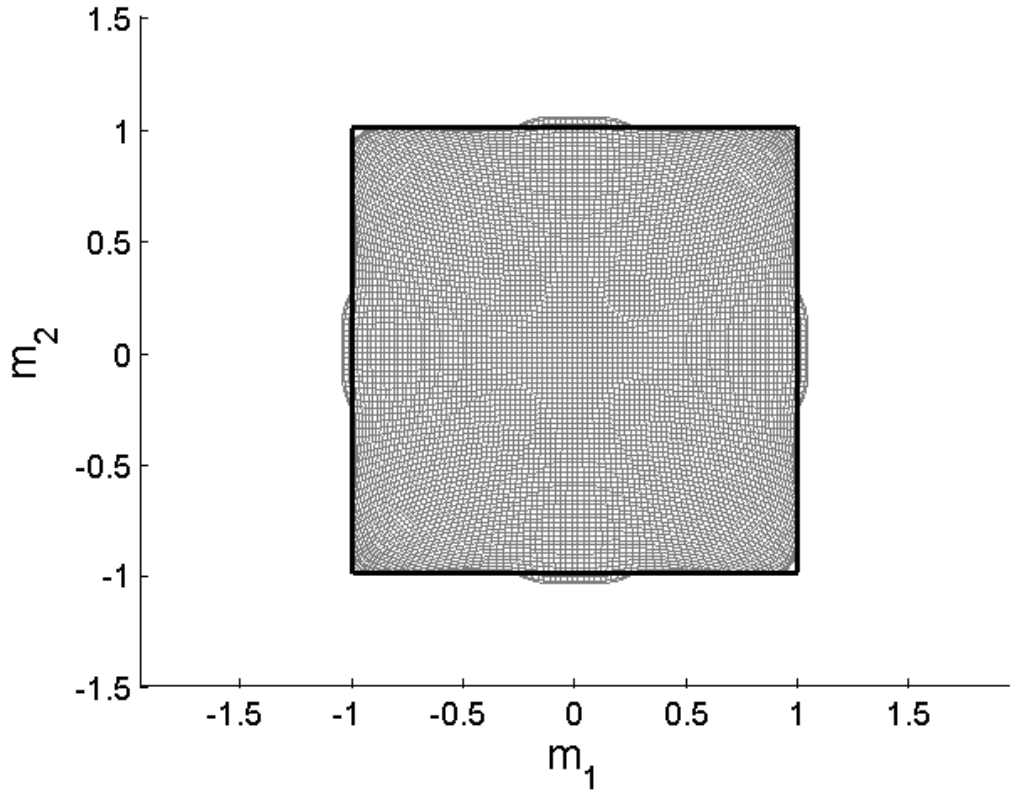


Figure 5.1: The mapping $\boldsymbol{m}$ after 300 iterations on a 100×100 grid with α in (4.6) equal to 0.2, for $\mathcal{E} = \mathcal{D}_1$ and $\mathcal{F} = [-1, 1]^2$. Unwanted "bulges" can be seen on the four edges. The size of these bulges do not decrease for an increased number of iterations.

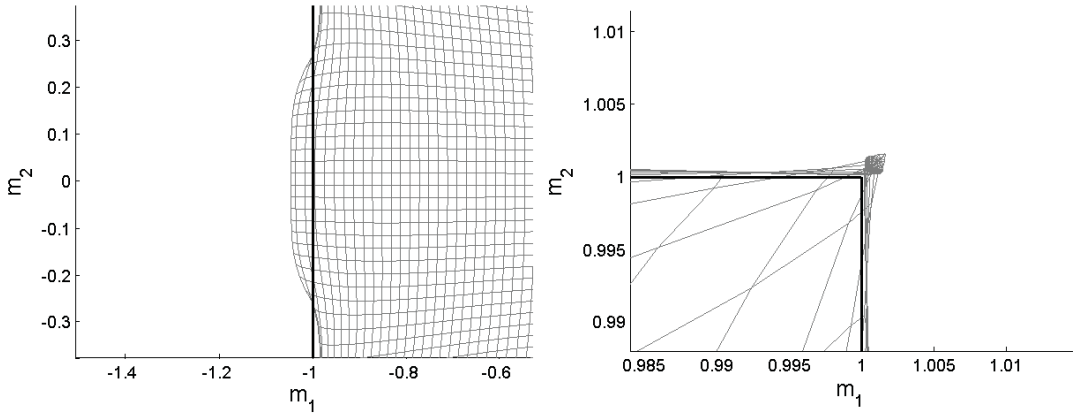


Figure 5.2: Two segments of Figure 5.1 have been enlarged. On the right one sees the upper right corner of Figure 5.1 and on the left one sees the bulge on the left edge of Figure 5.1. In the right plot it can be seen that the grid points in $\mathcal{E} \backslash B$ indeed get squeezed together.

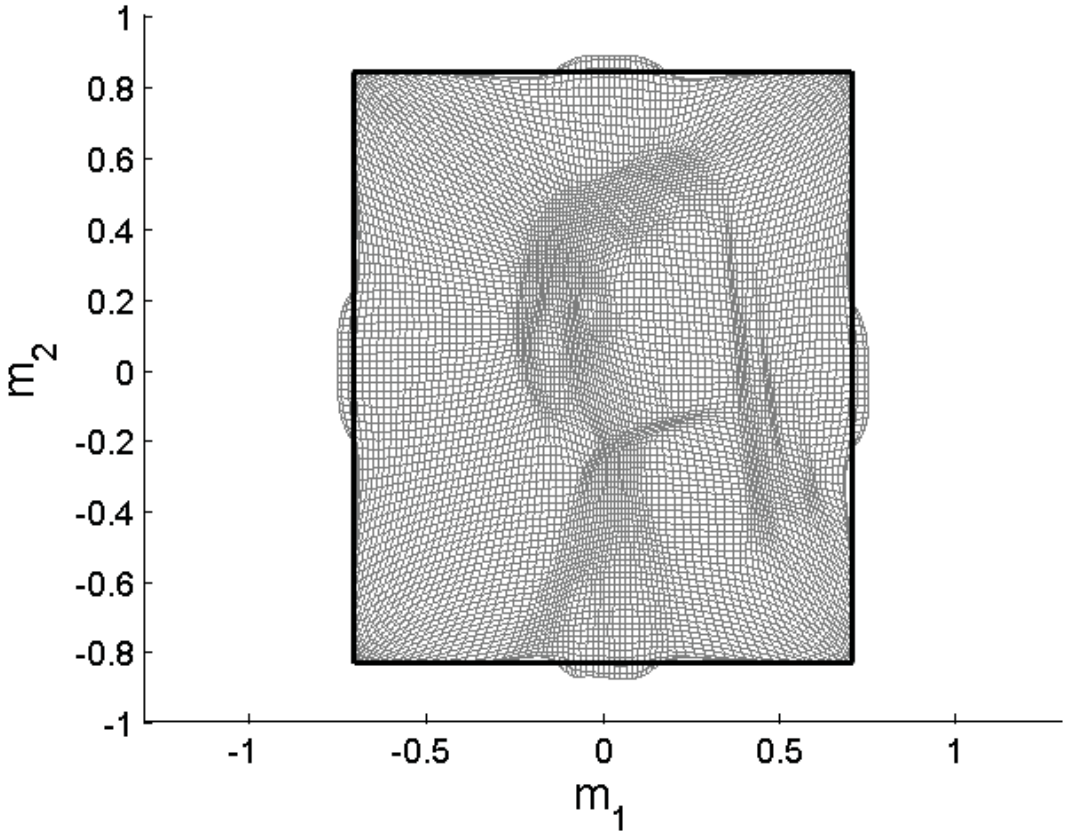


Figure 5.3: This is a plot of the mapping $\boldsymbol{m}$ after 300 iterations on a 100×100 grid with $\alpha = 0.2$. For this plot $\mathcal{F}$ and F correspond to the intensity output that gives the painting in Figure 5.4 when projected on a screen. Again the bulges on the side edges appear.

Figure 5.4: "The Girl with the Pearl Earring" by Johannes Vermeer.

5.2 Implementation for polar coordinates

This section we will focus on the implementation of the Least-squares method. We will cover the source $\mathcal{E} = \mathcal{D}_R$ with a polar coordinate grid. Let N_r and N_θ be the number of grid points along the r- and θ-coordinate lines, respectively. We number the grid points in the following way:

$$r_i = ih_r \qquad h_r := \frac{R}{N_r},$$
$$\theta_j = (j-1)h_\theta \qquad h_\theta := \frac{2\pi}{N_\theta}.$$

This grid nicely fits the geometry of $\mathcal{E}$. The grid points $(r_{N_r}, \theta_j),\ 1 \leq j \leq N_\theta$ all lie on $\partial\mathcal{E}$.

For the first two steps of the Least-squares method, i.e. the minimization of the boundary functional (4.4) and (4.3), are not very different in polar coordinates. The minimization of the boundary integral J_B can be done point-wise and amounts to determining some inner products and taking a minimum over a finite set. These operations translate trivially to the polar coordinate case.

The minimization J_I as described in Section 4.3 is a little bit more involved. The components (d_{ij}) in Problem 4.3.2 are the components of the finite difference approximation of the tensor $\nabla\hat{\boldsymbol{m}}$. The components d_{ij} are given by

$$d_{ij} := \delta_{\boldsymbol{e}_j} m_i - \Gamma^k_{ij} m_k, \tag{5.3}$$

where $\delta_{\boldsymbol{e}_j} m_i$ is the finite difference approximation of $\nabla_{\boldsymbol{e}_j} m_i$. Before we determine the coefficients, we must first determine which basis we use, in order to know what the Christoffel symbols in (5.3) are. We have chosen to use polar coordinates with the orthonormal anholonomic basis. We prefer the orthonormal basis over the coordinate basis, because in the orthonormal basis the matrix representation of the metric e_{ij} is just the identity matrix. This implies that we have $v_i = v^i$ for the components of a vector $\boldsymbol{v}$ and its corresponding covector $\hat{\boldsymbol{v}}$. Thus, in this basis we do not have to worry about the difference between contravariant and covariant components, because there is none. We recall from Example 2.4.9 that the only two nonzero Christoffel symbols for polar coordinate system with the orthonormal basis are $\Gamma^r_{\theta\theta} = -r^{-1}$ and $\Gamma^\theta_{r\theta} = r^{-1}$. Furthermore, we will use the central difference approximation for $\nabla_{\boldsymbol{e}_j} m_i$ and we will write $D_{k,l}$ for the matrix representation of (d_{ij}) in the grid point (r_k, θ_l). Substituting the Christoffel symbols and finite difference approximation in (5.3) we obtain

$$D_{k,l} = \begin{pmatrix} \dfrac{m^r_{k+1,l} - m^r_{k-1,l}}{2h_r} & \dfrac{m^r_{k,l+1} - m^r_{k,l-1}}{2h_\theta r_k} - \dfrac{m^\theta_{k,l}}{r_{k,l}} \\ \dfrac{m^\theta_{k+1,l} - m^\theta_{k-1,l}}{2h_r} & \dfrac{m^\theta_{k,l+1} - m^\theta_{k,l-1}}{2h_\theta r_k} + \dfrac{m^r_{k,l}}{r_{k,l}} \end{pmatrix}, \tag{5.4}$$

for $1 < k < N_r$ and $1 \leq l \leq N_\theta$. Note that we have in (5.4) immediately exploited the fact that $m^r = m_r$ and $m^\theta = m_\theta$. We will from now on use upper indices to indicate vector or covector components in order to reserve the lower indices for grid numbering. For the grid points corresponding to the boundary we use one-sided differences with second order accuracy. The rest of the minimization of J_I remains the same as in the Cartesian coordinate case. For each grid point (r_k, θ_l) the coefficients (d_{ij}) of Problem 4.3.2 are given by the matrix $D_{k,l}$ and for these coefficients Problem 4.3.2 gets solved algebraically as explained in Section 4.3.

So far, we have seen that for the first two steps (4.8a) and (4.8b) of the Least-squares method there are no qualitative difference between the implementation in Cartesian coordinates and in polar coordinates. For the third step (4.8c) of the Least-squares method the situation in polar coordinates is qualitatively different from the one in Cartesian coordinates. In Section 4.4 we found that, while in Cartesian coordinates the minimizer $\boldsymbol{m}$ for J in (4.8c) was the solution of the decoupled system of partial differential equations (4.35a) and (4.35b), in the polar coordinate system we end up with the coupled system of partial differential equations (4.43) and (4.44). The

fact that the source $\mathcal{E}$ is circular implies that the unit outward normal on $\partial\mathcal{E}$ is given by $\boldsymbol{n} = \boldsymbol{e}_r$. Using this equations (4.43) and (4.44) simplify to

$$\frac{\partial^2 m^r}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 m^r}{\partial \theta^2} + \frac{1}{r}\frac{\partial m^r}{\partial r} - \frac{2}{r^2}\frac{\partial m^\theta}{\partial \theta} - \frac{m^r}{r^2} = \frac{\partial P^{rr}}{\partial r} + \frac{1}{r}\frac{\partial P^{r\theta}}{\partial \theta} + \frac{P^{rr} - P^{\theta\theta}}{r} \quad \text{in } \mathcal{E}, \tag{5.5a}$$

$$\alpha\frac{\partial m^r}{\partial r} + (1-\alpha)m^r = \alpha P^{rr} + (1-\alpha)b^r \quad \text{on } \partial\mathcal{E}, \tag{5.5b}$$

and

$$\frac{\partial^2 m^\theta}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 m^\theta}{\partial \theta^2} + \frac{1}{r}\frac{\partial m^\theta}{\partial r} + \frac{2}{r^2}\frac{\partial m^r}{\partial \theta} - \frac{m^\theta}{r^2} = \frac{\partial P^{\theta r}}{\partial r} + \frac{1}{r}\frac{\partial P^{\theta\theta}}{\partial \theta} + \frac{P^{r\theta} + P^{\theta r}}{r} \quad \text{in } \mathcal{E}, \tag{5.6a}$$

$$\alpha\frac{\partial m^\theta}{\partial r} + (1-\alpha)m^\theta = \alpha P^{\theta r} + (1-\alpha)b^\theta \quad \text{on } \partial\mathcal{E}. \tag{5.6b}$$

To solve this coupled system of equations we will need to iterate between equations (5.5) and (5.6). We start with a solution $\boldsymbol{m}^n$ of the preceding iteration or an initial guess $\boldsymbol{m}^0$. We will then determine $\boldsymbol{m}^{n+1}$ in the following way. We start with solving equation (5.5) for m^r while keeping m^θ, the θ-component of $\boldsymbol{m}^n$ fixed. The solution of (5.5) obtained in this way we will call u^1. We then turn our attention to (5.6). We will solve (5.6) for m^θ while keeping m^r fixed and equal to u^1. The solution of (5.6) that we obtain in this way we call v^1 Then we solve again (5.5) for m^r with $m^\theta = v^1$ fixed. This gives us u^2 and so we proceed, iterating between (5.5) and (5.6) until a desired convergence is achieved. In practice one does not want to keep iterating until u^k and v^k have converged to machine precision, because this slows down the Least-squares method significantly as this iterative subprocess has to be performed each iteration of the Least-squares method. An optimal number of iterations K has to be sought such that the Least-squares method converges most fast. After halting the iterative process for u^k and v^k after K iterations the new mapping $\boldsymbol{m}^{n+1}$ is given by the r-component $m^r = u^K$ and the θ-component v^K. In practice it turns out that for most problems, u^k and v^k converge quite fast and that K lies somewhere between 5 and 15. It might also be beneficial to let K depend on n, the overal iteration number of the Least-squares method. However, this has not yet been investigated.

We will now discretize equations (5.5) and (5.6). We start by discretizing for the grid points (r_k, θ_l) with $2 \le k \le N_r$. We will treat the grid points directly adjacent to the origin afterwards. We start with equation (5.5a). Using second order central differences (5.5a) can be approximated by

$$\begin{aligned}
&\frac{1}{r_k h_r^2}\left[r_{k+\frac{1}{2}}(m^r_{k+1,l} - m^r_{k,l}) - r_{k-\frac{1}{2}}(m^r_{k,l} - m^r_{k-1,l})\right] + \frac{1}{r_k^2 h_\theta^2}\left[m^r_{k,l+1} - 2m^r_{k,l} + m^r_{k,l-1}\right] \\
&- \frac{1}{r_k^2 h_\theta}\left[m^\theta_{k,l+1} - m^\theta_{k,l-1}\right] - \frac{m^r_{k,l}}{r_j^2} \\
&= \frac{1}{2h_r}\left[P^{rr}_{k+1,l} - P^{rr}_{k-1,l}\right] + \frac{1}{2r_k h_\theta}\left[P^{r\theta}_{k,l+1} - P^{r\theta}_{k,l-1}\right] + \frac{1}{r_k}\left[P^{rr}_{k,l} - P^{\theta\theta}_{k,l}\right],
\end{aligned} \tag{5.7}$$

where make the identifications $(r_k, \theta_{N_\theta+1}) = (r_k, \theta_1)$ and $(r_k, \theta_0) = (r_k, \theta_{N_\theta})$ for all $1 \le k \le N_r$. Equation (5.7) holds for all grid points (r_k, θ_l), with $2 \le k \le N_r - 1$ and $1 \le l \le N_\theta$. We also discretize the boundary equation (5.5b) with second order accuracy and find

$$\frac{\alpha}{2h_r}\left[m^r_{N_r+1,l} - m^r_{N_r-1,l}\right] + (1-\alpha)m^r_{N_r,l} = \alpha P^{rr}_{N_r,l} + (1-\alpha)b^r_{N_r,l}, \tag{5.8}$$

which holds for $1 \le l \le N_\theta$. Solving this equation for the grid point outside our grid we find

$$m^r_{N_r+1,l} = 2h_r P^{rr}_{N_r,l} + 2h_r(1/\alpha - 1)(b^r_{N_r,l} - m^r_{N_r,l}) + m^r_{N_r-1,l}. \tag{5.9}$$

For the grid points on the boundary we can also write down (5.7):

$$\begin{aligned}
&\frac{1}{r_{N_r} h_r^2}\left[r_{N_r+\frac{1}{2}}(m^r_{N_r+1,l} - m^r_{N_r,l}) - r_{N_r-\frac{1}{2}}(m^r_{N_r,l} - m^r_{N_r-1,l})\right] \\
&+ \frac{1}{r_{N_r}^2 h_\theta^2}\left[m^r_{N_r,l+1} - 2m^r_{N_r,l} + m^r_{N_r,l-1}\right] - \frac{1}{r_{N_r}^2 h_\theta}\left[m^\theta_{N_r,l+1} - m^\theta_{N_r,l-1}\right] - \frac{m^r_{N_r,l}}{r_j^2} \\
&= \frac{1}{2h_r}\left[P^{rr}_{N_r+1,l} - P^{rr}_{N_r-1,l}\right] + \frac{1}{2r_{N_r}h_\theta}\left[P^{r\theta}_{N_r,l+1} - P^{r\theta}_{N_r,l-1}\right] + \frac{1}{r_{N_r}}\left[P^{rr}_{N_r,l} - P^{\theta\theta}_{N_r,l}\right],
\end{aligned}$$

for $1 \leq l \leq N_\theta$. We will now eliminate $m^r_{N_r+1,l}$ from this equation by replacing it with (5.9). Moreover we will replace the central difference approximation for $\partial_r P^{rr}$ by a one-sided difference approximation of the same order of accuracy. This gives us

$$\begin{aligned}
&\frac{2}{h_r^2}\left[m^r_{N_r-1,l} - m^r_{N_r,l}\right] - \left(\frac{2r_{N_r+\frac{1}{2}}(1/\alpha - 1)}{h_r r_{N_r}}\right) m^r_{N_r,l} \\
&+ \frac{1}{r_{N_r}^2 h_\theta^2}\left[m^r_{N_r,l+1} - 2m^r_{N_r,l} + m^r_{N_r,l-1}\right] - \frac{1}{r_{N_r}^2 h_\theta}\left[m^\theta_{N_r,l+1} - m^\theta_{N_r,l-1}\right] - \frac{m^r_{N_r,l}}{r^2_{N_r}} \\
&= \frac{1}{2h_r}\left[3P^{rr}_{N_r,l} - 4P^{rr}_{N_r-1,l} + P^{rr}_{N_r-2,l}\right] - \left(\frac{2r_{N_r+\frac{1}{2}}}{r_{N_r}h_r}\right)\left[P^{rr}_{N_r,l} + (1/\alpha - 1)b^r_{N_r,l}\right] \\
&+ \frac{1}{2r_{N_r}h_\theta}\left[P^{r\theta}_{N_r,l+1} - P^{r\theta}_{N_r,l-1}\right] + \frac{1}{r_{N_r}}\left[P^{rr}_{N_r,l} - P^{\theta\theta}_{N_r,l}\right].
\end{aligned} \tag{5.10}$$

In a similar way we discretize equations (5.6a) and (5.6b). This gives us

$$\begin{aligned}
&\frac{1}{r_k h_r^2}\left[r_{k+\frac{1}{2}}(m^\theta_{k+1,l} - m^\theta_{k,l}) - r_{k-\frac{1}{2}}(m^\theta_{k,l} - m^\theta_{k-1,l})\right] + \frac{1}{r_k^2 h_\theta^2}\left[m^\theta_{k,l+1} - 2m^\theta_{k,l} + m^\theta_{k,l-1}\right] \\
&+ \frac{1}{r_k^2 h_\theta}\left[m^r_{k,l+1} - m^r_{k,l-1}\right] - \frac{m^\theta_{k,l}}{r_k^2} \\
&= \frac{1}{2h_r}\left[P^{\theta r}_{k+1,l} - P^{\theta r}_{k-1,l}\right] + \frac{1}{2r_k h_\theta}\left[P^{\theta\theta}_{k,l+1} - P^{\theta\theta}_{k,l-1}\right] + \frac{1}{r_k}\left[P^{r\theta}_{k,l} + P^{\theta r}_{k,l}\right],
\end{aligned} \tag{5.11}$$

for $2 \leq k \leq N_r - 1$ and $1 \leq l \leq N_\theta$, and

$$\begin{aligned}
&\frac{2}{h_r^2}\left[m^\theta_{N_r-1,l} - m^\theta_{N_r,l}\right] - \left(\frac{2r_{N_r+\frac{1}{2}}(1/\alpha - 1)}{h_r r_{N_r}}\right) m^\theta_{N_r,l} \\
&+ \frac{1}{r_{N_r}^2 h_\theta^2}\left[m^\theta_{N_r,l+1} - 2m^\theta_{N_r,l} + m^\theta_{N_r,l-1}\right] + \frac{1}{r_{N_r}^2 h_\theta}\left[m^r_{N_r,l+1} - m^r_{N_r,l-1}\right] - \frac{m^\theta_{N_r,l}}{r^2_{N_r}} \\
&= \frac{1}{2h_r}\left[3P^{\theta r}_{N_r,l} - 4P^{\theta r}_{N_r-1,l} + P^{\theta r}_{N_r-2,l}\right] - \left(\frac{2r_{N_r+\frac{1}{2}}}{r_{N_r}h_r}\right)\left[P^{\theta r}_{N_r,l} + (1/\alpha - 1)b^\theta_{N_r,l}\right] \\
&+ \frac{1}{2r_{N_r}h_\theta}\left[P^{\theta\theta}_{N_r,l+1} - P^{\theta\theta}_{N_r,l-1}\right] + \frac{1}{r_{N_r}}\left[P^{r\theta}_{N_r,l} + P^{\theta r}_{N_r,l}\right],
\end{aligned} \tag{5.12}$$

for $1 \leq l \leq N_\theta$.

Let us now focus on the grid points adjacent to the origin, i.e. the grid points (r_1, θ_l), where $1 \leq l \leq N_\theta$. We must take care of the fact that the polar coordinate system has a singularity in the origin. For these grid points equations (5.7) and (5.11) will also contain the grid point at the origin. We will denote the value of m^r and m^θ at the origin by m^r_0 and m^θ_0, respectively. Due to the singularity at the origin, m^r_0 and m^θ_0 are not well-defined. We will therefore eliminate them from the equations. We will do this by considering a control area around the origin and use this to derive an expression for m^r_0 and m^θ_0 in terms of values defined at neighbouring grid points.

We first consider equation (5.5a). We multiply this equation by r^2 and integrate it over a control area around the origin. As the control area we take $\mathcal{D}_{h_r}$, i.e. the disk with radius h_r. We get

$$\int_{\mathcal{D}_{h_r}} \left(\Delta m^r - \frac{2}{r^2} \frac{\partial m^\theta}{\partial \theta} - \frac{m^r}{r^2} \right) r^2 \, \mathrm{d}A = \int_{\mathcal{D}_{h_r}} \left(\nabla \cdot \boldsymbol{P}^r + \frac{P^{rr} - P^{\theta\theta}}{r} \right) r^2 \, \mathrm{d}A, \tag{5.13}$$

where $\boldsymbol{P}^r = P^{rr}\boldsymbol{e}_r + P^{\theta\theta}\boldsymbol{e}_\theta$ and $\nabla \cdot \boldsymbol{P}^r$ is the divergence of $\boldsymbol{P}^r$. Notice that we used the fact that the first three terms on the left hand side of (5.5a) are really the Laplacian of m^r and the first two terms on the right hand side of (5.5a) are the divergence of the vector $\boldsymbol{P}^r$ as just defined. We will now rewrite most of the terms in equation (5.13) with use of the divergence theorem. For the first term in (5.13) we find by applying the divergence theorem two times that

$$\begin{aligned} \int_{\mathcal{D}_{h_r}} r^2 \Delta m^r \, \mathrm{d}A &= \oint_{\partial \mathcal{D}_{h_r}} r^2 (\nabla m^r \mid \boldsymbol{e}_r) \, \mathrm{d}s - \int_{\mathcal{D}_{h_r}} (\nabla (r^2) \mid \nabla m^r) \, \mathrm{d}A \\ &= \oint_{\partial \mathcal{D}_{h_r}} r^2 \frac{\partial m^r}{\partial r} \, \mathrm{d}s - \oint_{\partial \mathcal{D}_{h_r}} 2 r m^r \, \mathrm{d}s + \int_{\mathcal{D}_{h_r}} 2 m^r \, \mathrm{d}A. \end{aligned}$$

By the periodicity of $\boldsymbol{m}$ along the θ-coordinate lines we find for the second term in (5.13)

$$\int_{\mathcal{D}_{h_r}} 2 \frac{\partial m^\theta}{\partial \theta} \, \mathrm{d}A = \int_0^{h_r/2} 2r \int_0^{2\pi} \frac{\partial m^\theta}{\partial \theta} \mathrm{d}\theta \mathrm{d}r = \int_0^{h_r/2} 2r \big(m^\theta(r, 2\pi) - m^\theta(r, 0) \big) \, \mathrm{d}r = 0.$$

From this it follows that

$$\begin{aligned} \int_{\mathcal{D}_{h_r}} \left(\Delta m^r - \frac{2}{r^2} \frac{\partial m^\theta}{\partial \theta} - \frac{m^r}{r^2} \right) r^2 \, \mathrm{d}A = \int_{\mathcal{D}_{h_r}} m^r \, \mathrm{d}A - \oint_{\partial \mathcal{D}_{h_r}} 2 r m^r \, \mathrm{d}s \\ + \oint_{\partial \mathcal{D}_{h_r}} r^2 \frac{\partial m^r}{\partial r} \, \mathrm{d}s. \end{aligned} \tag{5.14}$$

Applying the divergence theorem on the first term on the right hand side of equation (5.13) gives us

$$\begin{aligned} \int_{\mathcal{D}_{h_r}} r^2 \nabla \cdot \boldsymbol{P}^r \, \mathrm{d}A &= \oint_{\partial \mathcal{D}_{h_r}} r^2 (\boldsymbol{P}^r \mid \boldsymbol{e}_r) \, \mathrm{d}s - \int_{\mathcal{D}_{h_r}} (\nabla (r^2) \mid \boldsymbol{P}^r) \, \mathrm{d}A \\ &= \oint_{\partial \mathcal{D}_{h_r}} r^2 P^{rr} \, \mathrm{d}s - \int_{\mathcal{D}_{h_r}} 2 r P^{rr} \, \mathrm{d}A. \end{aligned}$$

This shows that

$$\int_{\mathcal{D}_{h_r}} \left(\nabla \cdot \boldsymbol{P}^r + \frac{P^{rr} - P^{\theta\theta}}{r} \right) r^2 \, \mathrm{d}A = \oint_{\partial \mathcal{D}_{h_r}} r^2 P^{rr} \, \mathrm{d}s - \int_{\mathcal{D}_{h_r}} r \big(P^{rr} + P^{\theta\theta} \big) \, \mathrm{d}A. \tag{5.15}$$

Combining equations (5.13), (5.14) and (5.15), we obtain

$$\begin{aligned} \int_{\mathcal{D}_{h_r}} m^r \, \mathrm{d}A - \oint_{\partial \mathcal{D}_{h_r}} 2 r m^r \, \mathrm{d}s + \oint_{\partial \mathcal{D}_{h_r}} r^2 \frac{\partial m^r}{\partial r} \, \mathrm{d}s \\ = \oint_{\partial \mathcal{D}_{h_r}} r^2 P^{rr} \, \mathrm{d}s - \int_{\mathcal{D}_{h_r}} r \big(P^{rr} + P^{\theta\theta} \big) \, \mathrm{d}A. \end{aligned} \tag{5.16}$$

Now we turn our attention to (5.6a). We also multiply this equation by r^2, integrate and obtaining

$$\int_{\mathcal{D}_{h_r}} \left(\Delta m^\theta + \frac{2}{r^2} \frac{\partial m^r}{\partial \theta} - \frac{m^\theta}{r^2} \right) r^2 \, \mathrm{d}A = \int_{\mathcal{D}_{h_r}} \left(\nabla \cdot \boldsymbol{P}^\theta + \frac{2 P^{\theta r}}{r} \right) r^2 \, \mathrm{d}A, \tag{5.17}$$

where $\boldsymbol{P}^\theta = P^{\theta r}\boldsymbol{e}_r + P^{\theta\theta}\boldsymbol{e}_\theta$ and we used the symmetry of $\boldsymbol{P}$, i.e. $P^{r\theta} = P^{\theta r}$. Doing calculations similar to the ones above we find the identities

$$\begin{aligned}
\int_{\mathcal{D}_{h_r}} r^2 \Delta m^\theta \,\mathrm{d}A &= \oint_{\partial\mathcal{D}_{h_r}} r^2 \frac{\partial m^\theta}{\partial r}\,\mathrm{d}s - \oint_{\partial\mathcal{D}_{h_r}} 2rm^\theta \,\mathrm{d}s + \int_{\mathcal{D}_{h_r}} 2m^\theta \,\mathrm{d}A \\
\int_{\mathcal{D}_{h_r}} 2\frac{\partial m^r}{\partial\theta}\,\mathrm{d}A &= 0 \\
\int_{\mathcal{D}_{h_r}} r^2 \nabla\cdot\boldsymbol{P}^\theta \,\mathrm{d}A &= \oint_{\partial\mathcal{D}_{h_r}} r^2 P^{\theta r}\,\mathrm{d}s - \int_{\mathcal{D}_{h_r}} 2rP^{\theta r}\,\mathrm{d}A.
\end{aligned}$$

From this we find that (5.17) implies

$$\int_{\mathcal{D}_{h_r}} m^\theta \,\mathrm{d}A - \oint_{\partial\mathcal{D}_{h_r}} 2rm^\theta \,\mathrm{d}s + \oint_{\partial\mathcal{D}_{h_r}} r^2\frac{\partial m^\theta}{\partial r}\,\mathrm{d}s = \oint_{\partial\mathcal{D}_{h_r}} r^2 P^{\theta r}\,\mathrm{d}s. \tag{5.18}$$

None of the integrals in (5.16) and (5.18) contains derivatives with respect to θ or terms that otherwise have a singularity in the origin, therefore these integrals are all convergent. We will numerically approximate these integrals to obtain a finite difference equations for the grid points around the origin. The integrals on the left hand side of (5.16) can be approximated in the following way to second order accuracy:

$$\begin{aligned}
\int_{\mathcal{D}_{h_r}} m^r \,\mathrm{d}A &\approx \sum_{l=1}^{N_\theta}\left(\frac{m^r_{1,l}+m^r_{1,l+1}+m^r_0}{3}\frac{\pi h_r^2}{N_\theta}\right) = \left(\frac{h_r^2 h_\theta}{6}\right)\sum_{l=1}^{N_\theta}(2m^r_{1,l}+m^r_0) \\
-\oint_{\partial\mathcal{D}_{h_r}} 2rm^r \,\mathrm{d}s &\approx -2h_r\sum_{l=1}^{N_\theta}\left(\frac{m^r_{1,l}+m^r_{1,l+1}}{2}\right)h_\theta h_r = -\left(2h_r^2 h_\theta\right)\sum_{l=1}^{N_\theta} m^r_{1,l} \\
\oint_{\partial\mathcal{D}_{h_r}} r^2\frac{\partial m^r}{\partial r}\,\mathrm{d}s &\approx h_r^3\sum_{l=1}^{N_\theta}\frac{h_\theta}{2}\left[\frac{m^r_{2,l}-m^r_0}{2h_r}+\frac{m^r_{2,l+1}-m^r_0}{2h_r}\right] \\
&= \left(\frac{h_r^2 h_\theta}{2}\right)\left[\sum_{l=1}^{N_\theta}(m^r_{2,l}-m^r_0)\right]
\end{aligned}$$

Similarly we approximate the integrals on the right hand side of (5.16):

$$\begin{aligned}
\oint_{\partial\mathcal{D}_{h_r}} r^2 P^{rr}\,\mathrm{d}s &\approx h_r^2\sum_{l=1}^{N_\theta}\frac{P^{rr}_{1,l}+P^{rr}_{1,l+1}}{2}h_r h_\theta = \left(h_r^3 h_\theta\right)\sum_{l=1}^{N_\theta} P^{rr}_{1,l} \\
-\int_{\mathcal{D}_{h_r}} r\left(P^{rr}+P^{\theta\theta}\right)\mathrm{d}A &\approx -\sum_{l=1}^{N_\theta}\frac{P^{rr}_{1,l}+P^{\theta\theta}_{1,l}+P^{rr}_{1,l+1}+P^{\theta\theta}_{1,l+1}}{3}\left(\frac{\pi h_r^3}{N_\theta}\right) \\
&= -\sum_{l=1}^{N_\theta}\left(P^{rr}_{1,l}+P^{\theta\theta}_{1,l}\right)\left(\frac{h_r^3 h_\theta}{3}\right) = -\left(\frac{h_r^3 h_\theta}{3}\right)\sum_{l=1}^{N_\theta}\left(P^{rr}_{1,l}+P^{\theta\theta}_{1,l}\right).
\end{aligned}$$

From this we find that equation (5.16) can be approximated by the finite difference equation

$$\left(\frac{h_r^2 h_\theta}{6}\right)\sum_{l=1}^{N_\theta}(-10m^r_{1,l}+m^r_0)+\left(\frac{h_r^2 h_\theta}{2}\right)\left[\sum_{l=1}^{N_\theta}(m^r_{2,l}-m^r_0)\right] = \left(\frac{h_r^3 h_\theta}{6}\right)\left[\sum_{l=1}^{N_\theta}4P^{rr}_{1,l}-2P^{\theta\theta}_{1,l}\right],$$

which can be rewritten as

$$m^r_0 = \frac{1}{2N_\theta}\left[\sum_{l=1}^{N_\theta}(3m^r_{2,l}-10m^r_{1,l}-h_r(4P^{rr}_{1,l}-2P^{\theta\theta}_{1,l}))\right]. \tag{5.19}$$

Similarly to the approximation of the integrals in (5.16), the integrals in (5.18) can be approximated in the following way:

$$\begin{aligned}
\int_{\mathcal{D}_{h_r}} m^\theta \, \mathrm{d}A &\approx \left(\frac{h_r^2 h_\theta}{6}\right) \sum_{l=1}^{N_\theta} (2m_{1,l}^\theta + m_0^\theta) \\
-\int_{\partial\mathcal{D}_{h_r}} 2rm^\theta \, \mathrm{d}s &\approx -\left(2h_r^2 h_\theta\right) \sum_{l=1}^{N_\theta} m_{1,l}^\theta \\
\int_{\partial\mathcal{D}_{h_r}} r^2 \frac{\partial m^\theta}{\partial r} \, \mathrm{d}s &\approx \left(\frac{h_r^2 h_\theta}{2}\right) \left[\sum_{l=1}^{N_\theta} (m_{2,l}^\theta - m_0^\theta)\right] \\
\int_{\partial\mathcal{D}_{h_r}} r^2 P^{\theta r} \, \mathrm{d}s &\approx \left(h_r^3 h_\theta\right) \sum_{l=1}^{N_\theta} P_{1,l}^{\theta r}.
\end{aligned}$$

From this we see that (5.18) can be approximated by the finite difference equation

$$\left(\frac{h_r^2 h_\theta}{6}\right) \sum_{l=1}^{N_\theta} (-10m_{1,l}^\theta + m_0^\theta) + \left(\frac{h_r^2 h_\theta}{2}\right) \left[\sum_{l=1}^{N_\theta} (m_{2,l}^\theta - m_0^\theta)\right] = \left(h_r^3 h_\theta\right) \sum_{l=1}^{N_\theta} P_{1,l}^{\theta r},$$

which we rewrite to

$$m_0^\theta = \frac{1}{2N_\theta} \left[\sum_{l=1}^{N_\theta} (3m_{2,l}^\theta - 10m_{1,l}^\theta - 6h_r P_{1,l}^{\theta r})\right]. \tag{5.20}$$

For the grid points (r_1, θ_l), with $1 \le l \le N_\theta$ we can also write down equation (5.7) and (5.11), i.e.

$$\begin{aligned}
&\frac{1}{r_1 h_r^2} \left[r_{\frac{3}{2}}(m_{2,l}^r - m_{1,l}^r) - r_{\frac{1}{2}}(m_{1,l}^r - m_0^r)\right] + \frac{1}{r_1^2 h_\theta^2} \left[m_{1,l+1}^r - 2m_{1,l}^r + m_{1,l-1}^r\right] \\
&- \frac{1}{r_1^2 h_\theta} \left[m_{1,l+1}^\theta - m_{1,l-1}^\theta\right] - \frac{m_{1,l}^r}{r_j^2} \\
&= \frac{1}{2h_r} \left[-P_{3,l}^{rr} - 4P_{2,l}^{rr} + P_{1,l}^{rr}\right] + \frac{1}{2r_1 h_\theta} \left[P_{1,l+1}^{r\theta} - P_{1,l-1}^{r\theta}\right] + \frac{1}{r_1} \left[P_{1,l}^{rr} - P_{1,l}^{\theta\theta}\right],
\end{aligned} \tag{5.21}$$

and

$$\begin{aligned}
&\frac{1}{r_1 h_r^2} \left[r_{\frac{3}{2}}(m_{2,l}^\theta - m_{1,l}^\theta) - r_{\frac{1}{2}}(m_{1,l}^\theta - m_0^\theta)\right] + \frac{1}{r_1^2 h_\theta^2} \left[m_{1,l+1}^\theta - 2m_{1,l}^\theta + m_{1,l-1}^\theta\right] \\
&+ \frac{1}{r_1^2 h_\theta} \left[m_{1,l+1}^r - m_{1,l-1}^r\right] - \frac{m_{1,l}^\theta}{r_1^2} \\
&= \frac{1}{2h_r} \left[-P_{3,l}^{\theta r} - 4P_{2,l}^{\theta r} + P_{1,l}^{\theta r}\right] + \frac{1}{2r_1 h_\theta} \left[P_{1,l+1}^{\theta\theta} - P_{1,l-1}^{\theta\theta}\right] + \frac{1}{r_1} \left[P_{1,l}^{r\theta} + P_{1,l}^{\theta r}\right],
\end{aligned} \tag{5.22}$$

where we only replaced the central difference approximations of the radial derivatives of P^{rr} and $P^{\theta r}$ by one-sided difference of the same accuracy to avoid P_0^{rr} and $P_0^{\theta r}$.

If we substitute (5.19) into (5.21) and multiply by h_r we find the equations

$$\begin{aligned}
&\frac{1}{2h_r} \left[3m_{2,l}^r - 4m_{1,l}^r\right] + \frac{1}{h_r h_\theta^2} \left[m_{1,l+1}^r - 2m_{1,l}^r + m_{1,l-1}^r\right] - \frac{1}{h_r h_\theta} \left[m_{1,l+1}^\theta - m_{1,l-1}^\theta\right] \\
&- \frac{m_{1,l}^r}{h_r} + \sum_{l=1}^{N_\theta} \frac{3m_{2,l}^r - 10m_{1,l}^r}{4N_\theta h_r} \\
&= \frac{1}{2} \left[-P_{3,l}^{rr} - 4P_{2,l}^{rr} + P_{1,l}^{rr}\right] + \frac{1}{2h_\theta} \left[P_{1,l+1}^{r\theta} - P_{1,l-1}^{r\theta}\right] + \left[P_{1,l}^{rr} - P_{1,l}^{\theta\theta}\right] + \sum_{l=1}^{N_\theta} \frac{2P_{1,l}^{rr} - P_{1,l}^{\theta\theta}}{2N_\theta}
\end{aligned} \tag{5.23}$$

for the grid points (r_1, θ_l) with $1 \leq l \leq N_\theta$. Similarly, substituting (5.20) into (5.22) and multiplying by h_r, we find obtain

$$\begin{aligned}
&\frac{1}{2h_r}\left[3m^\theta_{2,l} - 4m^\theta_{1.l}\right] + \frac{1}{h_r h^2_\theta}\left[m^\theta_{1,l+1} - 2m^\theta_{1,l} + m^\theta_{1,l-1}\right] - \frac{1}{h_r h_\theta}\left[m^r_{1,l+1} - m^r_{1,l-1}\right] \\
&- \frac{m^\theta_{1,l}}{h_r} + \sum_{l=1}^{N_\theta} \frac{3m^\theta_{2,l} - 10m^\theta_{1,l}}{4N_\theta h_r} \\
&= \frac{1}{2}\left[-P^{\theta r}_{3,l} - 4P^{\theta r}_{2,l} + P^{\theta r}_{1,l}\right] + \frac{1}{2h_\theta}\left[P^{r\theta}_{1,l+1} - P^{r\theta}_{1,l-1}\right] + \left[P^{rr}_{1,l} - P^{\theta\theta}_{1,l}\right] + \sum_{l=1}^{N_\theta} \frac{3P^{\theta r}_{1,l}}{2N_\theta}
\end{aligned} \tag{5.24}$$

for the grid points (r_1, θ_l) with $1 \leq l \leq N_\theta$.

Now we have a finite difference equation for all the grid points in our grid. For the points on the boundary $\partial\mathcal{E}$ we have equations (5.10) and (5.12), for the grid points directly adjacent to the origin we have equations (5.23) and (5.24), and for the remaining grid points we have the equations (5.7) and (5.11). This completes the description of the Least-squares method in polar coordinates. We will in the next section compare the Least-square method in Cartesian coordinates with the Least-square method in polar coordinates.

5.3 Comparison between the Cartesian- and polar-implementation

We will compare the Least-squares method in Cartesian coordinates with the Least-squares method in polar coordinates for two different test cases. In all two test cases we will take as the light source $\mathcal{E} = \mathcal{D}_1$. We will take a light source with a uniform intensity, hence $E : \mathcal{E} \to \mathbb{R}_{>0}$ will be defined by $E(x) = 1/\pi$ for all $x \in \mathcal{E}$. We have normalized this function such that

$$\int_{\mathcal{E}} E(x^1, x^2)\sqrt{e}\mathrm{d}x^1\mathrm{d}x^2 = 1.$$

The first target will be a uniform square. We set $\mathcal{F}_1 = [-1, 1]^2 \subset \mathbb{R}^2$ and define $F_1 : \mathcal{F}_1 \to \mathbb{R}_{>0}$ by $F(y) = 1/4$ for all $y \in \mathcal{F}$. The function F is also normalized, therefore we have

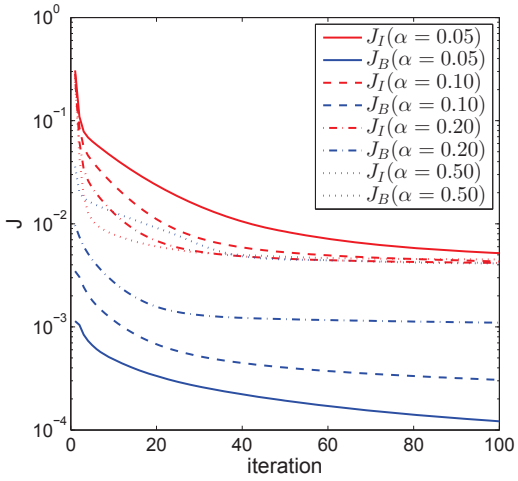
$$\int_{\mathcal{E}} E(x^1, x^2)\sqrt{e}\mathrm{d}x^1\mathrm{d}x^2 = \int_{\mathcal{F}_1} F_1(y^1, y^2)\sqrt{f}\mathrm{d}y^1\mathrm{d}y^2.$$

The second target will be the pair $(\mathcal{F}_2, F_2)$ corresponding to the intensity pattern on a plane in the far-field of the reflector corresponding to Figure 5.4, described by a pair $(\mathcal{H}_2 \subset \mathbb{R}^2, H_2)$. In Section 3.6, we discussed how to find the pair $(\mathcal{F}_2, F_2)$ from $(\mathcal{H}_2, H_2)$, we will take the distance d from reflector to plane equal to 1. The function F_2 will again be normalized.

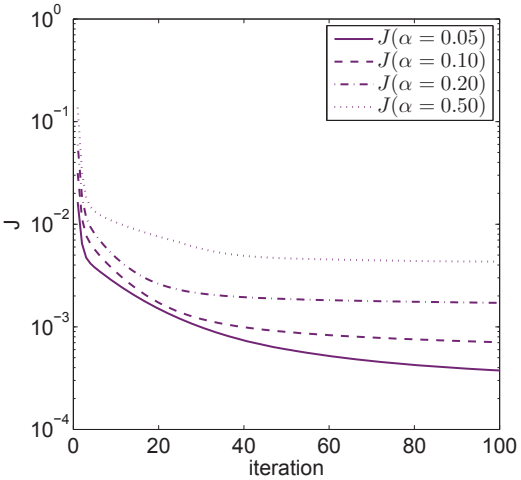
Before we can start to compare we first have to fix the value of $\alpha \in (0, 1)$ in (4.6). This actually a weak point of the Least-squares method, because it is unclear what value for α to choose. The Least-squares method will try to bring J in (4.6) to zero. Thus, if α is very small the boundary error J_B will be much smaller than J_I, similarly if α is very close to 1 then J_I will be much smaller than J_B. In general a choice for α should be based on the application in mind. On the one hand, if it is really important that mapping adheres strongly to the boundary $\partial\mathcal{F}$ then a relative low value of α should be chosen. On the other hand, if it is more important that the solution very closely satisfies the Monge-Ampère equation on the interior of $\mathcal{E}$ then a relative high value of α should be chosen.

In Figure 5.5 and Figure 5.6 the convergence for the test case $(\mathcal{F}_1, F_1)$ is shown for the the Least-squares method in Cartesian coordinates and in polar coordinates, respectively.

We will analyse test case $(\mathcal{F}_1, F_1)$ for $\alpha = 0.2$, because for this value of α the convergence is most fast for the Cartesian case as can be seen in Figure 5.5b. In Figure 5.7 the convergence of the

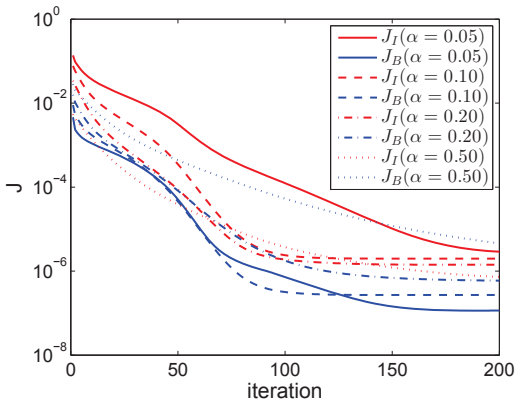


(a) Boundary error J_B and interior error J_I.

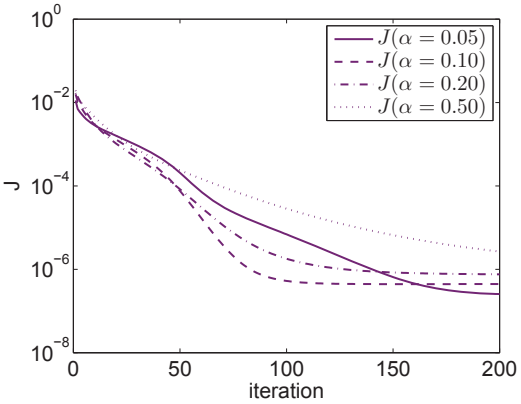


(b) Total error J.

Figure 5.5: Convergence for different values of α for test case $(\mathcal{F}_1, F_1)$ in Cartesian coordinates. We used a grid with $N_x = 200$, $N_y = 200$. We used 1000 points the approximate $\partial\mathcal{F}$, hence this will form no restriction.



(a) Boundary error J_B and interior error J_I.



(b) Total error J.

Figure 5.6: Convergence for different values of α for test case $(\mathcal{F}_1, F_1)$ in polar coordinates. We used a grid with $N_r = 200$, $N_\theta = 200$. We again used 1000 points to approximate $\partial\mathcal{F}$.

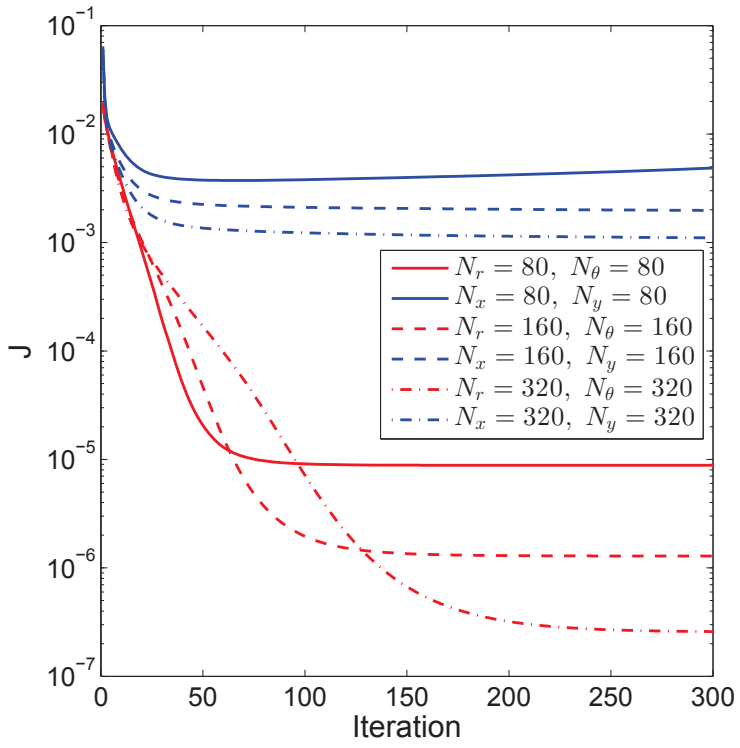


Figure 5.7: The converge of the Least-squares method in Cartesian coordinates and in polar coordinates for different grid sizes. The number of points to approximate $\partial\mathcal{F}$ is again taken equal to 1000.

Least-squares method is shown for different grid sizes in Cartesian and polar coordinates. It can be seen that convergence for the Least-squares method is significantly better in polar coordinates than in Cartesian coordinates. The convergence in polar coordinates does tend to take more iterations. In Figure 5.8 the functionals J_I and J_B are plotted as a function of the number of iterations for a 300×300 grid. In Figure 5.9 the mapping $\boldsymbol{m}$ is shown. In the figure it can be seen how the grid on $\mathcal{E}$ gets mapped to $\mathcal{F}$. Note that the grid points on $\mathcal{E}$ are not evenly distributed in the polar coordinate case. Closer to the origin the number of grid points increases, while far from the origin the density of grid points is lower than in the Cartesian case. In Figure 5.9a it can be seen that the mapping does not satisfy the boundary condition $\boldsymbol{m}(\partial\mathcal{E}) = \partial\mathcal{F}$ very well. On all four edges of the square small bulges appear. These bulges will be catastrophic when we extend the mapping to a larger domain in the next chapter. These bulges are not there in the polar case as can be seen in Figure 5.9b.

The results for test case $(\mathcal{F}_2, F_2)$ are comparable to those of the first test case. In Figure 5.10 we see that in the Cartesian case the same bulges appear as for the first test case. In the polar case these bulges are again not there. In Figure 5.11 the convergence of Least-squares method is shown for different grid sizes for the second test case. The convergence is again significantly better for the implementation in polar coordinates.

So, we see an overall better performance of the Least-squares method in polar coordinates when compared to the Least-squares method in Cartesian coordinates for a disk-shaped light source $\mathcal{E} = \mathcal{D}_1$. This is of importance in the construction of a physical reflector as we will see in the next chapter.

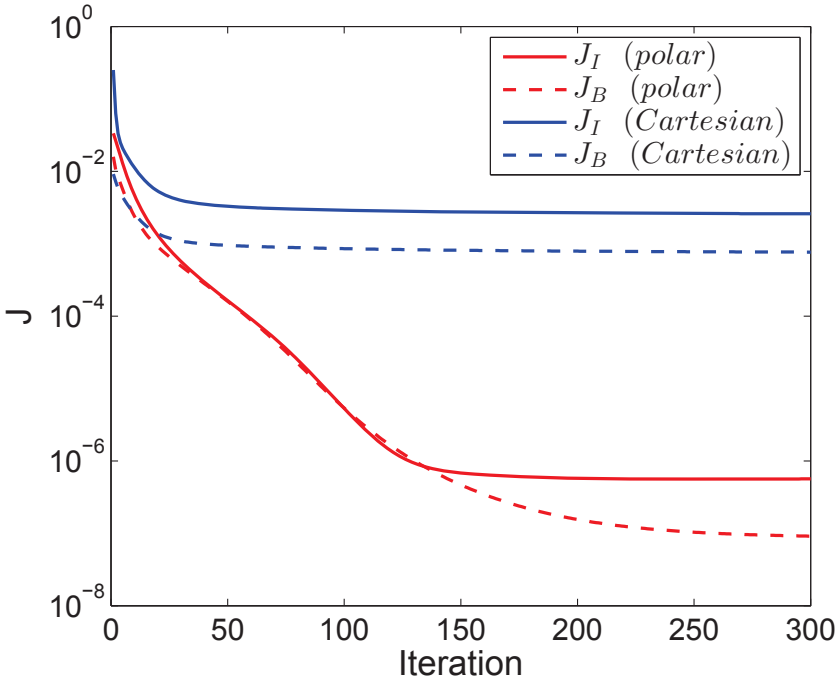
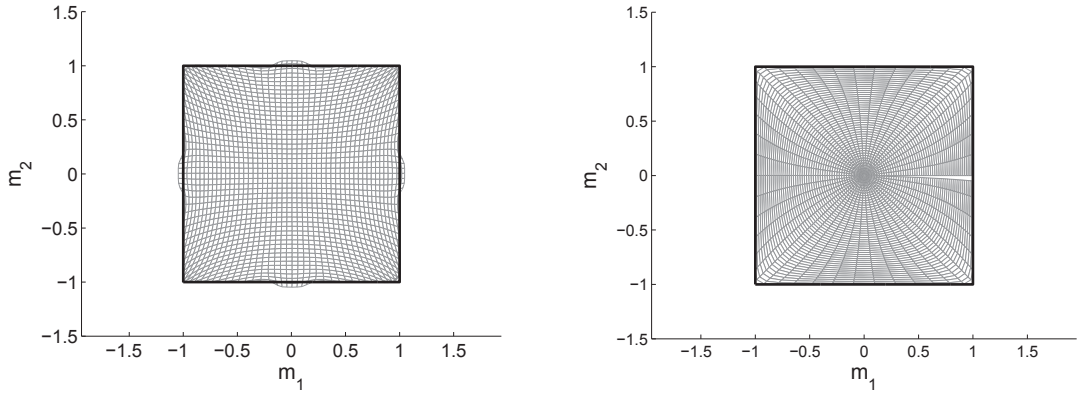


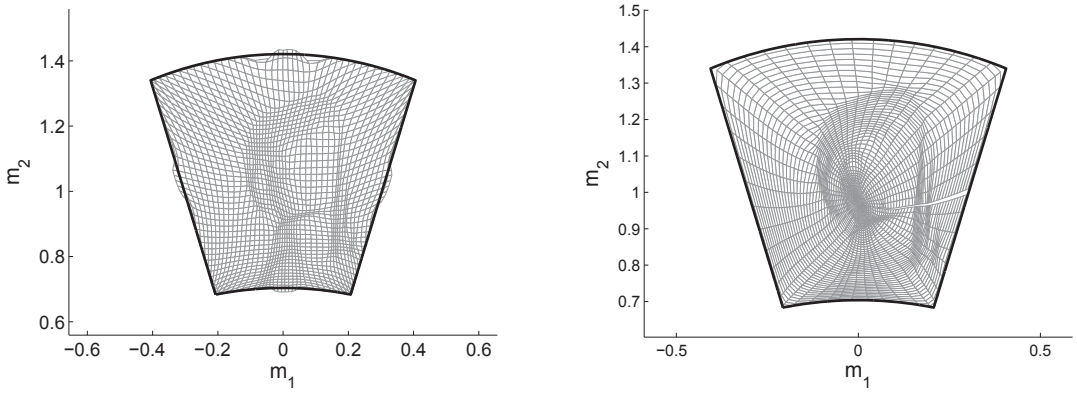
Figure 5.8: The converge of the Least-squares method in Cartesian coordinates for $N_x = 300$ and $N_y = 300$, and in polar coordinates for $N_r = 300$ and $N_\theta = 300$. The number of points to approximate $\partial\mathcal{F}$ is again taken equal to 1000 and $\alpha = 0.2$.



(a) The mapping for the Least-squares method in Cartesian coordinates.

(b) The mapping for the Least-squares method in polar coordinates.

Figure 5.9: The mapping $\boldsymbol{m}$ is shown after 300 iterations on a 200×200 grid. The number of points to approximate $\partial\mathcal{F}$ is again taken equal to 1000 and $\alpha = 0.2$.



(a) The mapping for the Least-squares method in Cartesian coordinates.

(b) The mapping for the Least-squares method in polar coordinates.

Figure 5.10: The mapping $\boldsymbol{m}$ is shown after 300 iterations on a 200×200 grid. The number of points to approximate $\partial\mathcal{F}$ is again taken equal to 1000 and $\alpha = 0.2$.

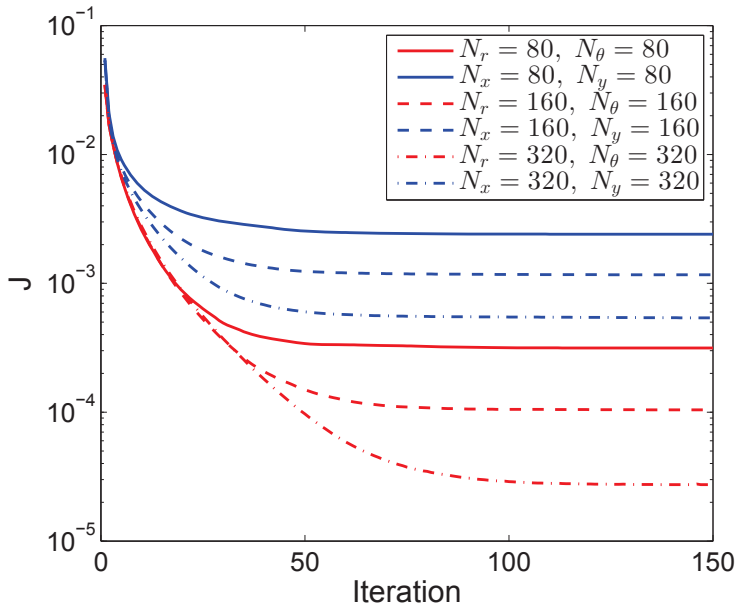


Figure 5.11: The converge of the least-squares method in Cartesian coordinates and in polar coordinates for different grid sizes. The number of points to approximate $\partial\mathcal{F}$ is again taken equal to 1000.

Chapter 6

Extension of the Reflector Surface

In Chapter 1 we discussed the steps still to take to produce a physical reflector. We considered the milling machine that we want to use to produce the reflector and found that this machine needs to be provided with the heights of the reflector on a polar coordinate grid. Moreover, we remarked that the deceleration or acceleration of the chisel is bounded and that this implies that $|\partial^2 v/\partial\theta^2|$ cannot be too large. In order to significantly reduce $|\partial^2 v/\partial\theta^2|$ we chose not to use the coordinate system aligned with the light source but a rotated cylindrical coordinate system in which the chisel of the milling machine is approximately normal to the reflector surface. This coordinate system is depicted in Figure 1.3.

Furthermore, the milling machine only produces disk-shaped reflectors and therefore needs to be provided with data for a disk-shaped reflector. However, while the function describing the reflector surface in the coordinate system aligned with the light source, i.e. the red coordinate system in Figure 1.3, had a disk-shaped support, the support of the function describing the reflector surface in the rotated coordinate system has approximately the shape of an ellipse. We therefore need to extrapolate the reflector surface to a disk containing this ellipse in order to be able to provide the milling machine with workable data. Attempts to extrapolate the reflector in the rotated cylindrical coordinate system failed and therefore we try to extrapolate the reflector surface already in the coordinates of the light source. We will extrapolate by using an adapted form of the least-squares method in order to be able to specify in which direction the extrapolated parts of the reflector reflect incident light.

6.1 Boundary value problem for reflector extension

We will determine the extrapolated reflector in two steps. First we will use the least-squares method to determine the original reflector, and second we will use a slightly adapted version of the least-squares method to determine the extrapolated part of the reflector. The two-part boundary value problem that we will solve is the following.

Problem 6.1.1. Find $\boldsymbol{m}_I \in T(\mathcal{E}_I)_{C^1}$ that satisfies

$$\frac{\det(\nabla\hat{\boldsymbol{m}}_{\text{Int}}(x))}{e} = \frac{E_{\text{Int}}(x)}{F_{\text{Int}}(\boldsymbol{m}'(x))}, \qquad \text{in } \mathcal{E}_{\text{Int}}, \tag{6.1a}$$

$$\boldsymbol{m}_{\text{Int}}(\partial\mathcal{E}_{\text{Int}}) = \partial\mathcal{F}_{\text{Int}}, \tag{6.1b}$$

and find $\boldsymbol{m} \in T\mathcal{E}_C$ that satisfies

$$\frac{\det(\nabla\hat{\boldsymbol{m}}(x))}{e} = \frac{E_{\text{Ext}}(x)}{F_{\text{Ext}}(\boldsymbol{m}(x))}, \qquad \text{in } \mathcal{E}_{\text{Ext}}, \tag{6.2a}$$

$$\boldsymbol{m} = \boldsymbol{m}_{\text{Int}}, \qquad \text{in } \overline{\mathcal{E}}_{\text{Int}}, \tag{6.2b}$$

$$\boldsymbol{m}(\partial\mathcal{E}_{\text{Ext}}) = \partial\mathcal{F}_{\text{Ext}}, \tag{6.2c}$$

and for which $\nabla \hat{\boldsymbol{m}}$ and hence $\nabla \hat{\boldsymbol{m}}_{\text{Int}}$ are positive semi-definite tensors. In this problem the functions E_{Int}, E_{Ext}, F_{Int} and F_{Ext} are strictly positive functions such that

$$\int_{\mathcal{E}_{\text{Int}}} E_{\text{Int}}(x^1, x^2)\sqrt{e}\ \mathrm{d}x^1\mathrm{d}x^2 = \int_{\mathcal{F}_{\text{Int}}} F_{\text{Int}}(y^1, y^2)\sqrt{f}\ \mathrm{d}y^1\mathrm{d}y^2, \tag{6.3a}$$

$$\int_{\mathcal{E}_{\text{Ext}}} E_{\text{Ext}}(x^1, x^2)\sqrt{e}\ \mathrm{d}x^1\mathrm{d}x^2 = \int_{\mathcal{F}_{\text{Ext}}} F_{\text{Ext}}(y^1, y^2)\sqrt{f}\ \mathrm{d}y^1\mathrm{d}y^2, \tag{6.3b}$$

where x^1, x^2 are coordinates on $\mathcal{E} := \mathcal{E}_{\text{Int}} \cup \mathcal{E}_{\text{Ext}}$ with corresponding metric e_{ij} and y^1, y^2 are coordinates on $\mathcal{F} := \mathcal{F}_{\text{Int}} \cup \mathcal{F}_{\text{Ext}}$ with corresponding metric f_{ij}. Furthermore, the sets $\mathcal{E}_{\text{Int}}, \mathcal{E}_{\text{Ext}} \subset \mathbb{R}^2$ are such that $\mathcal{E}_{\text{Int}}$ and $\mathcal{E}$ are convex, closed and bounded. We assume the sets $\mathcal{F}_{\text{Int}}, \mathcal{F}_{\text{Ext}} \subset \mathbb{R}^2$ also to be such that $\mathcal{F}_{\text{Int}}$ and $\mathcal{F}_{\text{Ext}}$ are closed and convex.

In Figure 6.1 a graphical representation is given of the domains and mappings involved in Problem 6.1.1. Let us define $E : \mathcal{E} \to \mathbb{R}_{>0}$ by

$$E\big|_{\mathcal{E}_{\text{Int}}} := E_{\text{Int}} \quad \text{and} \quad E\big|_{\mathcal{E}_{\text{Ext}}} := E_{\text{Ext}},$$

and let us define $F : \mathcal{F} \to \mathbb{R}_{>0}$ analogously. The mapping $\boldsymbol{m}$ is a map from $\mathcal{E}$ to $\mathcal{F}$ and satisfies $\boldsymbol{m}(\partial\mathcal{E}) = \partial\mathcal{F}$, however, note that in general $\boldsymbol{m}$ is not a solution to the boundary value problem

$$\begin{cases} \dfrac{\det(\nabla \hat{\boldsymbol{m}}(x))}{e} = \dfrac{E(x)}{F(\boldsymbol{m}(x))}, & \text{in } \mathcal{E}, \\ \boldsymbol{m}(\partial\mathcal{E}) = \partial\mathcal{F}. \end{cases}$$

This results from the fact that in general $\boldsymbol{m}$ does not have to satisfy $\boldsymbol{m}(\mathcal{E}_{\text{Int}}) = \mathcal{F}_{\text{Int}}$ and $\boldsymbol{m}(\partial\mathcal{E}_{\text{Int}}) = \partial\mathcal{F}_{\text{Int}}$. Furthermore, it must be remarked that the mapping $\boldsymbol{m} \notin T\mathcal{E}_{C^1}$, but $\boldsymbol{m} \in T\mathcal{E}_C$. We have imposed that $\boldsymbol{m} = \boldsymbol{m}_{\text{Int}}$ on $\mathcal{E}_{\text{Int}}$ and hence on $\partial\mathcal{E}_{\text{Int}}$ however we have placed no demands on the derivatives of $\boldsymbol{m}$ on $\partial\mathcal{E}_{\text{Int}}$, therefore we can only expect $\boldsymbol{m}$ to be continuous on $\partial\mathcal{E}_{\text{Int}}$ and not necessary differentiable. This is indeed what we will see in the numerical tests.

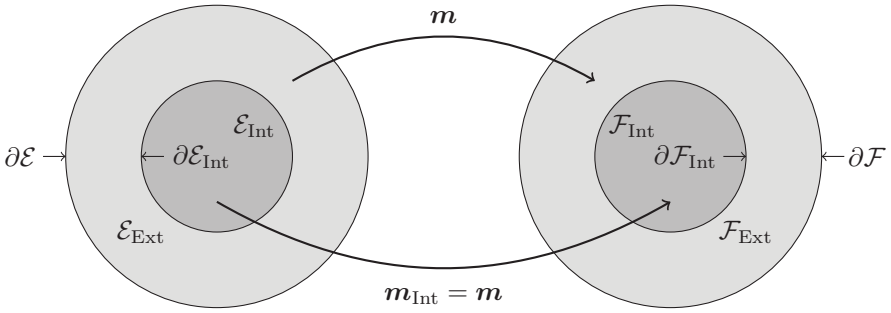


Figure 6.1: Schematic representation of the two mappings and the domains involved. In the test cases that we will consider, the sets and $\mathcal{E}_{\text{Int}}$ and $\mathcal{E}_{\text{Ext}}$ will have the circular shape presented here, but the sets $\mathcal{F}_{\text{Int}}$ and $\mathcal{F}_{\text{Ext}}$ will have different noncircular shapes. However, they will have the relation to each other as represented here in a topological sense.

To solve the boundary value problem (6.2) we need to make some minor adaptations to the least-squares method. The first two steps of each iteration of the method, i.e. the minimization of J_B and J_I as described in Sections 4.2 and 4.3, will stay unaltered as they are performed pointwise. The third step of an iteration, i.e. the minimization of J as described in Section 4.4 does need some changes. This time we have to take the boundary condition (6.2b) into account. We therefore have to minimize J over the space

$$\mathcal{V}_{\boldsymbol{m}_{\text{Int}}} := \left\{ \boldsymbol{v} \in T\mathcal{E}_{C^2} \mid \boldsymbol{v}(x) = \boldsymbol{m}_{\text{Int}}(x) \quad x \in \mathcal{E}_{\text{Int}} \right\}, \tag{6.4}$$

instead of over the space in equation (4.7). We can follow the same derivation as in Section 4.4, which gives us the equations (4.34) again, i.e.

$$\begin{aligned} D_j D^j m^i &= D_j P^{ij} && \text{in } \mathcal{E}, \\ \alpha(D^j m^i) n_j + (1-\alpha) m^i &= \alpha P^{ij} n_j + (1-\alpha) b^i && \text{on } \partial\mathcal{E}. \end{aligned}$$

However, here $\boldsymbol{m} \in \mathcal{V}_{\boldsymbol{m}_{\text{Int}}}$ and $\boldsymbol{m}_{\text{Int}}$, being a solution to boundary value problem (6.1), already satisfies the first of these two equations. This implies that the minimizer is given by $\boldsymbol{m} \in \mathcal{V}_{\boldsymbol{m}_{\text{Int}}}$ that satisfies

$$\begin{aligned} D_j D^j m^i &= D_j P^{ij} && \text{in } \mathcal{E}_{\text{Ext}}, \\ m^i &= (m')^i && \text{in } \mathcal{E}_{\text{Int}}, \\ \alpha(D^j m^i) n_j + (1-\alpha) m^i &= \alpha P^{ij} n_j + (1-\alpha) b^i && \text{on } \partial\mathcal{E}. \end{aligned}$$

We demand the mapping $\boldsymbol{m}$ to be at least continuous therefore the condition $m^i = (m_1)^i$ on $\mathcal{E}_{\text{Int}}$ effectively works as a boundary condition on $\partial\mathcal{E}_{\text{Int}}$. Thus the minimizer of J over $\mathcal{V}_{\boldsymbol{m}_{\text{Int}}}$ is on $\mathcal{E}_{\text{Int}}$ given by $\boldsymbol{m}_{\text{Int}}$ and on $\mathcal{E}_{\text{Ext}}$ by the solution to the boundary value problem

$$\begin{aligned} D_j D^j m^i &= D_j P^{ij} && \text{in } \mathcal{E}_{\text{Ext}}, && \text{(6.5a)} \\ m^i &= (m')^i && \text{on } \partial\mathcal{E}_{\text{Int}}, && \text{(6.5b)} \\ \alpha(D^j m^i) n_j + (1-\alpha) m^i &= \alpha P^{ij} n_j + (1-\alpha) b^i && \text{on } \partial\mathcal{E}. && \text{(6.5c)} \end{aligned}$$

Thus, in order to solve Problem 6.1.1, we use the least-squares method with the only adaptation that we, instead of boundary value problem (4.34), now solve boundary value problem (6.5a) to determine the minimizer for J in the third substep of each iteration. We will from now on call this version of the least-squares method the adapted least-squares method.

Problem 6.1.1 is Problem 4.1.2 with the extra constraint that $\boldsymbol{m}$ must equal $\boldsymbol{m}_{\text{Int}}$ on $\mathcal{E}_{\text{Int}}$. We discussed earlier that for problem 4.1.2 there possibly consists a unique solution under certain smoothness criteria of the boundary and the functions E and F. We can therefore expect the functional J to converge to zero when we apply the least-squares method to Problem 4.1.2. However, when we apply the adapted least-squares method as just described to Problem 6.1.1, the functional J can only converge to zero if the minimizer of Problem 4.1.2 is also the minimizer of Problem 6.1.1. So, suppose if $\boldsymbol{m}$ is the solution of Problem 4.1.2, for which $J = 0$, then we need to have

$$\boldsymbol{m}\big|_{\mathcal{E}_{\text{Int}}} = \boldsymbol{m}_{\text{Int}},$$

in order for J to be able to diminish to zero when we apply the adapted least-squares method to Problem 6.1.1. Thus in general we can not expect the functional J to diminish to zero when we apply the Adapted least-squares method to the Problem 6.1.1. This is a serious weak point of the method. Although $\boldsymbol{m}_{\text{Int}}$ still will be a fairly accurate solution of the Monge-Ampère equation mapping from $\mathcal{E}_{\text{Int}}$ to $\mathcal{F}_{\text{Int}}$, we cannot expect for $\boldsymbol{m}$ restricted to $\mathcal{E}_{\text{Ext}}$ to be an accurate solution of the Monge-Ampère equation, mapping from $\mathcal{E}_{\text{Ext}}$ to $\mathcal{F}_{\text{Ext}}$. However, just as in the case of least-squares method we can test the adapted least-squares method by calculating the reflector corresponding to the mapping and submitting this reflector to a virtual light bundle leaving our source and determining resulting light intensity distribution on the projection screen. This we will do in the next section.

6.2 Numerical results of the adjusted least-squares method

We will apply the Adjusted least-squares method to two test cases. For each test case we will again take $\mathcal{E}_{\text{Int}} = \mathcal{D}_{R_{\text{Int}}=1}$ and $\mathcal{E}_{\text{Ext}} = \mathcal{D}_{R_{\text{Ext}}=1.5}$. The first test case will be the simplest. We take $\mathcal{F}^1_{\text{Int}}$

to be the figure which has the boundary $\partial\mathcal{F}^1_{\text{Int}}$ described in polar coordinates $\rho \in \mathbb{R}_{>0}, \phi \in [0, 2\pi)$ by

$$\rho_{\text{Int}}(\phi) = 1 + 0.1\cos(3\phi).$$

(This test case is taken from [5].) We will take $F^1_{\text{Int}} : \mathcal{F}^1_{\text{Int}} \to \mathbb{R}_{>0}$ constant such that the integral of F^1_{Int} over $\mathcal{F}^1_{\text{Int}}$ equals 1. We define the extension $\mathcal{F}^1_{\text{Ext}}$ again by its boundary. Let the boundary $\partial\mathcal{F}^1_{\text{Ext}}$ be given by the curve

$$\rho_{\text{Ext}}(\phi) = \frac{R_{\text{Ext}}}{R_{\text{Int}}} + 0.1\cos(3\phi).$$

We take F^1_{Ext} to assume the same constant value on the whole of $\mathcal{F}^1_{\text{Ext}}$ as F^1_{Int} assumes on $\mathcal{F}^1_{\text{Int}}$. It is easily verified that $\mathcal{F}^1_{\text{Int}}, F^1_{\text{Int}}$ and $\mathcal{F}^1_{\text{Ext}}, F^1_{\text{Ext}}$ such defined satisfy equations (6.3). The sets $\mathcal{F}^1_{\text{Int}}$ and $\mathcal{F}^1_{\text{Ext}}$ and their boundaries are shown in Figure 6.2.

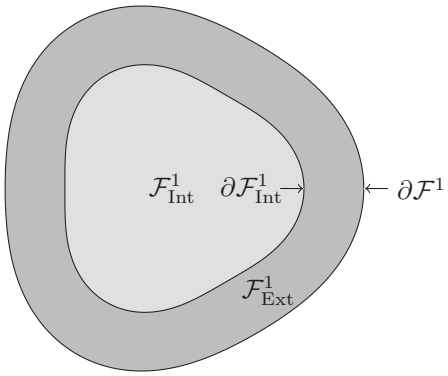


Figure 6.2: The target space for test case 1.

As the second test case we again take the set $\mathcal{F}^2_{\text{Int}}$ to be the set corresponding to the projection of the painting by Vermeer, shown in Figure 5.4, on a projection plane at about 2 meters away from the reflector. We assume that the radii R_{Int} and R_{Ext} of $\mathcal{E}_{\text{Int}}$ and $\mathcal{E}_{\text{Ext}}$ are 1 and 1.5 centimeters respectively. We take the set $\mathcal{F}^2_{\text{Int}}$ such that the picture on the projection screen is roughly 20 centimeters wide and 40 centimeters high. We take the extension $\partial\mathcal{F}^2$ to be equidistant to $\partial\mathcal{F}^2_{\text{Int}}$ and we let F^2_{Ext} decrease linearly in the normal direction from the value of F^2_{Int} on $\partial\mathcal{F}^2_{\text{Int}}$ to 20% of the maximum value of F^2_{Int}. We do this in order to ensure that $F^2_{\text{Ext}} > 0$ and we do not divide by 0 in equation (6.2a).

We will use the adapted least-squares method to find a minimizer for the functional J. Let us denote such a minimizer of J by $\boldsymbol{m}$. This minimizer is by definition of the adapted least-squares method an element of the set defined in equation (6.4), hence $\boldsymbol{m}$ equals $\boldsymbol{m}_{\text{Int}}$ on $\mathcal{E}_{\text{Int}}$. Let us now suppose that we have an ideal situation where $\boldsymbol{m}_{\text{Int}}$ is the exact solution to boundary value problem (6.1). The functionals J_I and J_B are then given by

$$J_I(\boldsymbol{m}, \boldsymbol{P}) = \frac{1}{2}\iint_{\mathcal{E}_{\text{Int}}} \|\nabla\hat{\boldsymbol{m}} - \boldsymbol{P}\|^2\sqrt{e}\ \mathrm{d}x^1\mathrm{d}x^2,$$
$$J_B(\boldsymbol{m}, \boldsymbol{b}) = \frac{1}{2}\oint_{\partial\mathcal{E}} \|\boldsymbol{m} - \boldsymbol{b}\|^2\ \mathrm{d}s,$$

because the integrand of the integral J_I is 0 on $\mathcal{E}_{\text{Int}}$ when $\boldsymbol{m}_{\text{Int}}$ is the exact solution to (6.1). We argued earlier that in general there will not exist a solution to Problem 6.1.1, therefore $J = \alpha J_I + (1-\alpha)J_B > 0$. Moreover, $\boldsymbol{m}$ is only continuous and not differentiable for $r = R_{\text{Int}}$. The partial derivatives

$$\frac{\partial m^r}{\partial r} \quad \text{and} \quad \frac{\partial m^\theta}{\partial r} \tag{6.6}$$

will therefore have a jump at $r = R_{\text{Int}}$ and these will cause jumps in the second derivatives of the reflector surface. Large values and jumps in second derivatives of the reflector surface we want to avoid, because these cause trouble for our milling machine. In order to measure the non-differentiability of $\boldsymbol{m}$ with respect to r at $r = R_{\text{Int}}$, let us define the functions

$$f^r_{\text{n-d}}(\theta) := \left| \lim_{\rho \downarrow R_{\text{Int}}} \left[\frac{\partial m^r}{\partial r}(\rho, \theta) \right] - \lim_{\rho \uparrow R_{\text{Int}}} \left[\frac{\partial m^r}{\partial r}(\rho, \theta) \right] \right|, \tag{6.7a}$$

$$f^\theta_{\text{n-d}}(\theta) := \left| \lim_{\rho \downarrow R_{\text{Int}}} \left[\frac{\partial m^r}{\partial \theta}(\rho, \theta) \right] - \lim_{\rho \uparrow R_{\text{Int}}} \left[\frac{\partial m^\theta}{\partial r}(\rho, \theta) \right] \right|, \tag{6.7b}$$

where $\lim_{\rho \downarrow R_{\text{Int}}}$ and $\lim_{\rho \uparrow R_{\text{Int}}}$ denote the one-sided limits with $r > R_{\text{Int}}$ and $r < R_{\text{Int}}$, respectively. The functions $f^r_{\text{n-d}}$ and $f^\theta_{\text{n-d}}$ give the size of the jumps in the derivatives as a function of θ. We can use these functions to give a measure to the non-differentiability at $r = R_{\text{Int}}$. Let us define the following errors:

$$\varepsilon^r_2 := \sqrt{\int_0^{2\pi} \left(f^r_{\text{n-d}}(\theta) \right)^2 R_{\text{Int}} \, \mathrm{d}\theta}, \qquad \varepsilon^\theta_2 := \sqrt{\int_0^{2\pi} \left(f^\theta_{\text{n-d}}(\theta) \right)^2 R_{\text{Int}} \, \mathrm{d}\theta}, \tag{6.8a}$$

$$\varepsilon^r_\infty := \sup_{[0,2\pi]} \left| f^r_{\text{n-d}}(\theta))^2 \, R_{\text{Int}} \right|, \qquad \varepsilon^\theta_\infty := \sup_{[0,2\pi]} \left| f^\theta_{\text{n-d}}(\theta))^2 \, R_{\text{Int}} \right|. \tag{6.8b}$$

The Monge-Ampère equation (6.2a) is better satisfied when J_I is smaller and the boundary condition (6.2c) is better satisfied when J_B is smaller. If J_I would be 0, then $\boldsymbol{m}$ would satisfy the Monge-Ampère equation on the whole of $\mathcal{E}$ and $\boldsymbol{m}$ would also be differentiable for $r = R_{\text{Int}}$, at least when E and F are smooth enough. Thus, we suspect that errors (6.8) are smaller when J_I is smaller. The value of $\alpha \in (0, 1)$ determines the relative importance between J_I and J_B. A value for α close to 0 implies that a small value for J is mainly achieved by minimizing J_B and a value for α close to 1 implies that a small value for J is mainly achieved by minimizing J_I. We can therefore expect that if we take a larger value for α the differentiability errors for the minimizer $\boldsymbol{m}$ of J will tend to be smaller .

We will use forward and backward finite differences to approximate the one-sided limits in (6.7). This gives us

$$f^r_{\text{n-d}}(\theta_l) = \left| \frac{3(m^r_{N_{R_{\text{Int}}},l} + m^r_{N_{R_{\text{Int}}},l}) - 4(m^r_{N_{R_{\text{Int}}}+1,l} + m^r_{N_{R_{\text{Int}}}-1,l}) + (m^r_{N_{R_{\text{Int}}}+2,l} + m^r_{N_{R_{\text{Int}}}-2,l})}{2h_r} \right| + \mathcal{O}(h_r^2),$$

$$f^\theta_{\text{n-d}}(\theta_l) = \left| \frac{3(m^\theta_{N_{R_{\text{Int}}},l} + m^\theta_{N_{R_{\text{Int}}},l}) - 4(m^\theta_{N_{R_{\text{Int}}}+1,l} + m^\theta_{N_{R_{\text{Int}}}-1,l}) + (m^\theta_{N_{R_{\text{Int}}}+2,l} + m^\theta_{N_{R_{\text{Int}}}-2,l})}{2h_r} \right| + \mathcal{O}(h_r^2).$$

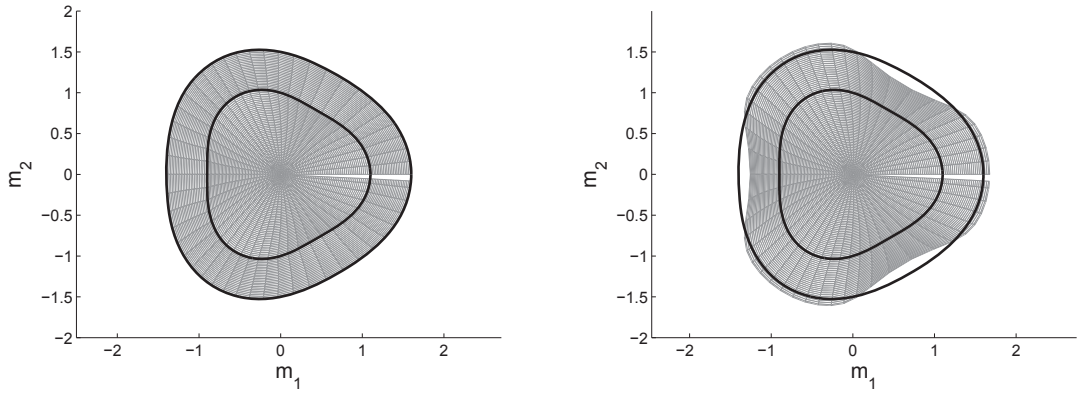
Furthermore, we will use the trapezoidal/midpoint rule to determine the integrals for ε^r_2 and ε^θ_2. Note that there is no difference between the midpoint rule and the trapezoidal rule when begin point and end point of the integration interval coincide.

In Table 6.1, the differentiability errors are given for different values of α. It can be seen that the numerical results are in agreement with our expectations. We denote the value of α used in the least-squares method to solve equations (6.1) by α_{Int} and we denote the value of α used in the adapted least-squares method to solve equations (6.2) by α_{Ext}. We see that the errors decrease if we either increase α_{Int} or α_{Ext}. In Figure 6.3 it can be seen that in the case that α_{Int} and α_{Ext} are close to 1, and the errors are relatively small, the mapping will not satisfy the boundary condition $\boldsymbol{m}(\partial\mathcal{E}) = \partial\mathcal{F}$ particularly well. However, when α_{Int} and α_{Ext} are close to 0, and the errors are relatively large, the boundary condition $\boldsymbol{m}(\partial\mathcal{E}) = \partial\mathcal{F}$ is satisfied very well.

For the second test case the same relationship is found between the α_{Int}, α_{Ext} and the differentiability errors. This is shown in Table 6.2. In Figure 6.6 the functions $f^r_{\text{n-d}}$ and $f^\theta_{\text{n-d}}$ are shown for the two cases of Table 6.2 with the largest and the smallest differentiability errors. It can be seen

α_{Int}	α_{Ext}	ε_2^r $(\cdot 10^{-2})$	ε_2^θ $(\cdot 10^{-2})$	ε_∞^r $(\cdot 10^{-2})$	$\varepsilon_\infty^\theta$ $(\cdot 10^{-2})$
0.2	0.02	3.097	2.329	0.3194	0.2449
0.2	0.2	2.377	1.175	0.2815	0.2132
0.2	0.8	0.2441	0.1686	0.09679	0.07446
0.2	0.98	0.03453	0.03535	0.04556	0.04357
0.8	0.02	3.065	2.263	0.2991	0.2409
0.8	0.2	2.329	1.660	0.2559	0.2060
0.8	0.8	0.2353	0.1531	0.08003	0.06722
0.8	0.98	0.01832	0.01774	0.02963	0.02851

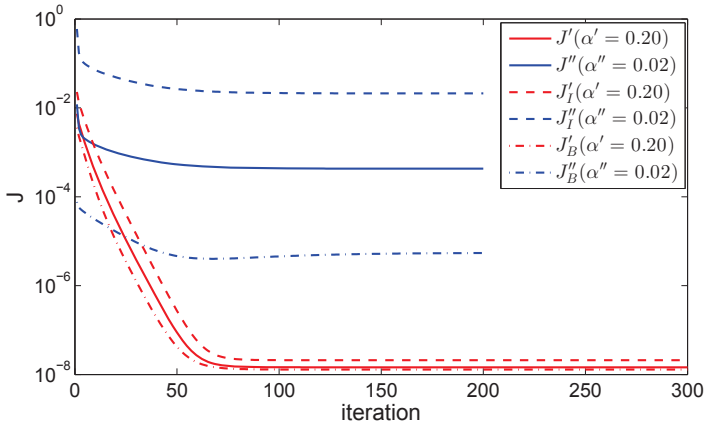
Table 6.1: The differentiability errors for test case 1 for different values of α_{Int} and α_{Ext}. α_{Int} is the value α used when solving the boundary value problem for $\boldsymbol{m}_{\text{Int}}$ in Problem 6.1.1 and α_{Ext} is the value for α used when solving the boundary value problem for $\boldsymbol{m}$ in Problem 6.1.1. We use a 100×100 grid on $\mathcal{E}_{\text{Int}}$ and a 150×150 grid on $\mathcal{E}$.



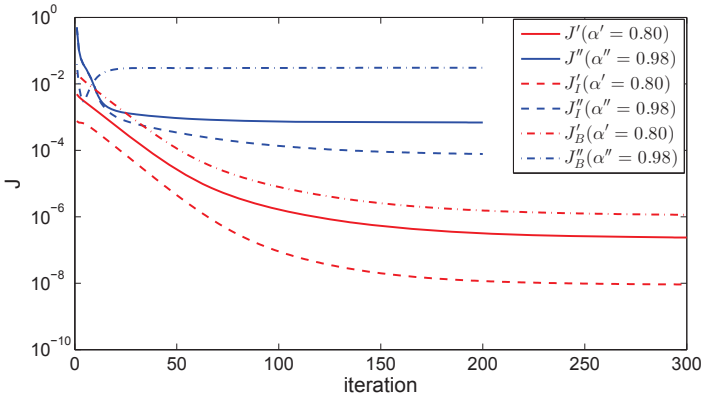
(a) The mapping corresponding to $\alpha_{\text{Int}} = 0.2$ and $\alpha_{\text{Ext}} = 0.02$.

(b) The mapping corresponding to $\alpha_{\text{Int}} = 0.8$ and $\alpha_{\text{Ext}} = 0.98$.

Figure 6.3: The mapping for the two extreme cases of Table 6.1. On the left side the one with the largest errors on the right side the one with the smallest errors. Only half of the grid lines have been plotted.



(a) The functionals to be minimized for $\alpha_{\text{Int}} = 0.2$ and $\alpha_{\text{Ext}} = 0.02$.



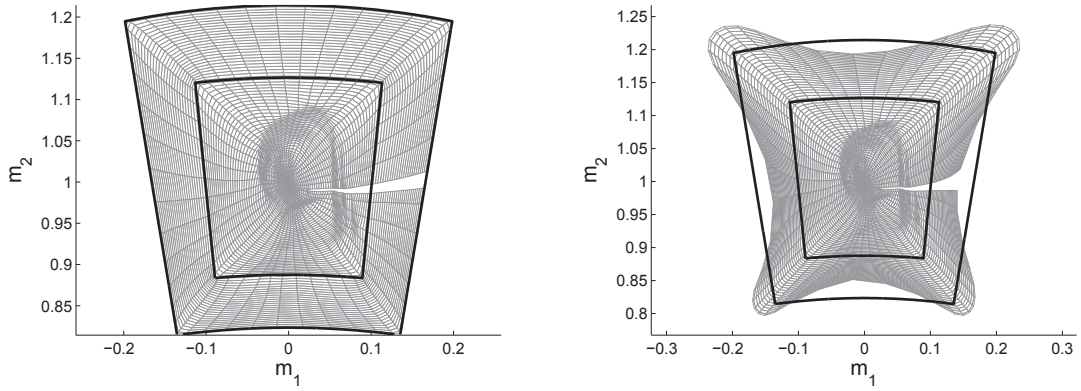
(b) The functionals to be minimized for $\alpha_{\text{Int}} = 0.8$ and $\alpha_{\text{Ext}} = 0.98$.

Figure 6.4: The functionals J, J_I and J_B for the two extreme cases of Table 6.1. On the top the one with the largest errors on the bottom the one with the smallest errors. We used 300 iterations for the least-squares method and 200 iterations for the adapted least-squares method. In this plot $\alpha' = \alpha_{\text{Int}}$ and $\alpha'' = \alpha_{\text{Ext}}$.

α_{Int}	α_{Ext}	ε_2^r $(\cdot 10^{-2})$	ε_2^θ $(\cdot 10^{-2})$	ε_∞^r $(\cdot 10^{-2})$	$\varepsilon_\infty^\theta$ $(\cdot 10^{-2})$
0.2	0.02	0.8666	0.2346	0.3074	0.1035
0.2	0.2	0.7854	0.1964	0.3000	0.09733
0.2	0.8	0.5226	0.1140	0.2616	0.07821
0.2	0.98	0.3577	0.07721	0.2154	0.06676
0.8	0.02	0.5872	0.1870	0.2085	0.08479
0.8	0.2	0.5258	0.1554	0.2012	0.07786
0.8	0.8	0.3198	0.07911	0.1638	0.05816
0.8	0.98	0.1988	0.04844	0.1311	0.05411

Table 6.2: The differentiability errors for test case 2 for different values of α_{Int} and α_{Ext}. α_{Int} is the value α used when solving the boundary value problem for $\boldsymbol{m}_{\text{Int}}$ in Problem 6.1.1 and α_{Ext} is the value for α used when solving the boundary value problem for $\boldsymbol{m}$ in Problem 6.1.1. We use a 100×100 grid on $\mathcal{E}_{\text{Int}}$ and a 150×150 grid on $\mathcal{E}$.

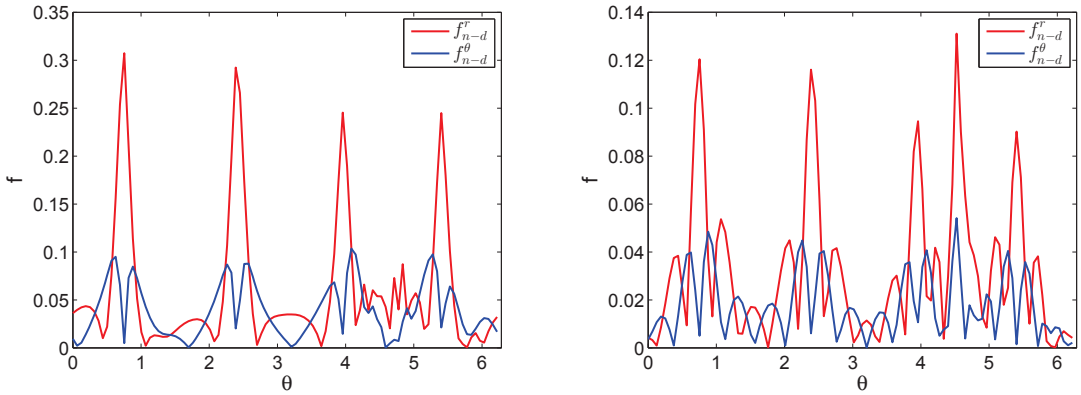
that $f^r_{\text{n-d}}$ has four peaks located at the corners of $\partial\mathcal{F}^2_{\text{Int}}$ and $f^\theta_{\text{n-d}}$ has peaks just before and after each corner. Moreover, $f^r_{\text{n-d}}$ also peaks at the bottom of $\mathcal{F}^2_{\text{Int}}$. This peak is more apparent for the case with $\alpha_{\text{Int}} = 0.8$ and $\alpha_{\text{Ext}} = 0.98$ than for the one with $\alpha_{\text{Int}} = 0.2$ and $\alpha_{\text{Ext}} = 0.02$, because in the former the peaks at the corners are less severe. This fifth peak results from the fact that the function F^2_{Int} attains a relatively high value at the bottom of $\mathcal{F}^2_{\text{Int}}$, because more light is demanded there for the torso of "The Girl with the Pearl Earring". In Figure 6.5 the mapping is shown for the two extreme cases of Table 6.2. The results are again similar to the ones obtained for test case 1. For large values of α_{Int} and α_{Ext} the boundary condition on $\partial\mathcal{E}$ is not very well satisfied. The extension of the grid clusters around the corners in order to better satisfy the Monge-Ampère equations at these locations. For $\alpha_{\text{Int}} = 0.2$ and $\alpha_{\text{Ext}} = 0.02$ the boundary condition on $\partial\mathcal{E}$ is satisfied very well. This can also be seen in the convergence plots for the two cases, which are depicted in Figure 6.8. In Figure 6.7 one of the corners of the mapping is shown for the two cases with largest and smallest differentiability errors. In the plot for $\alpha_{\text{Int}} = 0.2$ and $\alpha_{\text{Ext}} = 0.02$ the differentiability errors are relatively large and it can be seen that there are kinks in the grid lines corresponding to constant θ, near the corner where the original grid of $\mathcal{E}_{\text{Int}}$ is attached to the grid extension for $\mathcal{E}_{\text{Ext}}$. These kinks are almost absent for $\alpha_{\text{Int}} = 0.8$ and $\alpha_{\text{Ext}} = 0.98$. An overal higher grid density can be seen for $\alpha_{\text{Int}} = 0.2,\ \alpha_{\text{Ext}} = 0.02$ near the corner in $\mathcal{F}^2_{\text{Int}}$. This implies that the resulting light intensity distribution will be higher at the corresponding part of the projection screen for $\alpha_{\text{Int}} = 0.2,\ \alpha_{\text{Ext}} = 0.02$ than for $\alpha_{\text{Int}} = 0.8,\ \alpha_{\text{Ext}} = 0.98$. This is indeed what we see in Figure 6.9. In this Figure the ray-trace results are shown for the extended reflector, both for a reflector determined with $\alpha_{\text{Int}} = 0.2$ and $\alpha_{\text{Ext}} = 0.02$ as for a reflector determined with $\alpha_{\text{Int}} = 0.8$ and $\alpha_{\text{Ext}} = 0.98$. In this figure it can also be seen that the boundary conditions on the boundary $\partial\mathcal{E}$ are satisfied less well for the case with $\alpha_{\text{Int}} = 0.8$ and $\alpha_{\text{Ext}} = 0.98$. For the case with $\alpha_{\text{Int}} = 0.2$ and $\alpha_{\text{Ext}} = 0.02$ a black gap has appeared on the bottom of the picture corresponding to the fifth peak in Figure 6.6. For the case with $\alpha_{\text{Int}} = 0.8$ and $\alpha_{\text{Ext}} = 0.98$ this black gap does not appear and also the amount of unwanted extra light around the corners of the picture is less than for the case with $\alpha_{\text{Int}} = 0.2$ and $\alpha_{\text{Ext}} = 0.02$. Thus the the case with values of α close to 1 give significantly better results than the case with values of α closer to 0.



(a) The mapping corresponding to $\alpha_{\text{Int}} = 0.2$ and $\alpha_{\text{Ext}} = 0.02$.

(b) The mapping corresponding to $\alpha_{\text{Int}} = 0.8$ and $\alpha_{\text{Ext}} = 0.98$.

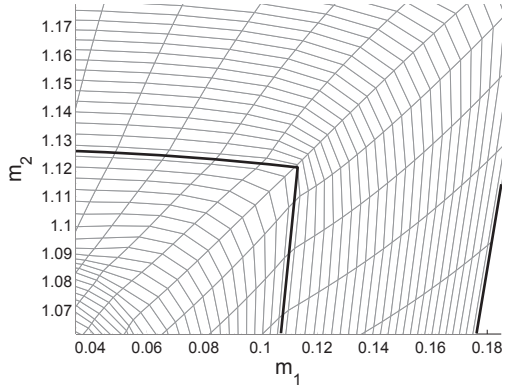
Figure 6.5: The mapping for the two extreme cases of Table 6.2. On the left side the one with the largest errors on the right side the one with the smallest errors. Only half of the grid lines have been plotted.



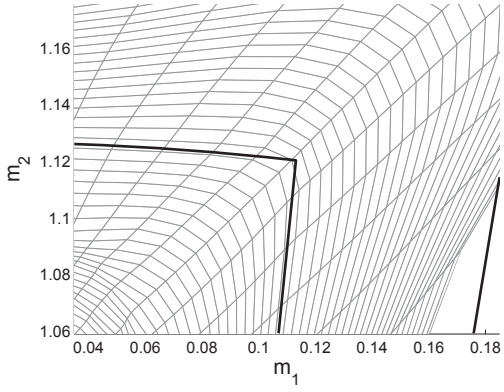
(a) The functions for $\alpha_{\text{Int}} = 0.2$ and $\alpha_{\text{Ext}} = 0.02$.

(b) The functions for $\alpha_{\text{Int}} = 0.8$ and $\alpha_{\text{Ext}} = 0.98$.

Figure 6.6: The functions $f^r_{\text{n-d}}$ and $f^\theta_{\text{n-d}}$ for the two extreme cases of Table 6.2.

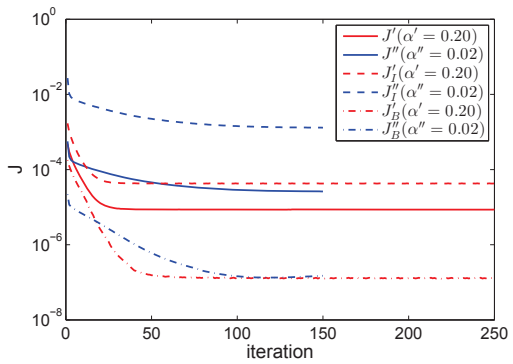


(a) The mapping corresponding to $\alpha_{\text{Int}} = 0.2$ and $\alpha_{\text{Ext}} = 0.02$.

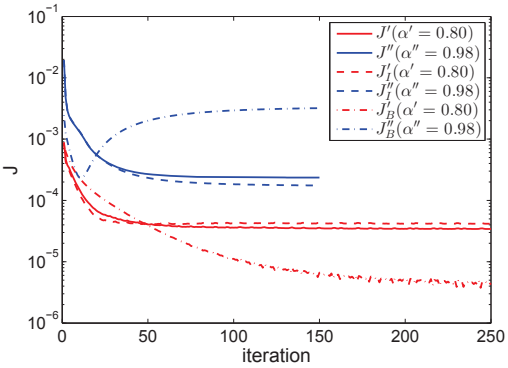


(b) The mapping corresponding to $\alpha_{\text{Int}} = 0.8$ and $\alpha_{\text{Ext}} = 0.98$.

Figure 6.7: The right upper corner of the mapping for the two extreme cases of Table 6.2.

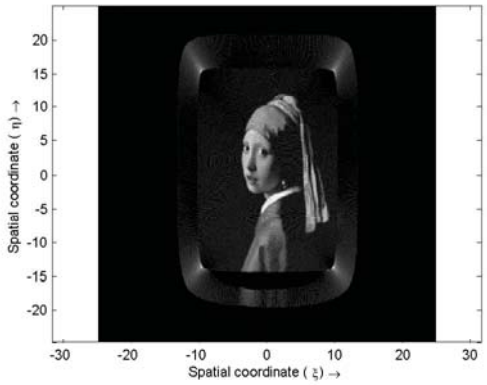


(a) The functionals to be minimized for $\alpha_{\text{Int}} = 0.2$ and $\alpha_{\text{Ext}} = 0.02$.

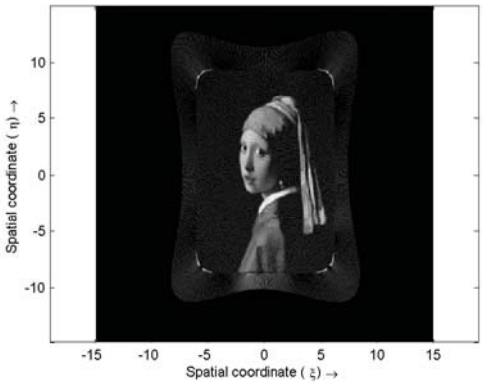


(b) The functionals to be minimized for $\alpha_{\text{Int}} = 0.8$ and $\alpha_{\text{Ext}} = 0.98$.

Figure 6.8: The functionals J, J_I and J_B for the two extreme cases of Table 6.2. On the left side the one with the largest errors on the right side the one with the smallest errors. We used 250 iterations for the least-squares method and 150 iterations for the adapted least-squares method. In this plot $\alpha' = \alpha_{\text{Int}}$ and $\alpha'' = \alpha_{\text{Ext}}$.



(a) For $\alpha_{\text{Int}} = 0.2$ and $\alpha_{\text{Ext}} = 0.02$.



(b) For $\alpha_{\text{Int}} = 0.8$ and $\alpha_{\text{Ext}} = 0.98$.

Figure 6.9: Ray-trace results for the extended reflector on a 1200×1200 polar coordinate grid.

Chapter 7

Conclusions and Final Remarks

In this chapter we will summarize this thesis and make some suggestions for further research. The goal of this graduation project was to take the necessary steps in order to be enable the manufacturing of a reflector which transforms a parallel homogeneous light bundle into an output that projects Figure 1.2 on a projection screen. We will discuss to which extent these steps have been taken and which hurdles still lie ahead.

7.1 Summary

We started out with a chapter on Tensor Calculus in which we introduced the necessary concepts in order to be able to derive the Monge-Ampère equation for the reflector system in a coordinate independent manner. This we did in Chapter 3. The derivation of the Monge-Ampère equation concluded with Theorem 3.3.4 in which finally the coordinate independent Monge-Ampère equation was found, describing the conservation of energy in the reflector system. From this we derived a coordinate specific expression for the Monge-Ampère equation for different coordinate systems on the light source $\mathcal{E}$. We did this in Section 3.4 for polar coordinates with a holonomic basis, polar coordinates with an anholonomic basis and Cartesian coordinates. For Cartesian coordinates we found the form of the Monge-Ampère equation earlier found in [5], showing that our coordinate independent Monge-Ampère equation is consistent with the results obtained earlier. In the rest of Chapter 3 we focussed on formulating an inverse problem for the function u describing the reflector surface. We showed that instead of the implicit condition $\nabla u(\mathcal{E}) = \mathcal{F}$, expressing conservation of global energy, we could equally well use the more explicit boundary condition $\nabla u(\partial\mathcal{E}) = \partial\mathcal{F}$. Furthermore, we discussed in this chapter the relationship between the source $\mathcal{E}$, the gradient space $\mathcal{F}$, the subset of the unit-sphere $\mathcal{G}$ and the projection screen $\mathcal{H}$. We learned that $y = \psi \circ \nabla u$ is a continuously differentiable bijection between $\mathcal{E}$ and $\mathcal{G}$ and that the Monge-Ampère equation can be viewed from the viewpoint of integration by substitution using the continuously differentiable bijection $y : \mathcal{E} \to \mathcal{G}$.

In Chapter 4 we introduced the least-squares method, previously introduced in Cartesian coordinates in [5], for arbitrary coordinate systems. Furthermore, we showed in Section 4.3 that the minimization problem for J_I can still be solved algebraically when we also take into account the trace condition on $\boldsymbol{P}$, which was accidently left out in [5]. In Section 4.4 we used the Calculus of Variations to derive a boundary value problem for $\boldsymbol{m}$ the solution of which minimizes J. While in Cartesian coordinates the boundary value problem consisted of two decoupled scalar equations, which involved the Laplacian of m^x and m^y, we found using an arbitrary coordinate system that this was really a vector equation involving the vector Laplacian. This vector equation mixes the components $\boldsymbol{m}$ and is therefore in general coupled. We found that for polar coordinates this boundary value problem indeed results in two coupled equations.

In Chapter 5 we implemented the least-squares method in polar coordinates with an orthonormal basis. We compared the Cartesian least-squares method with its counter-part in polar co-

ordinates for a disk-shaped source $\mathcal{E} = \mathcal{D}_R$. We found an overall better performance by the least-squares method in polar coordinates. The convergence in polar coordinates was better up to 4 orders of 10. The unwanted bulges often appearing on the edges of $\mathcal{F}$ disappear when using the least-squares method in polar coordinates. Also, the strange behaviour of the grid lines near the corners, which was found in Cartesian coordinates, is no longer present in polar coordinates.

7.2 Recommendations for further research

In order to provide the milling machine that will produce the reflector with workable data still some final steps have to be taken. The determined extrapolated reflector should be determined in the rotated coordinate system of the milling machine, which was depicted in Figure 1.3. It should be analysed in further detail what the resulting chisel accelerations are for the extrapolated reflector in this coordinate system. Subsequently, it should be verified if these chisel accelerations are attainable by the milling machine. Before actually manufacturing the reflector, it would be wise to verify the extrapolated reflector with professional ray-tracing software. If the results of this test are also satisfactory, then the reflector can be produced.

When the boundary of the target $\partial\mathcal{F}$ is not smooth, there seems to be a trade-off between the value of the functional corresponding to the interior J_I and the functional corresponding to the boundary J_B. It seems that for non-smooth $\partial\mathcal{F}$ it is impossible to get both J_I and J_B to equal zero simultaneously. However, if the boundaries $\partial\mathcal{E}$ and $\partial\mathcal{F}$ are smooth enough, and the intensity functions E and F are smooth enough, we suspect this trade-off to disappear. What exactly these smoothness conditions on $\partial\mathcal{E}$, $\partial\mathcal{F}$, E and F should be is something open to further research. The available literature on the Monge-Ampère equation should be consulted on this point. If the required smoothness conditions have been clarified, numerical tests should be performed on a test case satisfying the smoothness conditions to check if the apparent trade-off between J_I and J_B does indeed disappear.

Finally, it is of great interest to generalize the methods presented in this graduation thesis to more general light sources. In practice one often encounters point light sources and extended light sources that do not emit a parallel bundle of light but radiate in certain set of directions with a corresponding intensity. Starting out with the point light source case, for which the corresponding Monge-Ampère-type equation has been determined in [9]. Some first steps have been set in devising numerical methods to solve the inverse reflector problem for the point light source, see for example [17].

Bibliography

[1] Florack, L.M.J. *Course Notes Tensor Calculus and Differential Geometry*. (Unpublished, (in april 2015) available online: http://www.bmia.bmt.tue.nl/people/lflorack/Extensions/2F800.html), 2015

[2] Spivak, M. *A Comprehensive Introduction to Differential Geometry, Vol. 1, 2nd ed.* Berkeley, CA: Publish or Perish Press, 1979

[3] Lee, J.M. *Manifolds and Differential Geometry* American Mathematical Society, 2009

[4] Frankel, T. *The Geometry of Physics, 3rd ed.* Cambridge University Press, 2012

[5] Prins, C. R. Ph.D. Thesis, *Inverse Methods for Illumination Optics*. ISBN 978-90-386-3662-7, 2014

[6] Courant, R.; Hilbert, D. *Methods of Mathematical Physics, volume 2*. John Wiley & Sons, 1989 edition, 1953

[7] Misner, C. W.; Thorne, K. S.; and Wheeler, J. A. *Gravitation*. San Francisco: W. H. Freeman, 1973

[8] Grinfeld, P. *Introduction to Tensor Analysis and the Calculus of Moving Surfaces*. Springer, New York, 2013

[9] Oliker, V.; Newman, E. *On the Energy Conservation Equation in the Reflector Mapping Problem*. Appl. Math. Lett. 6 91-5, 1993

[10] Oliker, V.; Newman, E. *On the Design of a Reflector Antenna II*. Calc. Var. PDE 20 329-341, 2004

[11] Villani, C. *Topics in Optimal Transportation* The American Mathematical Society 2003

[12] Hecht, E. *Optics, 2nd ed.* Addison Wesley 1987

[13] Marsden, J.E.; Tromba A.J. *Vector Calculus, 5th ed.*. W.H. Freeman and Company, 2003

[14] Boyd, S.; Vandenberghe L. *Convex Optimization*. Cambridge University Press, 2004

[15] Tao, T. *Analysis II, 2nd ed.*. Hindustan Book Agency, 2009

[16] Spivak, M. *Calculus on Manifolds*. Addison-Wesley Publishing Company, 1965

[17] Brix, K.; Hafizogullari, Y.; Platen, A. *Solving the Monge-Ampère Equations for the Inverse Reflector Porblem*. Mathematical Models and Methods in Applied Sciences, Volume 25, Issue 5, pp. 803-837, 2015