A least-squares method for the inverse reflector problem in arbitrary orthogonal coordinate systems

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computer Science

CASA-Report 16-17
August 2016

A least-squares method for the inverse reflector problem in arbitrary orthogonal coordinate systems

by

R. Beltman, J.H.M. ten Thije Boonkkamp, W.L. IJzerman

Centre for Analysis, Scientific computing and Applications
Department of Mathematics and Computer Science
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven, The Netherlands
ISSN: 0926-4507

A LEAST-SQUARES METHOD FOR THE INVERSE REFLECTOR PROBLEM IN ARBITRARY ORTHOGONAL COORDINATE SYSTEMS

R. BELTMAN*, J. H. M. TEN THIJE BOONKKAMP*, AND W. L. IJZERMAN*†

Abstract. In this article we solve the inverse reflector problem for a light source emitting a parallel light bundle and a target in the far-field of the reflector by use of a least-squares method. We derive the Monge-Ampère equation, expressing conservation of energy, while assuming an arbitrary coordinate system. We generalize a Cartesian coordinate least-squares method presented earlier by C.R. Prins et al. [13] to arbitrary orthogonal coordinate systems. This generalized least-squares method provides us the freedom to choose a coordinate system suitable for the shape of the light source. This results in significantly increased numerical accuracy. Decrease of errors by factors up to 10^4 is reported. We present the generalized least-squares method and compare its numerical results with the Cartesian version for a disk-shaped light source.

1. Introduction. In the last decades LED lighting technology rapidly developed. The costs of LED lighting constantly decrease, as is expressed by Haitz' law which states that the cost per lumen (power perceived by the human eye) falls by a factor of 10 every decade [1]. Furthermore, LED lighting surpasses traditional lighting in efficacy (lumen per Watt) [2]. As a result, LED lighting systems are used in illumination optics ever more frequently. LED lighting systems are LEDs integrated in an optical system consisting of lenses, reflectors, diffusers and absorbers.

Two classes of methods are used to design these optical systems: *forward methods* and *inverse methods*. In forward methods the optimal optical system is determined through a process of trial and error. A given optical system is tested, the light output of the system is determined by Monte-Carlo ray tracing [4] and subsequent adjustments are made to improve the system. This process then iterates to a more or less satisfactory solution, of which the quality depends to a large extent on the skill of the designer. This method is widely applicable and straightforward, but time consuming. By contrast, in inverse methods the light output of the optical system is related to the geometry of the optical elements by a partial differential equation, the solution of which directly gives the shape of the optical elements. Inverse methods are less straightforward to apply but lead to far more accurate results and are time efficient. Moreover, with inverse methods a diversity of new designs are possible that, due to their complexity, are completely unattainable by direct methods.

The rise of LED lighting has increased the interest in inverse methods because LED lighting operates at much lower temperatures than conventional lighting. This clears the path for the use of easy to mold transparent plastics instead of glass. The optimal shape of these plastic elements can be exactly determined by the inverse method. Moreover, due to active development in diamond turning techniques the arbitrarily shaped elements can be fabricated with increasingly high precision [3].

In this paper we consider an optical system consisting of an incoming parallel bundle of light and a reflecting surface. Parallel bundles occur frequently in LED lighting systems as the result of a converging lens placed on top of divergently emitting LEDs. Given the intensity distribution of the incoming parallel bundle and a desired output distribution, a partial differential equation can be derived for the reflector surface. This partial differential equation turns out to be an equation of the Monge-

*CASA, Eindhoven University of Technology, PO Box 513 5600 MB Eindhoven, The Netherlands.

†Philips Lighting, High Tech Campus 44, 5656 AE Eindhoven, The Netherlands.

Ampère type.

Monge-Ampère type equations also arises in the context of optimal mass transport (OMT). The inverse reflector problem and OMT problem are closely related [5]. OMT concerns, roughly speaking, the problem of filling a hole with a heap of sand from another location. The goal is to do this while minimizing the transportation cost. In inverse optical problems we do not consider a hole and heap of sand, but instead a light source with an emittance and a target with a desired light intensity distribution. It was shown that this problem can be viewed as an OMT problem [6].

Numerical methods for solving OMT problems have been scarce until recently. Benamou and Brenier introduced an augmented Lagrangian method to solve the OMT problem [7]. This approach was further developed by Haber et al. [8]. A numerical method for the Monge-Ampère equation using finite differences was introduced by Froese et al. [9, 10]. This method is robust, but requires a convex target set. Brix et al. [11] solved the inverse reflector problem for a point source by using a collocation method with a tensor-product B-spline basis. For a comprehensive overview of the literature on numerical methods for the inverse reflector problem we refer to the thesis of C.R. Prins [12] and the aforementioned article by Brix et al.

In a recent publication, C. R. Prins et al. [13] introduced a least-squares method (LS method) to solve the OMT problem related to the inverse reflector problem. The LS method solves the inverse reflector problem, i.e., the problem of finding the reflector surface that reflects a parallel bundle of light such that a prescribed luminous intensity pattern is achieved on a projection screen in the far-field of the reflector. The method can handle very complicated source and target intensities. The LS method was used, for example, to determine the reflector surface that reflects a parallel bundle of light to form the luminous intensity pattern corresponding to a gray-scale image of a famous painting by Vermeer.

The LS method determines the shape of the reflector surface by covering the light source with a rectangular grid and computing the height of the reflector in each grid point. This works fine for rectangular light sources, however, for differently shaped light sources the rectangular grid also contains grid points outside of the light source. For these grid points the emittance of the light source is taken to be zero. This approach to non-rectangular light sources is far from optimal and gives results much less satisfying than obtained for rectangular light sources. Most importantly, the boundary condition, which states that the boundary of the source must be mapped to the boundary of the target, is at places very badly satisfied and this makes the method inapplicable for non-rectangular sources. This poses a severe restriction on the applicability of the method in illumination optics. The parallel bundles encountered in illumination optics often result from a converging lens and frequently have disk-shaped cross sections, therefore a numerical method that can handle disk-shaped light sources in a satisfactory way is highly desirable.

The goal of this paper is to present an improved generalized version of the LS method (GLS method) that is applicable to arbitrarily shaped light sources emitting a parallel bundle. We use some concepts from tensor calculus to formulate the inverse reflector problem in coordinate-free form, derive the corresponding coordinate-independent Monge-Ampère equation and generalize the LS method to arbitrary orthogonal coordinate systems. In one of the minimization steps of the GLS method a pair of boundary value problems is solved. In Cartesian coordinates these problems are decoupled, however, in general they are coupled. We present how to deal with this issue. Furthermore, we compare the LS method from [13] with the GLS method presented in this paper. We show that for disk-shaped light sources the GLS method in polar coordinates outperforms the LS method significantly.

This paper is structured as follows. In Section 2 we derive the Monge-Ampère equation describing the reflector surface and formulate the reflector problem for an arbitrary coordinate system. In Section 3 we introduce the GLS method by generalizing the LS method to arbitrary orthogonal coordinate systems. We shed light on the different minimization steps in this method and show how they are different from the Cartesian version of the method. In Section 4 we compare the LS and GLS methods. We will consider two test cases. In both cases we will take a disk-shaped light source and therefore choose polar coordinates as the orthogonal coordinate system for the GLS method. In the first test case the light source is mapped to a square gradient set and in the second test case we consider a target distribution corresponding to a lithograph by the artist M.C. Escher (Figure 1.1). In Section 5 we summarize and discuss the results. This paper contains some appendices. In Appendix A we introduce some concepts from Tensor calculus needed in this paper, and give accurate pointers to classical literature on these matters. In Appendix B and Appendix C one can find some proofs of results given in the main text. The reading of these proofs should not be necessary for understanding the rest of the paper.

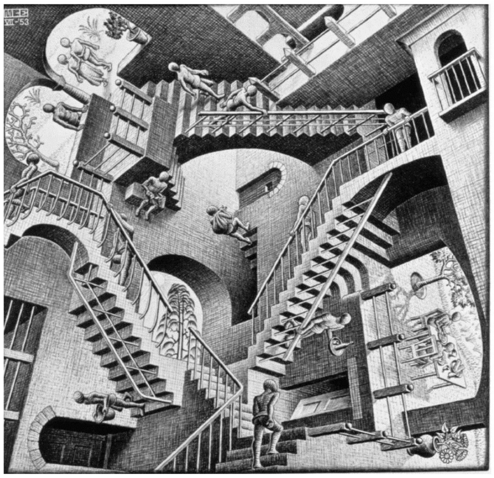

Figure 1.1: Lithograph *Relativity* (1953) by the Dutch artist M.C. Escher who was frequently inspired by mathematics [14]. This lithograph, with its great detail, will serve as the ultimate test.

2. Monge-Ampère equation and inverse reflector problem. Let us consider the optical system of interest. The system consists of a light source and a reflector surface. We embed our optical system in three dimensional Euclidean space. We describe the light source by a set $\mathcal{E} \subset \mathbb{R}^2 \times \{-a\}$, a subset of a plane below and parallel to the x-y plane at a distance $a > 0$. We assume an arbitrary coordinate system on $\mathcal{E}$ with at each point $\boldsymbol{x} \in \mathcal{E}$ corresponding basis vectors $\boldsymbol{e}_1(\boldsymbol{x})$ and $\boldsymbol{e}_2(\boldsymbol{x})$.

We denote by $\boldsymbol{e}^1(\boldsymbol{x})$ and $\boldsymbol{e}^2(\boldsymbol{x})$ the dual basis vectors defined by $\boldsymbol{e}^i(\boldsymbol{e}_j) = \delta^i_j$. Let $(\cdot,\cdot)$ denote the Euclidean inner product on the ambient space $\mathbb{R}^3$ and let $\|\cdot\|$ be the corresponding norm. We denote by $e_{ij} = (\boldsymbol{e}_i, \boldsymbol{e}_j)$ the metric on $\mathcal{E}$ and by $e = \det(e_{ij})$ the determinant of the metric (Appendix A). We assume that the light source emits a parallel bundle of light along the z-axis. The emittance of the light source at a point $\boldsymbol{x} \in \mathcal{E}$ is given by $E(\boldsymbol{x})$ [lm/m^2], where $E : \mathcal{E} \to (0,\infty)$ is the emittance function, which we assume to be continuous. $E(\boldsymbol{x})\sqrt{e}(\boldsymbol{x})\boldsymbol{e}^1(\boldsymbol{x}) \wedge \boldsymbol{e}^2(\boldsymbol{x})$ expresses the light flux through the infinitesimal area element $\sqrt{e}(\boldsymbol{x})\boldsymbol{e}^1(\boldsymbol{x}) \wedge \boldsymbol{e}^2(\boldsymbol{x})$ (Appendix A) on $\mathcal{E}$ centered around $\boldsymbol{x}$. For details on photometric quantities, see for example [15]. The light rays leaving the source will all hit upon the reflector surface. We describe the reflector surface by a function $u : \mathcal{E} \to (-a, \infty)$. A ray leaving from the point $\boldsymbol{x} \in \mathcal{E}$ will travel a distance $a + u(\boldsymbol{x})$ in the z-direction before hitting upon the reflector surface. The function $u : \mathcal{E} \to (-a,\infty)$ is the Monge parameterization of the reflector surface [16]. Note that by definition $u > -a$, because the reflector surface is situated above the source and not allowed to intersect with the source. In what follows we need the function u to be strictly convex and twice continuously differentiable. We will see that strict convexity of u implies that a pair of rays leaving $\mathcal{E}$ from different points will be reflected in different directions. We assume the target to be positioned in the far-field of the reflector. Thus, we assume the rays after reflection to be all originating from one point and we discard the size of the reflector in this respect. In our embedding of the reflector system we let this point coincide with the origin of $\mathbb{R}^3$.

The direction of reflection is given by the law of reflection, which in vector form is given by

$$\boldsymbol{r} = \boldsymbol{i} - 2(\boldsymbol{i}, \boldsymbol{n})\boldsymbol{n}, \tag{2.1}$$

where $\boldsymbol{i}$ is the direction of the incoming ray, $\boldsymbol{n}$ is the direction of the normal on the reflector surface and $\boldsymbol{r}$ is the direction of the ray after reflection. These vectors all have unit length. The direction of an incoming ray will not depend on the point $\boldsymbol{x} \in \mathcal{E}$ at which it leaves the source, however, the normal $\boldsymbol{n}$ on the reflector surface does depend on $\boldsymbol{x}$. The vector $\boldsymbol{i}$ is the unit vector normal to the light source directed at the reflector. We denote this vector by the unit vector $\boldsymbol{e}_3$ as it will complement the local two-dimensional bases on $\mathcal{E}$ to a three-dimensional basis for $\mathbb{R}^3$. The unit normal on the reflector surface pointing down towards the light source can be expressed in terms of the gradient of u and $\boldsymbol{e}_3$ and when we substitute this in (2.1), we obtain

$$\boldsymbol{r}(\boldsymbol{x}) = \boldsymbol{e}_3 + 2\frac{\nabla u(\boldsymbol{x}) - \boldsymbol{e}_3}{\|\nabla u(\boldsymbol{x}) - \boldsymbol{e}_3\|^2}. \tag{2.2}$$

The gradient $\nabla u(\boldsymbol{x})$ is a vector lying in the plane of $\mathcal{E}$ and we interpret it here as a vector in $\mathbb{R}^3$ orthogonal to $\boldsymbol{e}_3$. For all $\boldsymbol{x} \in \mathcal{E}$ the vector $\boldsymbol{r}(\boldsymbol{x})$ is of unit length and by the far-field approximation we may furthermore assume it to have its initial point at the origin. This implies that the vectors $\boldsymbol{r}(\boldsymbol{x})$ lie on the unit sphere, $\mathcal{S}^2$. We can therefore interpret the map given by $\boldsymbol{x} \mapsto \boldsymbol{r}(\boldsymbol{x})$ to be mapping a point on the light source to a point on the unit sphere. We will denote this mapping by $r : \mathcal{E} \to \mathcal{S}^2$.

The reflected light will shine in a set of directions $\mathcal{G} \subset \mathcal{S}^2$. We assume a local coordinate system on $\mathcal{G}$ with basis vectors $\boldsymbol{g}_1$, $\boldsymbol{g}_2$, dual basis vectors $\boldsymbol{g}^1$, $\boldsymbol{g}^2$, corresponding metric g_{ij} and let $g = \det(g_{ij})$. Here a logical choice for $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ would be the pushforward [26, p.89] of $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ by r. Let us describe the luminous intensity in the directions $\mathcal{G}$ by a continuous function $G : \mathcal{G} \to (0,\infty)$. The luminous intensity

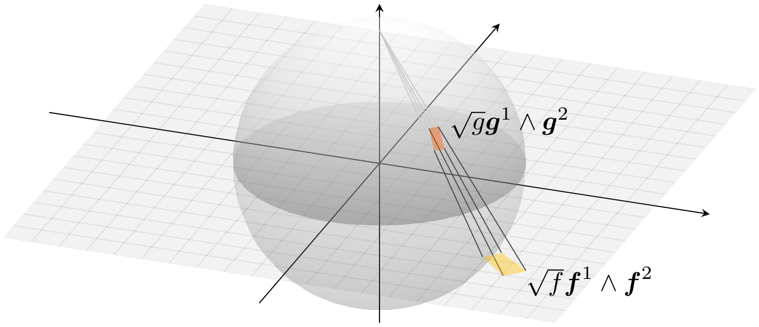
$\sqrt{g}\boldsymbol{g}^1 \wedge \boldsymbol{g}^2$
$\sqrt{f}\boldsymbol{f}^1 \wedge \boldsymbol{f}^2$

Figure 2.1: The inverse stereographic projection maps the surface element $\sqrt{f}\boldsymbol{f}^1 \wedge \boldsymbol{f}^2$ on $\mathbb{R}^2 \times \{0\}$ to the surface element $\sqrt{g}\boldsymbol{g}^1 \wedge \boldsymbol{g}^2$ on $\mathcal{S}^2 \backslash \boldsymbol{e}_3$.

is the light intensity per steradian [lm/sr]. The light flux through an infinitesimal surface area element on $\mathcal{G}$ centered around $\boldsymbol{z} \in \mathcal{G}$ is given by $G(\boldsymbol{z})\sqrt{g}(\boldsymbol{z})\boldsymbol{g}^1(\boldsymbol{z}) \wedge \boldsymbol{g}^2(\boldsymbol{z})$. In practice the couple $\mathcal{G}$ and G will be such that a desired intensity pattern is projected on a screen in the far-field of the reflector. As long as $\mathcal{G}$ is confined to one half of $\mathcal{S}^2$ there is one-to-one correspondence between the couple $\mathcal{G}$ and G and the intensity pattern in the far-field. Details can be found in [12]. We call $\mathcal{G}$ the target set.

The problem we want to solve is informally stated as follows. *Given a light source $\mathcal{E}$ with emittance function E, determine the shape of the reflector such that, after reflection, the intensity pattern in the far-field is given by the target set $\mathcal{G}$ with luminous intensity function G.* This problem is known as the *inverse reflector problem*. Before we will state this problem more formally, we will first, under the assumption $u \in C^2(\mathcal{E})$, derive a partial differential equation from the principle of conservation of luminous flux. The luminous flux through $U \subset \mathcal{E}$ results in a luminous flux through the set $r(U) \subset \mathcal{S}^2$. By conservation of luminous flux these two fluxes must be equal and therefore we have

$$\int_U E\sqrt{e}\boldsymbol{e}^1 \wedge \boldsymbol{e}^2 = \int_{r(U)} G\sqrt{g}\boldsymbol{g}^1 \wedge \boldsymbol{g}^2, \tag{2.3}$$

for every Lebesgue measurable set $U \subset \mathcal{E}$. We can use (2.3) to derive the partial differential equation. To see this we must closely examine the map $r : \mathcal{E} \to \mathcal{S}^2$. From (2.2) it can be seen that $\boldsymbol{r}(\boldsymbol{x})$ only depends on the gradient of u in the point $\boldsymbol{x}$. We can therefore interpret r as the composition $s \circ \nabla u$, i.e., the composition of ∇u and another map which we will denote by s. By the far-field approximation $\boldsymbol{r}$ has its initial point at the origin. This implies that we should interpret ∇u also as a vector with its initial point at the origin. The vector ∇u is by definition parallel to $\mathcal{E}$ and because it has its initial point in the origin it lies in the plane $\mathbb{R}^2 \times \{0\}$. From equation (2.2) we see that s maps a vector $\boldsymbol{v}$ in this plane to the unit sphere according to

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

Closer inspection reveals that this map is the inverse of the stereographic projection pictured in Figure 2.1 [17, p.26]. It is the bijection between $\mathcal{S}^2 \backslash \boldsymbol{e}_3$, i.e., the unit-sphere without its north pole, and $\mathbb{R}^2 \times \{0\}$, the plane intersecting its equator.

$$\begin{tikzcd}\mathcal{E} \arrow[r, "\nabla u"] \arrow[rr, bend right, "r"'] & \mathcal{F} \arrow[r, "s"] & \mathcal{G}\end{tikzcd}$$

Figure 2.2: The mappings and sets involved in the Inverse reflector problem.

We proceed with examining the map $r : \mathcal{E} \to \mathcal{S}^2$ in order to derive the sought partial differential equation. The vector $\nabla u(\boldsymbol{x})$ lies in the plane $\mathbb{R}^2 \times \{0\}$ and has its initial point at the origin, therefore, when we identify the vector $\nabla u(\boldsymbol{x})$ with its endpoint, we can interpret $\nabla u(\mathcal{E})$ as a subset of $\mathbb{R}^2 \times \{0\}$. We will present some features of ∇u in the following lemma.

Lemma 2.1. *Let $u \in C^2(\mathcal{E})$ be strictly convex. The map $\nabla u : \mathcal{E} \to \nabla u(\mathcal{E})$ is a continuously differentiable bijection, satisfying the equation*

$$|\mathrm{D}\nabla u(\boldsymbol{x})| = \frac{\det(H_{ij}(u(\boldsymbol{x})))}{e(\boldsymbol{x})}, \tag{2.4}$$

where $\mathrm{D}\nabla u(\boldsymbol{x})$ is the Jacobian of ∇u in $\boldsymbol{x} \in \mathcal{E}$. Furthermore, $H_{ij}(u(\boldsymbol{x}))\boldsymbol{e}^i \otimes \boldsymbol{e}^j$ is the Hessian tensor (Appendix A) of u in the point $\boldsymbol{x} \in \mathcal{E}$.

The proof of this lemma can be found in Appendix B. We have a similar result for the stereographic projection.

Lemma 2.2. *The inverse of the stereographic projection $s : \mathbb{R}^2 \times \{0\} \to \mathcal{S}^2 \backslash \boldsymbol{e}_3$ is continuously differentiable and hence $s : \nabla u(\mathcal{E}) \to s(\nabla u(\mathcal{E}))$ is a continuously differentiable bijection. Moreover, for the Jacobian of s in $\boldsymbol{y} \in \nabla u(\mathcal{E})$ we have*

$$|\mathrm{D}s(\boldsymbol{y})| = \frac{4}{(1 + \|\boldsymbol{y}\|^2)^2}.$$

The proof of this lemma can also be found in Appendix B. We can use the results of the preceding two lemmas to derive the differential equation expressing conservation of energy. This is stated in the following theorem.

Theorem 2.3. *Assume $\mathcal{E} \subset \mathbb{R}^2 \times \{-a\}$ is convex, closed and bounded, $u \in C^2(\mathcal{E})$ strictly convex and $r = s \circ \nabla u$, where s is the inverse of the stereographic projection. Let $E \in C(\mathcal{E})$ and $G \in C(r(\mathcal{E}))$ be strictly positive and bounded functions. Furthermore, assume we have a coordinate system on $\mathcal{E}$ with metric e_{ij}. Then the function u satisfies the differential equation*

$$\frac{E(\boldsymbol{x})}{G(r(\boldsymbol{x}))} = \frac{4\det(H_{ij}(u))}{e(1 + \|\nabla u\|^2)^2}, \tag{2.5}$$

for every $\boldsymbol{x} \in \mathcal{E}$. This equation is a Monge-Ampère type equation.

Proof. We first remark that (2.5) is independent of the choice of coordinate system. The left hand side of (2.5) and ∇u are clearly independent of the coordinate system in use. Moreover, in the proof of Lemma 2.1 we saw that $\det(H_{ij}(u))/e$ is an invariant, hence the right hand side of (2.5) is also independent of the coordinate system in use.

Both s and ∇u are continuously differentiable injections, therefore we can apply integration by substitution [20, Thm. 7.26]. For every Lebesgue measurable open subset $U \subset \mathcal{E}$ we have

$$\int_{r(U)} G\sqrt{g}\boldsymbol{g}^1 \wedge \boldsymbol{g}^2 = \int_U (G \circ r)|\mathrm{D}s||\mathrm{D}\nabla u|\sqrt{e}\boldsymbol{e}^1 \wedge \boldsymbol{e}^2.$$

Using equation (2.3) we find

$$\int_U E\sqrt{e}\boldsymbol{e}^1 \wedge \boldsymbol{e}^2 = \int_U (G \circ r)|\mathrm{D}s||\mathrm{D}\nabla u|\sqrt{e}\boldsymbol{e}^1 \wedge \boldsymbol{e}^2,$$

for every Lebesgue measurable $U \subset \mathcal{E}$. The continuity of the functions E, G, r, $|\mathrm{D}s|$ and $|\mathrm{D}\nabla u|$ and Lemma 2.1 and Lemma 2.2 imply that equation (2.5) holds in all of $\mathcal{E}$. □

We are now in the position to state the inverse reflector problem in more formal terms. *Given a convex, closed and bounded light source $\mathcal{E}$ with strictly positive and bounded emittance $E \in C(\mathcal{E})$ and a closed target set $\mathcal{G} \subset \mathcal{S}^2$ with desired strictly positive and bounded luminous intensity $G \in C(\mathcal{G})$ such that*

$$\int_{\mathcal{E}} E\sqrt{e}\boldsymbol{e}^1 \wedge \boldsymbol{e}^2 = \int_{\mathcal{G}} G\sqrt{g}\boldsymbol{g}^1 \wedge \boldsymbol{g}^2, \tag{2.6}$$

find a function $u \in C^2(\mathcal{E})$ that satisfies $r(\mathcal{E}) = \mathcal{G}$ and the Monge-Ampère type equation (2.5). The condition $r(\mathcal{E}) = \mathcal{G}$ needs to be satisfied for equation (2.5) to have meaning. We can use the continuously differentiable map s to reformulate the problem in terms of a gradient set $\mathcal{F}$ and function F on this set instead of the target set $\mathcal{G}$ and the luminous intensity G. Using the fact that s^{-1} exists, we define $\mathcal{F} := s^{-1}(\mathcal{G})$, and, using the differentiability of s, we define $F \in C(\mathcal{F})$ by

$$F(\boldsymbol{y}) = (G \circ s)(\boldsymbol{y})|\mathrm{D}s(\boldsymbol{y})| = \frac{4G(s(\boldsymbol{y}))}{(1 + \|\boldsymbol{y}\|^2)^2},$$

for all $\boldsymbol{y} \in \mathcal{F}$. Furthermore, suppose that we have a local coordinate system on $\mathcal{F}$ with basis vectors $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, dual basis vectors $\boldsymbol{f}^1$ and $\boldsymbol{f}^2$, corresponding metric f_{ij} and $f = \det(f_{ij})$. Using integration by substitution we see that (2.6) implies

$$\int_{\mathcal{E}} E\sqrt{e}\boldsymbol{e}^1 \wedge \boldsymbol{e}^2 = \int_{\mathcal{F}} F\sqrt{f}\boldsymbol{f}^1 \wedge \boldsymbol{f}^2. \tag{2.7}$$

The conditions $r(\mathcal{E}) = \mathcal{G}$ translates in the condition $\nabla u(\mathcal{E}) = \mathcal{F}$. These definitions allow us to reformulate the inverse reflector problem.

Inverse reflector problem. *Given a convex, closed and bounded light source $\mathcal{E}$ with strictly positive and bounded emittance $E \in C(\mathcal{E})$ and a closed gradient set $\mathcal{F}$ with strictly positive, bounded and bounded away from zero, function $F \in C(\mathcal{F})$ that satisfy (2.7), find a function $u \in C^2(\mathcal{E})$ that satisfies $\nabla u(\partial\mathcal{E}) = \partial\mathcal{F}$, $e^{ij}H_{ij}(u) > 0$ and the Monge-Ampère type equation*

$$\frac{E(\boldsymbol{x})}{F(\nabla u(\boldsymbol{x}))} = \frac{\det(H_{ij}(u))}{e}. \tag{2.8}$$

Note that we replaced the implicit boundary condition $\nabla u(\mathcal{E}) = \mathcal{F}$ with the more explicit boundary condition $\nabla u(\partial\mathcal{E}) = \partial\mathcal{F}$. The explicit boundary condition is better manageable numerically. We will show in Appendix C that for strictly convex u these two conditions are equivalent. The reason that we demand F to be bounded away from zero, is to be able to show this equivalence. Furthermore, we added the earlier absent constraint $e^{ij}H_{ij}(u) > 0$ demanding the trace of the Hessian to be strictly positive. The fraction E/F is by definition strictly positive, hence the determinant of the Hessian is strictly positive too. From this it follows that the Hessian matrix is strictly positive definite or strictly negative definite, corresponding to either a convex or concave solution, respectively. By demanding the trace of the Hessian to be positive we make sure that only a convex reflector surface is admitted. In this paper we restrict ourselves to this convex solution, however, the algorithm can be easily adapted to find the concave solution instead. In [12, p.96] it is described how one can easily find the concave solution from the convex solution and vice versa. A theorem by Brenier [5, p.66] states that a weak formulation of the Inverse reflector problem admits a unique convex solution. As we earlier argued, the Hessian is strictly positive definite, therefore u is strictly convex [25]. Thus, we conclude that a solution to the Inverse reflector problem as stated above and the same problem but with $\nabla u(\partial\mathcal{E}) = \partial\mathcal{F}$ replaced by $\nabla u(\mathcal{E}) = \mathcal{F}$ are truly equivalent as they both only admit strictly convex solutions. It is, however, not clear that, for all pairs $(\mathcal{E}, E)$ and $(\mathcal{F}, F)$, the unique weak solution of Brenier's theorem is twice continuously differentiable.

3. Least-squares method in arbitrary coordinates. In [13] Prins et al. proposed the LS method to solve the Inverse reflector problem. We will in this section introduce the GLS method, i.e., the generalization of the LS method to arbitrary orthogonal coordinate systems.

We assume an arbitrary coordinate system on the source $\mathcal{E}$ with orthogonal coordinates x^1, x^2, local orthogonal basis vectors $\boldsymbol{e}_1, \boldsymbol{e}_2$ and a metric $e_{ij} = (\boldsymbol{e}_i, \boldsymbol{e}_j)$. The orthogonality of the basis vectors imply $e_{ij} = 0$ for $i \neq j$. We will not try to solve the Inverse reflector problem directly for u, but instead look for a mapping $\boldsymbol{m} : \mathcal{E} \to \mathcal{F}$ representing ∇u such that:

(i) $\boldsymbol{m}$ solves the following boundary value problem

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

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

(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. If $\boldsymbol{m}$ satisfies (ii), then $\hat{\boldsymbol{m}} = \mathrm{d}u$ and hence the tensor

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

must be, by definition (Appendix A), the Hessian of some function and therefore needs to be symmetric (Appendix A). This condition is actually enough to ensure that $\boldsymbol{m}$ equals the gradient of some function. The symmetry of $\nabla\hat{\boldsymbol{m}}$ implies $\nabla \times \boldsymbol{m} = 0$. To see this let us interpret $\boldsymbol{m}$ as a vector in $\mathbb{R}^3$. The component $m^3 = 0$ and hence the curl is given by [26, p.96]

$$\nabla \times \boldsymbol{m} = \frac{1}{\sqrt{e}}\Big((\nabla_{\boldsymbol{e}_2}m_1 - \Gamma^i_{12}m_i) - (\nabla_{\boldsymbol{e}_1}m_2 - \Gamma^i_{21}m_i)\Big)\boldsymbol{e}_3.$$

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 *conservative.* A conservative field on a simply connected domain always equals the gradient of some function, see for example [24, p.551]. Thus we conclude that $\boldsymbol{m}$ equals the gradient of some function $u \in C^2(\mathcal{E})$ if and only if $\nabla\hat{\boldsymbol{m}}$ is symmetric.

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 [25, 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

$$\boldsymbol{H}(u)(\boldsymbol{x}, \boldsymbol{x}) = H_{ij}x^i x^j = x_k e^{ki} H_{ij} x^j = \boldsymbol{x}^T(e^{ki}H_{ij})\boldsymbol{x}.$$

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

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

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.[1] Thus asking for more than $\nabla\hat{\boldsymbol{m}}$ to be positive semi-definite would be too restrictive. The numerical method that we will introduce solves the following boundary value problem (BVP):

Transport BVP. *Find a continuously differentiable* $\boldsymbol{m}$ *that satisfies*

$$\frac{\det(\nabla\hat{\boldsymbol{m}}(\boldsymbol{x}))}{e(\boldsymbol{x})} = \frac{E(\boldsymbol{x})}{F(\boldsymbol{m}(\boldsymbol{x}))}, \quad \boldsymbol{x} \in \mathcal{E}, \tag{3.1a}$$

$$\boldsymbol{m}(\partial\mathcal{E}) = \partial\mathcal{F}, \tag{3.1b}$$

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 and bounded on* $\mathcal{E}$ *and* $\mathcal{F}$*, respectively, such that (2.7) is satisfied and* F *is bounded away from zero.* If u is a solution to the Inverse reflector problem, then $\boldsymbol{m} = \nabla u$ will be a solution to Transport BVP. The reverse statement is not true because a solution $\boldsymbol{m}$ of Transport BVP may be such that the u in $\boldsymbol{m} = \nabla u$ is convex but not strictly convex. Transport BVP thus allows also for convex solutions which are not strictly convex.

We will numerically solve Transport BVP by starting with an initial guess $\boldsymbol{m}^0$ and improving this initial guess in an iterative manner. We will try to approximate $\boldsymbol{m}$ satisfying equation (3.1a) by minimizing the functional

$$J_{\mathrm{I}}(\boldsymbol{m}, \boldsymbol{P}) := \frac{1}{2}\int_{\mathcal{E}} \|\nabla\hat{\boldsymbol{m}} - \boldsymbol{P}\|^2 \sqrt{e}\,\boldsymbol{e}^1 \wedge \boldsymbol{e}^2, \tag{3.2a}$$

over the space

$$\mathcal{P}(\boldsymbol{m}) := \left\{\boldsymbol{P} \in \boldsymbol{T}_2^0(\mathcal{E})_{C^1} \mid [\det(p_{ij}(\boldsymbol{x})) = \frac{e(\boldsymbol{x})E(\boldsymbol{x})}{F(\boldsymbol{m}(\boldsymbol{x}))},\ \boldsymbol{P}(\boldsymbol{x}) \text{ is spsd}\right\}, \tag{3.2b}$$

where "spsd" stands for symmetric positive semi-definite and $\boldsymbol{P} = p_{ij}\boldsymbol{e}^i \otimes \boldsymbol{e}^j$. Furthermore, we use $\boldsymbol{T}_2^0(\mathcal{E})_{C^1}$ to denote the space of continuously differentiable tensor fields of contravariant rank 0 and covariant rank 2, which assign to each point $\boldsymbol{x} \in \mathcal{E}$ a tensor in the tangent space at $\boldsymbol{x}$ to $\mathcal{E}$, which we denote by $\boldsymbol{T}_2^0(T_{\boldsymbol{x}}\mathcal{E})$. It seems as if we demand more smoothness than necessary, because Transport BVP and (3.2a) suggest

[1] 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.

that $\boldsymbol{P}$ only needs 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 (3.2a) is defined in the following way. Let $\boldsymbol{A}, \boldsymbol{B} \in \boldsymbol{T}_2^0(T_{\boldsymbol{x}}\mathcal{E})$, where $\boldsymbol{A} = a_{ij}\boldsymbol{e}^i \otimes \boldsymbol{e}^j$ and $\boldsymbol{B} = b_{ij}\boldsymbol{e}^i \otimes \boldsymbol{e}^j$, then

$$\boldsymbol{A} : \boldsymbol{B} := e^{ik}e^{jl}a_{ij}b_{kl}, \tag{3.3}$$

defines an inner product on $\boldsymbol{T}_2^0(T_{\boldsymbol{x}}\mathcal{E})$. This inner product on $\boldsymbol{T}_2^0(T_{\boldsymbol{x}}\mathcal{E})$ is induced by the metric. The fact that this is indeed an inner product follows from the symmetry, linearity and positivity of the metric $\boldsymbol{e}$. Let $\|\cdot\|$ be the norm associated with this inner product.[2] It is clear that if $J_{\mathrm{I}} = 0$, $\boldsymbol{m}$ will satisfy equation (3.1a) and $\nabla\hat{\boldsymbol{m}}$ will be symmetric positive semi-definite.

To satisfy the boundary condition (3.1b) we will minimize another functional simultaneously. The functional we will minimize is given by

$$J_{\mathrm{B}}(\boldsymbol{m}, \boldsymbol{b}) := \frac{1}{2}\oint_{\partial\mathcal{E}} \|\boldsymbol{m} - \boldsymbol{b}\|^2 \mathrm{d}s, \tag{3.4a}$$

and will be minimized over the space

$$\mathcal{B} := \left\{\boldsymbol{b} \in \boldsymbol{T}_0^1(\partial\mathcal{E})_C \mid \boldsymbol{b}(x) \in \partial\mathcal{F}\right\}, \tag{3.4b}$$

for an arc-length parameterization of the boundary and with $\boldsymbol{T}_0^1(\mathcal{E})_C$ the space of continuous vector fields on $\mathcal{E}$. Analogously to the functional J_{I} we notice that if $J_{\mathrm{B}} = 0$, $\boldsymbol{m}$ satisfies equation (3.1b).

Our goal is to minimize J_{I} and J_{B} simultaneously. In order to do this we define a third functional:

$$J(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b}) := \alpha J_{\mathrm{I}}(\boldsymbol{m}, \boldsymbol{P}) + (1-\alpha)J_{\mathrm{B}}(\boldsymbol{m}, \boldsymbol{b}) \tag{3.5}$$

with $\alpha \in (0, 1)$. This functional we will minimize for $\boldsymbol{m}$ over the space $\mathcal{M} := \boldsymbol{T}_2^0(\mathcal{E})_{C^2}$. One iteration of the numerical method consists of three steps. Assume that $\boldsymbol{m}^n$ is given. In order to determine $\boldsymbol{m}^{n+1}$ three steps are performed subsequently:

$$\boldsymbol{b}^{n+1} = \underset{\boldsymbol{b}\in\mathcal{B}}{\operatorname{argmin}}\, J_{\mathrm{B}}(\boldsymbol{m}^n, \boldsymbol{b}), \tag{3.6a}$$

$$\boldsymbol{P}^{n+1} = \underset{\boldsymbol{P}\in\mathcal{P}(\boldsymbol{m}^n)}{\operatorname{argmin}}\, J_{\mathrm{I}}(\boldsymbol{m}^n, \boldsymbol{P}), \tag{3.6b}$$

$$\boldsymbol{m}^{n+1} = \underset{\boldsymbol{m}\in\mathcal{M}}{\operatorname{argmin}}\, J(\boldsymbol{m}, \boldsymbol{P}^{n+1}, \boldsymbol{b}^{n+1}). \tag{3.6c}$$

To solve these minimization problems we will cover our light source with a grid. The grid will be an orthogonal curvilinear grid with as grid lines a finite set of the coordinate lines of the coordinate system in use. The continuous minimization problems (3.6) will then be translate to discontinuous problems on this grid.

The first minimization step, step (3.6a), can be performed in an efficient pointwise way as discussed in [13]. No changes are made to this minimization step and therefore we do not further discuss it here. The minimization step (3.6b) is discussed

[2] 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.

quite extensively. A new geometrical interpretation of this minimization is presented, which provides increased insight and clarifies the intricate expressions of [13]. This allows us to algebraically determine the minimizer for this problem. Also minimization problem (3.6c) is covered in great detail, because this minimization problem becomes substantially more involved for arbitrary coordinate systems. We start with minimization problem (3.6b).

3.1. Minimization of J_{I}. The integrand of J_{I} does not contain derivatives of $\boldsymbol{P}$, therefore we can carry out the minimization for each grid point $\boldsymbol{x} \in \mathcal{E}$ individually. For each grid point $\boldsymbol{x} \in \mathcal{E}$ we want to minimize $\|\nabla \hat{\boldsymbol{m}}(\boldsymbol{x}) - \boldsymbol{P}(\boldsymbol{x})\|^2/2$. 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 $\boldsymbol{D} = d_{ij} \boldsymbol{e}^i \otimes \boldsymbol{e}^j$, 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

$$\begin{aligned}
&\frac{1}{2}\|\boldsymbol{D} - \boldsymbol{P}\|^2 \\
&\quad = \frac{1}{2}\big(d_{ij} - p_{ij}\big)\big(d_{kl} - p_{kl}\big)e^{ik}e^{jl} \\
&\quad = \frac{1}{2e}\bigg(e^{11}e_{22}(d_{11} - p_{11})^2 + (d_{12} - p_{12})^2 + (d_{21} - p_{21})^2 + e^{22}e_{11}(d_{22} - p_{22})^2\bigg),
\end{aligned}$$

where we used the fact that the basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$ is orthogonal and hence (e_{ij}) is diagonal. The tensor $\boldsymbol{P}(\boldsymbol{x})$ 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

$$(e^{ij} p_{jk}) = \begin{pmatrix} e^{11}p_{11} & e^{11}p_{12} \\ e^{22}p_{12} & e^{22}p_{22} \end{pmatrix}, \tag{3.7}$$

where we used that $p_{21} = p_{12}$. Let a transformation matrix be given by $T = \operatorname{diag}\big(\sqrt{e_{11}}, \sqrt{e_{22}}\big)$. We use this transformation to make $(e^{ij} p_{jk})$ symmetric:

$$T(e^{ij} p_{jk})T^{-1} = \begin{pmatrix} e^{11}p_{11} & p_{12}/\sqrt{e} \\ p_{12}/\sqrt{e} & e^{22}p_{22} \end{pmatrix}. \tag{3.8}$$

A quick calculation shows that the eigenvalues of the matrix $(e^{ij} p_{jk})$, and hence also of the matrix $T(e^{ij} p_{jk})T^{-1}$, are given by

$$\mu_{\pm} = \frac{1}{2e}\left(e_{22}p_{11} + e_{11}p_{22} \pm \sqrt{(e_{22}p_{11} + e_{11}p_{22})^2 - 4e\det(p_{ij})}\right), \tag{3.9}$$

which are both real since the matrix $T(e^{ij} p_{jk})T^{-1}$ is symmetric. It is a familiar result that a matrix is positive semi-definite if and only if its eigenvalues are nonnegative. This implies that $(e^{ij} p_{jk})$ is positive semi-definite if and only if the matrix in (3.8) is positive semi-definite. The matrix in (3.8) 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 (3.8) are nonnegative, i.e., if and only if $e^{11}p_{11} + e^{22}p_{22} \geq 0$ and $(p_{11}p_{22} - p_{12}^2)/e \geq 0$. 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 $\det(p_{ij}) \geq 0$.

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. Let us now, to get rid of the metric altogether, introduce the variables

$$\bar{p}_{11} := e^{11}p_{11}, \qquad \bar{p}_{12} := p_{12}/\sqrt{e}, \qquad \bar{p}_{22} := e^{22}p_{22}, \tag{3.10}$$

$$\bar{d}_{11} := e^{11}d_{11}, \qquad \bar{d}_{22} := e^{22}d_{22}, \qquad \bar{d}_{12} := (d_{12}+d_{21})/(2\sqrt{e}). \tag{3.11}$$

In these new variables we can give a more convenient reformulation of the minimization problem. To do this we also drop a constant term $(d_{12}-d_{21})^2/(4e)$ from the function to minimize. We may do this as it does not effect the minimizers. The reformulated problem is given as follows.

Minimization Problem. *Given the symmetric matrix*

$$\bar{D} = \begin{pmatrix} \bar{d}_{11} & \bar{d}_{12} \\ \bar{d}_{12} & \bar{d}_{22} \end{pmatrix},$$

with $\bar{d}_{11}$, $\bar{d}_{12}$ and $\bar{d}_{22}$ as defined in (3.11), find the symmetric matrix

$$\bar{P} = \begin{pmatrix} \bar{p}_{11} & \bar{p}_{12} \\ \bar{p}_{12} & \bar{p}_{22} \end{pmatrix},$$

that minimizes the function

$$H(\bar{P}) := \frac{1}{2}\|\bar{D}-\bar{P}\|^2, \tag{3.12}$$

under the constraints $\det(\bar{P}) = E/F$ *and* $\operatorname{tr}(\bar{P}) \geq 0$*, where the norm used in (3.12) is the Frobenius norm for matrices, defined as* $\|A\| = \sqrt{\sum_{i,j} a_{ij}^2}$ *for a matrix* $A = (a_{ij})$.

From the relations (3.10) the minimizer (p_{11}, p_{12}, p_{22}) can be found once the minimizer $(\bar{p}_{11}, \bar{p}_{12}, \bar{p}_{22})$ of Minimization Problem has been found. Furthermore, we have $H(\bar{P}) = \|\boldsymbol{D}-\boldsymbol{P}\|^2/2 - (d_{12}-d_{21})^2/(4e)$. We solve Minimization Problem algebraically by using the method of Lagrange multipliers. Besides this we give a graphical 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.

3.1.1. Lagrange minimizers and their geometric representation. We find the minimizers of Minimization Problem with the help of the *Lagrange function*

$$\Lambda(\bar{P};\lambda) = H(\bar{P}) + \lambda\left(\det\bar{P} - \frac{E}{F}\right). \tag{3.13}$$

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

$$\bar{p}_{11} + \lambda\bar{p}_{22} = \bar{d}_{11}, \tag{3.14a}$$

$$\lambda\bar{p}_{11} + \bar{p}_{22} = \bar{d}_{22}, \tag{3.14b}$$

$$(1-\lambda)\bar{p}_{12} = \bar{d}_{12}, \tag{3.14c}$$

$$\bar{p}_{11}\bar{p}_{22} - \bar{p}_{12}^2 = E/F. \tag{3.14d}$$

In the Lagrange function (3.13) the condition $\operatorname{tr}(\bar{P}) \geq 0$ has not been taken into account, hence a solution of (3.14a)-(3.14d) might have $\operatorname{tr}(\bar{P}) < 0$. In what follows, we will show that there always exists a solution to (3.14a)-(3.14d) such that $\operatorname{tr}(\bar{P}) \geq 0$.

Let us now give a geometric interpretation of the Lagrange minimizers. The Lagrange minimizers correspond to a joint tangent plane of a hyperboloid and an ellipsoid. We introduce the function value $H(\bar{P}) = C_H$. By definition $C_H \geq 0$. Every value of C_H corresponds to an iso-surface of the function H. By definition of H we have

$$\left(\frac{\bar{p}_{11} - \bar{d}_{11}}{\sqrt{2C_H}}\right)^2 + \left(\frac{\bar{p}_{12} - \bar{d}_{12}}{\sqrt{C_H}}\right)^2 + \left(\frac{\bar{p}_{22} - \bar{d}_{22}}{\sqrt{2C_H}}\right)^2 = 1. \tag{3.15}$$

Equation (3.15) describes an ellipsoid in $\mathbb{R}^3$ with center $(\bar{d}_{11}, \bar{d}_{12}, \bar{d}_{22})$ and semi-axes $\sqrt{2C_H}$, $\sqrt{C_H}$ and $\sqrt{2C_H}$. Thus the iso-surfaces of H can be interpreted as ellipsoids in $\mathbb{R}^3$.

The constraint $\det(\bar{P}) = E/F$ describes an hyperboloid in $\mathbb{R}^3$ with symmetry axes given by $\bar{p}_{11} = \bar{p}_{22}$ and $\bar{p}_{12} = 0$. To see this we will rotate our coordinate system to align the symmetry axes with our coordinate axes. We perform the rotation given by

$$\begin{aligned} p_1 &:= (\bar{p}_{11} - \bar{p}_{22})/\sqrt{2}, & p_2 &:= \operatorname{tr}(\bar{P})/\sqrt{2}, & p_3 &:= \bar{p}_{12}, \\ d_1 &:= (\bar{d}_{11} - \bar{d}_{22})/\sqrt{2}, & d_2 &:= \operatorname{tr}(\bar{D})/\sqrt{2}, & d_3 &:= \bar{d}_{12}. \end{aligned}$$

Using this transformation, the constraint $\det(\bar{P}) = E/F$ can be rewritten as

$$\left(\frac{p_1}{\sqrt{2E/F}}\right)^2 - \left(\frac{p_2}{\sqrt{2E/F}}\right)^2 + \left(\frac{p_3}{\sqrt{E/F}}\right)^2 = -1. \tag{3.16}$$

This equation describes a hyperboloid of two separate sheets. One sheet is located in the half-space $p_2 > 0$ and the other one is located in the half-space $p_2 < 0$. The distance from the origin to the extremum of the sheet with $\operatorname{tr}(\bar{P}) > 0$ and the extremum of the sheet with $\operatorname{tr}(\bar{P}) < 0$ is both $\sqrt{2E/F}$.

Equation (3.15) transforms to

$$\left(\frac{p_1 - d_1}{\sqrt{2C_H}}\right)^2 + \left(\frac{p_2 - d_2}{\sqrt{2C_H}}\right)^2 + \left(\frac{p_3 - d_3}{\sqrt{C_H}}\right)^2 = 1.$$

We see (Figure 3.1) that the principal axes of both the ellipsoids and the hyperboloids are such that the p_1- and p_2-principal axis are equally long and $\sqrt{2}$ times the length of the p_3-principal axis. This fact will play a role in the minimization problem.

The local minimizers of the Lagrange function (3.13) are exactly the points where an iso-surface of H is tangent to the hyperboloid. This can be seen from the equations (3.14) in the following way. Equation (3.14d) implies that a local minimizer of the Lagrange function is a point on the hyperboloid. A minimizer of the Lagrange function Λ is a local minimum of H when confined 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, because this iso-surface corresponds to the smallest value of H on the hyperboloid. The plane $p_2 = \operatorname{tr}(\bar{P})/\sqrt{2} = 0$ lies precisely between the two sheets of the hyperboloid. Thus, only the points where an iso-surface of H is tangent to the sheet of the hyperboloid with $\operatorname{tr}(\bar{P}) > 0$ are actual minimizers of Minimization Problem. In Figure 3.2 this is illustrated. The global minimizer corresponds to the smallest ellipsoid that is tangent to the upper sheet of the hyperboloid.

In the remaining part of this section we will algebraically solve the system of equations (3.14). We will verify the algebraic solutions that we find by these graphical representations. This allows us to get more intuition for the problem and visualize symmetries that are not directly apparent from the equations (3.14a) - (3.14d).

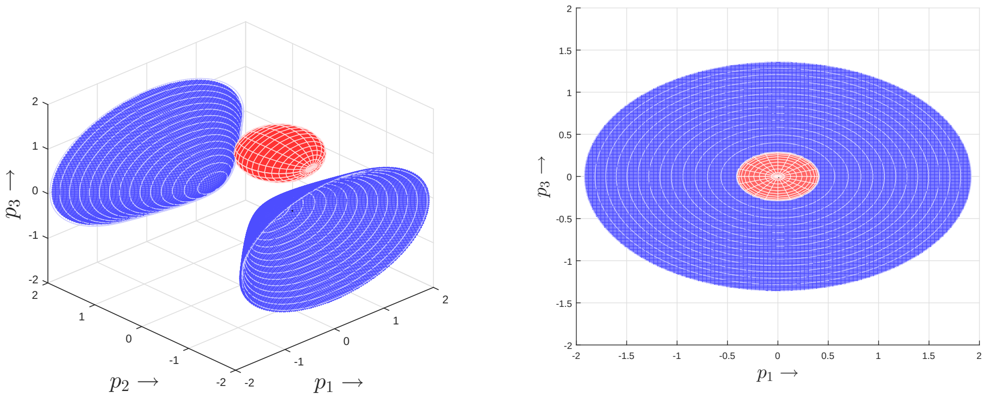

Figure 3.1: In this figure an example of an ellipsoidal iso-surface of H and a hyperboloid are shown from two different perspectives. We see that the principal p_1- and p_3-axis have the same proportion for the hyperboloid and the ellipsoid.

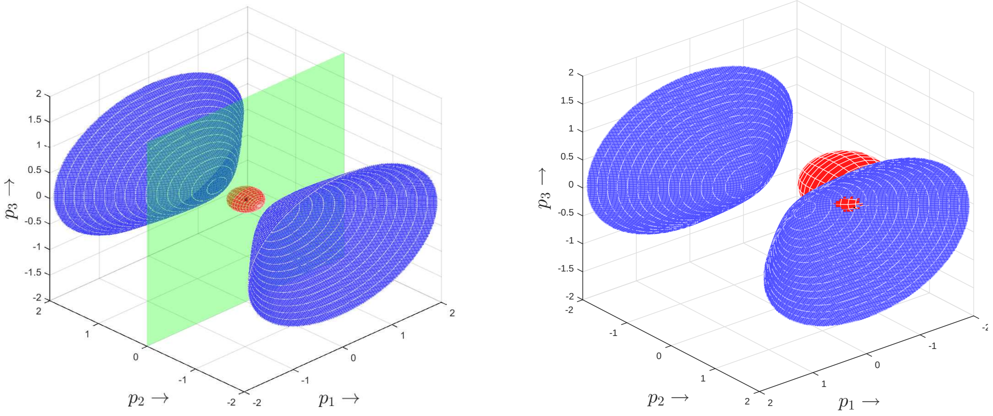

Figure 3.2: On the left side the two sheets of the hyperboloid and the dividing plane $p_2 = \mathrm{tr}(\bar{P})/\sqrt{2} = 0$ are shown. On the right side an example of an ellipsoid which is tangent to the hyperboloid with $\mathrm{tr}(\bar{P}) > 0$ is shown. Some red of the ellipsoid can be seen through the hyperboloid. This point is the minimizer.

3.1.2. Determining the minimizers. We will show that for each given $\bar{D}$ we can find $\bar{P}$ that is the solution of Minimization Problem. If $\lambda \neq \pm 1$, we can invert equations (3.14a) - (3.14b). Doing this we obtain

$$\bar{p}_{11} = \frac{\lambda \bar{d}_{22} - \bar{d}_{11}}{\lambda^2 - 1}, \qquad \bar{p}_{12} = \frac{\bar{d}_{12}}{1 - \lambda}, \qquad \bar{p}_{22} = \frac{\lambda \bar{d}_{11} - \bar{d}_{22}}{\lambda^2 - 1}. \tag{3.17}$$

However, these equations only hold if $\lambda \neq \pm 1$. From equations (3.14a) - (3.14b) we have the following immediate logical implications:

$$\lambda = 1 \implies (\bar{d}_{11} = \bar{d}_{22} \wedge \bar{d}_{12} = 0), \qquad \lambda = -1 \implies (\bar{d}_{11} = -\bar{d}_{22}).$$

From these implications we see there are only two situations that have to be dealt with separately, namely the cases $(\bar{d}_{11} = \bar{d}_{22} \wedge \bar{d}_{12} = 0)$ and $(\bar{d}_{11} = -\bar{d}_{22})$. When we are not in one of these two cases, the solution (3.17) holds. We will now treat the three different cases in turn, starting out with the general case.

Lemma 3.1. *If $(\bar{d}_{11} \neq \bar{d}_{22} \vee \bar{d}_{12} \neq 0)$ and $(\bar{d}_{11} \neq -\bar{d}_{22})$, the global minimizer to Minimization Problem is given by equations (3.17). In these expressions λ is given by one of the following four expressions:*

$$\begin{aligned} \lambda_i &= -\sqrt{\frac{y}{2}} + (-1)^i \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} + \frac{a_1}{2a_4\sqrt{2y}}}, \quad i = 1, 2, \\ \lambda_i &= \sqrt{\frac{y}{2}} + (-1)^i \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} - \frac{a_1}{2a_4\sqrt{2y}}}, \quad i = 3, 4. \end{aligned} \tag{3.18}$$

In (3.18) y is given by the following two sets of equations:

$$\begin{aligned} y &= A + \frac{Q}{A} - \frac{b_2}{3}, \qquad A = -\operatorname{sgn}(R)\big(|A| + \sqrt{R^2 - Q^3}\big)^{1/3}, \\ R &= \frac{2b_2^3 - 9b_1b_2 + 27b_0}{54}, \quad Q = \frac{b_2^3 - 3b_1}{9}, \end{aligned} \tag{3.19}$$

and

$$\begin{aligned} a_4 &= \frac{E}{F}, \qquad a_2 = -2a_4 - \det(\bar{D}), \quad a_1 = \|\bar{D}\|^2, \quad a_0 = a_4 - \det(\bar{D}), \\ b_0 &= -\frac{a_1^2}{8a_4^2}, \quad b_1 = \frac{a_2^2 - 4a_0a_4}{4a_4^2}, \qquad b_2 = \frac{a_2}{a_4}. \end{aligned} \tag{3.20}$$

At least one of the four choices for λ is such that the requirement $\operatorname{tr}(\bar{P}) > 0$ is satisfied by (3.17).

Proof. Substituting the expressions (3.17) in (3.14d) we obtain the following quartic polynomial $\Pi(\lambda) := a_4\lambda^4 + a_2\lambda^2 + a_1\lambda + a_0 = 0$, where the coefficients are as given in (3.20). In [13] it is shown that this polynomial admits the four solutions (3.18). Since $a_4 = E/F > 0$ we have $\lim_{\lambda \to \pm\infty} \Pi(\lambda) = \infty$. Furthermore, we can rewrite $\Pi(\lambda)$ as

$$\Pi(\lambda) = a_4(\lambda^2 - 1)^2 - (a_0 - a_4)(\lambda^2 + 1) + a_1\lambda.$$

From this we see that $\Pi(-1) = -(\bar{d}_{11} + \bar{d}_{22})^2$. By assumption $\bar{d}_{11} \neq -\bar{d}_{22}$, hence $\Pi(-1) < 0$. From this inequality combined with the fact that $\Pi(\lambda) \to +\infty$ for $\lambda \to \pm\infty$ it follows by the Intermediate Value Theorem that Π must have at least two real roots, one smaller than -1 and one larger than -1. From (3.14) it follows that $\operatorname{tr}(\bar{P}) = \operatorname{tr}(\bar{D})/(1 + \lambda)$. This shows that for one of the two real roots it holds that $\operatorname{tr}(\bar{P}) > 0$, while for the other real root it holds that $\operatorname{tr}(\bar{P}) < 0$.

We now have established the fact that one of the four λ in (3.18) is such that (3.17) is a minimum of the Lagrange function such that $\operatorname{tr}(\bar{P}) > 0$, thereby it follows that a global minimizer exists. Moreover, the minimizer is given by (3.17), with λ

given by one of the real roots of (3.18). The global minimizer will be found by checking for which of the four λ_i $(i = 1, \ldots, 4)$ the function H is minimal. □

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

Lemma 3.2. *When $\bar{d}_{11} = -\bar{d}_{22}$, the global minimizer to Minimization Problem is given by*

$$\bar{p}_{11} = \frac{1}{2}\left(\bar{d}_{11} + \sqrt{\bar{d}_{11}^2 + 4E/F + \bar{d}_{12}^2}\right), \qquad \bar{p}_{12} = \frac{\bar{d}_{12}}{2}, \qquad \bar{p}_{22} = \bar{p}_{11} - \bar{d}_{11}. \tag{3.21}$$

Proof. When $\bar{d}_{11} = -\bar{d}_{22}$, the Lagrange conditions (3.14a) and (3.14b) imply that $(\lambda + 1)(\bar{p}_{11} + \bar{p}_{22}) = 0$. From this it follows that we have either $\lambda = -1$ or $\bar{p}_{11} = -\bar{p}_{22}$, or both. When $\bar{p}_{11} = -\bar{p}_{22}$, it follows from (3.14d) that $-\bar{p}_{11}^2 - \bar{p}_{12}^2 = E/F$. However, this situation cannot occur because $E/F > 0$. We conclude that $\lambda = -1$ must hold. The Lagrange conditions (3.14a) - (3.14d) now simplify to

$$\bar{p}_{11} - \bar{p}_{22} = \bar{d}_{11}, \quad 2\bar{p}_{12} = \bar{d}_{12}, \quad \bar{p}_{11}\bar{p}_{22} = \frac{E}{F} + \frac{\bar{d}_{12}^2}{4}.$$

Combining the first and third of these equations gives us

$$\bar{p}_{11}^2 - \bar{d}_{11}\bar{p}_{11} - \frac{E}{F} - \frac{\bar{d}_{12}^2}{4} = 0.$$

This polynomial has for any $\bar{d}_{11}$ and $\bar{d}_{12}$ always two real solutions, which are given by $\bar{p}_{11} = (\bar{d}_{11} \pm \sqrt{\bar{d}_{11}^2 + 4E/F + \bar{d}_{12}^2})/2$. However, if the minus sign holds we see that $\mathrm{tr}(\bar{P}) = -\sqrt{\bar{d}_{11}^2 + 4E/F + \bar{d}_{12}^2} < 0$. Thus, when $\bar{d}_{11} = -\bar{d}_{22}$, the global minimizer to Minimization Problem is given by (3.21). In Figure 3.3 these findings are illustrated. □

Now we only have to deal with the case $(\bar{d}_{11} = \bar{d}_{22} \wedge \bar{d}_{12} = 0)$.

Lemma 3.3. *Suppose $\bar{d}_{11} = \bar{d}_{22}$ and $\bar{d}_{12} = 0$. When $\bar{d}_{11} < 2\sqrt{E/F}$, the solution to Minimization Problem is the global minimum given by*

$$\bar{p}_{11} = \sqrt{E/F}, \quad \bar{p}_{12} = 0, \quad \bar{p}_{22} = \sqrt{E/F}, \tag{3.22}$$

otherwise, when $\bar{d}_{11} \geq 2\sqrt{E/F}$, the solution is a continuum of global minimizers given by

$$\bar{p}_{11} \in \left[\frac{\bar{d}_{11} - a}{2}, \frac{\bar{d}_{11} + a}{2}\right], \quad \bar{p}_{12} = \pm\sqrt{\bar{d}_{11}\bar{p}_{11} - \bar{p}_{11}^2 - \frac{E}{F}}, \quad \bar{p}_{22} = \bar{d}_{11} - \bar{p}_{11}, \tag{3.23}$$

where $a = \sqrt{\bar{d}_{11}^2 - 4E/F}$.

Proof. In the case that $\bar{d}_{11} = \bar{d}_{22}$ and $\bar{d}_{12} = 0$, Lagrange conditions (3.14a) and (3.14b) imply that $(1 - \lambda)(\bar{p}_{11} - \bar{p}_{22}) = 0$. From this it follows that we must either have $\lambda = 1$ or $\lambda \neq 1$ and then $\bar{p}_{11} = \bar{p}_{22}$. Let us first deal with the case $\lambda \neq 1$. When $\lambda \neq 1$, the Lagrange conditions (3.14c) and (3.14d) read

$$(1 - \lambda)\bar{p}_{12} = \bar{d}_{12} = 0, \quad \bar{p}_{11}^2 - \bar{p}_{12}^2 = E/F.$$

As $\lambda \neq 1$, the first of these equations implies that $\bar{p}_{12} = 0$. This fact combined with the second equation implies that $\bar{p}_{11} = \bar{p}_{22} = \pm\sqrt{E/F}$. The condition $\mathrm{tr}(\bar{P}) > 0$

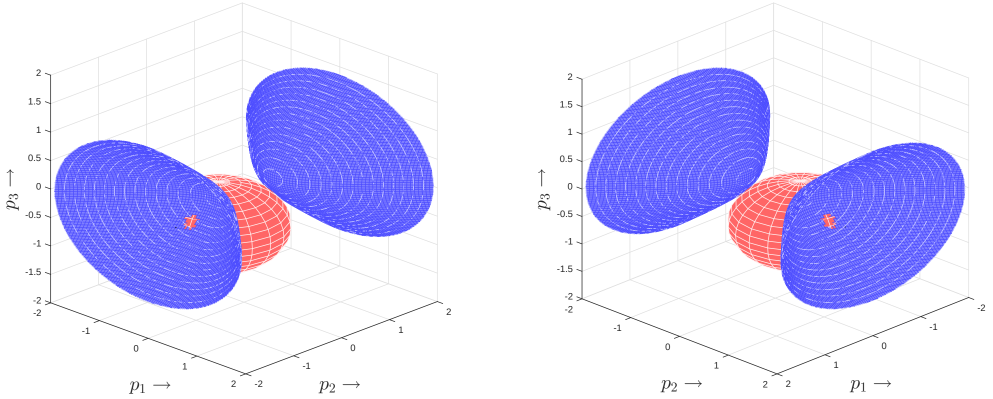

Figure 3.3: These figures corresponds to Lemma 3.2. The ellipsoid is centered around the same point $(d_1 = (\bar{d}_{11} - \bar{d}_{22})/\sqrt{2} = \sqrt{2}\bar{d}_{11}, d_2 = (\bar{d}_{11} + \bar{d}_{22})/\sqrt{2} = 0, d_3 = \bar{d}_{12})$ in both figures. This results in two local minima with the same function value for H. In the figure on the left we see the minimum on the hyperboloid sheet with $\operatorname{tr}(\bar{P}) < 0$ and in the figure on the right we see the minimum on the sheet with $\operatorname{tr}(\bar{P}) > 0$. These are the two minima that have been found in the proof of Lemma 3.2, the minimum in the figure on the left was discarded as it did not satisfy $\operatorname{tr}(\bar{P}) > 0$.

is only satisfied when the plus sign holds, hence we find one minimizer. This is the minimizer given by equation (3.22).

Now suppose that $\lambda = 1$. Lagrange condition (3.14a) implies $\bar{p}_{22} = \bar{d}_{11} - \bar{p}_{11}$ and from Lagrange condition (3.14d) we obtain $\bar{p}_{12} = \pm\sqrt{\bar{p}_{11}\bar{p}_{22} - E/F}$. Substituting the former expression in the latter gives us $\bar{p}_{12} = \pm\sqrt{\bar{d}_{11}\bar{p}_{11} - \bar{p}_{11}^2 - E/F}$, which is only real if $\bar{p}_{11}^2 - \bar{d}_{11}\bar{p}_{11} + E/F \leq 0$, that is, when $\bar{p}_{11} \in [(\bar{d}_{11} - a)/2, (\bar{d}_{11} + a)/2]$, where $a = \sqrt{\bar{d}_{11}^2 - 4E/F}$. This gives us the continuum of minimizers (3.23). However, $\bar{p}_{11}$ is only real if $\bar{d}_{11} \notin (-2\sqrt{E/F}, 2\sqrt{E/F})$. Moreover, because $\operatorname{tr}(\bar{P}) = \bar{d}_{11}$, we see that $\operatorname{tr}(\bar{P}) > 0$ is only satisfied when $\bar{d}_{11} > 0$. From this it follows that the continuum of minimizers can only be a solution to Minimization Problem when $\bar{d}_{11} \geq 2\sqrt{E/F}$. Thus, when $\bar{d}_{11} < 2\sqrt{E/F}$, the global minimizer is given by (3.22). To decide for $\bar{d}_{11} \geq 2\sqrt{E/F}$ whether the global minimizer is given by (3.22) or by an element of the continuum (3.23), we must compare the values of the function being minimized, i.e. H, for the local minimizers.

$H(\bar{P})$ has the same value for every element of the continuum of minimizers, because otherwise not all the elements of the continuum would have been local minima. For the value of $H(\bar{P})$ in the continuum we have

$$\begin{aligned} H_{\text{cont}} &= \frac{1}{2}\left\| \begin{pmatrix} \bar{d}_{11} - \bar{p}_{11} & -\sqrt{\bar{d}_{11}^2/4 - E/F} \\ -\sqrt{\bar{d}_{11}^2/4 - E/F} & \bar{p}_{11} \end{pmatrix} \right\|^2 \\ &= \frac{1}{2}\left(2\left(\bar{d}_{11}\bar{p}_{11} - \bar{p}_{11}^2 - \frac{E}{F}\right) + (\bar{d}_{11}^2 - \bar{p}_{11}^2)^2 + p_{11}^2\right) \\ &= \frac{\bar{d}_{11}^2}{2} - \frac{E}{F}. \end{aligned}$$

For the local minimizer at the extremum of the hyperboloid, given by (3.22), we have

$$H_{\text{ext}} = \frac{1}{2}\left\| \begin{pmatrix} \bar{d}_{11} - \sqrt{E/F} & 0 \\ 0 & \bar{d}_{11} - \sqrt{E/F} \end{pmatrix} \right\|^2 = d_{11}^2 - 2d_{11}\sqrt{\frac{E}{F}} + \frac{E}{F}.$$

This implies that $H_{\text{cont}} - H_{\text{ext}} = -\bar{d}_{11}^2/2 + 2d_{11}\sqrt{E/F} - 2E/F$. This polynomial in $\bar{d}_{11}$ has its maximal value in $d_{11} = 2\sqrt{E/F}$ where it equals 0, therefore it is negative for every $\bar{d}_{11} > 2\sqrt{E/F}$. This implies that if $\bar{d}_{11} \geq 2\sqrt{E/F}$, the solution to Minimization Problem is given by the continuum of minimizers (3.23). □

In Figure 3.4 examples of the results from Lemma 3.3 are geometrically shown. Recall that the extrema of the two sheets of the hyperboloid are located at

$$(p_1, p_2, p_3) = ((\bar{p}_{11} - \bar{p}_{22})/\sqrt{2}, (\bar{p}_{11} + \bar{p}_{22})/\sqrt{2}, \bar{p}_{12}) = \pm(0, \sqrt{2E/F}, 0).$$

Thus Lemma 3.3 implies that the global minimizer is at $(0, \sqrt{2E/F}, 0)$ if $\bar{d}_2 < 2\sqrt{2E/F}$, $\bar{d}_{11} = \bar{d}_{22}$ and $\bar{d}_{12} = 0$, i.e., when the center of the ellipsoid is located in $(0, p_2, 0)$, where $\bar{p}_2 = \sqrt{2}\bar{d}_{11} < 2\sqrt{2E/F}$. Or to put it in words, in the case that $\bar{d}_{11} = \bar{d}_{22}$ and $\bar{d}_{12} = 0$, if the distance from the center of the ellipsoid to the origin is less that two times the distance to the extremum of the sheet with $\text{tr}(\bar{P}) > 0$ of the hyperboloid, or if the center of the ellipsoid is situated beneath the plane $p_2 = \text{tr}(\bar{P})/\sqrt{2} = 0$, then the global minimizer is given by the extremum of the upper sheet of the hyperboloid. If $\bar{d}_{11} = \bar{d}_{22}$, $\bar{d}_{12} = 0$, the center of the ellipsoid is located above the plane $p_2 = \text{tr}(\bar{P})/\sqrt{2} = 0$ 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 the graph on the right in Figure 3.4. In the graph on the left in Figure 3.4, the center of the ellipsoid is farther away from the origin than the extremum of the sheet with $\text{tr}(\bar{P}) > 0$ of the hyperboloid, but it is closer 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.

Summarizing, we have proven the following theorem.

Theorem 3.4. *Minimization Problem, can be solved algebraically. In the general case, when* $(\bar{d}_{11} \neq \bar{d}_{22} \vee \bar{d}_{12} \neq 0)$ *and* $(\bar{d}_{11} \neq -\bar{d}_{22})$*, the solution to Minimization Problem is given by (3.17), with* λ *given by one of the four possibilities in (3.18). At least two of the* λ*'s in (3.18) are real. Explicit calculation of the function value* $H(\bar{P})$ *shows which of the real* λ*'s gives the global minimizer. In the case that* $(\bar{d}_{11} = -\bar{d}_{22})$*, there is a unique solution to Minimization Problem. This global minimizer is given by (3.21). Finally, in the case that* $(\bar{d}_{11} = \bar{d}_{22} \wedge \bar{d}_{12} = 0)$*, there is unique solution to Minimization Problem if* $\bar{d}_{11} < 2\sqrt{E/F}$ *and it is given by (3.22). If* $\bar{d}_{11} \geq 2\sqrt{E/F}$*, there is a whole continuum of solutions to Minimization Problem, which is given by (3.23).*

3.2. Minimization of J. In this section we focus on the last step of the least-squares method, i.e. (3.6c). We will minimize the functional J, defined in equation (3.5), for $\boldsymbol{m} \in \mathcal{M}$, while keeping $\boldsymbol{P}$ and $\boldsymbol{b}$ constant. Again we do this for an arbitrary coordinate system on $\mathcal{E}$ with basis vectors $\boldsymbol{e}_1$, $\boldsymbol{e}_2$ and corresponding metric $\boldsymbol{e} = e_{ij}\boldsymbol{e}^i \otimes \boldsymbol{e}^j$. We derive a coordinate-independent boundary value problem for the mapping $\boldsymbol{m}$ and subsequently derive from this the boundary value problem in Cartesian and polar coordinates. We will see that in the Cartesian case we end up with the same boundary value problem for $\boldsymbol{m}$ as derived in [12, p.142-144].

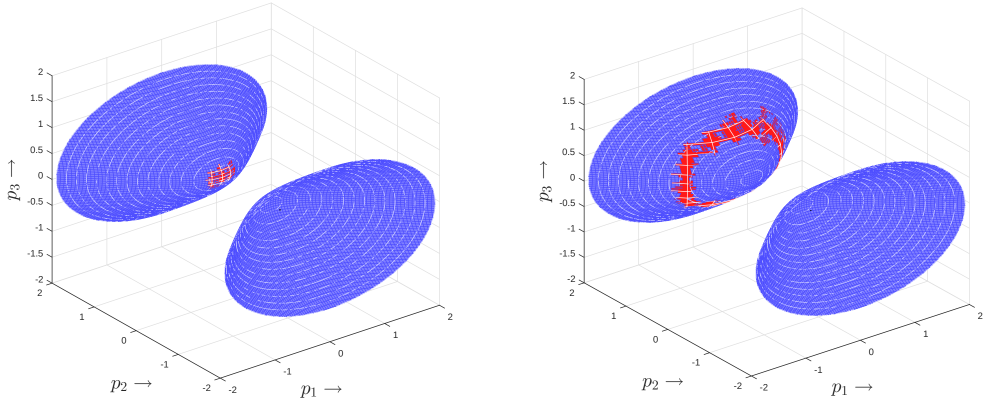

Figure 3.4: These figures corresponds to the minimizers in Lemma 3.3. The ellipsoid is located behind the hyperboloid. We see the sheet of the hyperboloid with $\operatorname{tr}(\bar{P}) > 0$ on the left. The ellipsoid is centered around a point $(d_1 = (\bar{d}_{11} - \bar{d}_{22})/\sqrt{2} = 0, d_2 = (\bar{d}_{11} + \bar{d}_{22})/\sqrt{2} = \sqrt{2}\bar{d}_{11}, d_3 = \bar{d}_{12} = 0)$. In the figure on the left $d_2 < 2\sqrt{2E/F}$ and we find the extremum of the hyperboloid as minimizer. In the figure on the right $d_2 \geq 2\sqrt{2E/F}$ and we find an elliptical continuum of minimizers.

3.2.1. 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 Gâteaux derivative of the J must be 0 in every direction, i.e.

$$\delta J(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b}; \boldsymbol{\eta}) := \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left(J(\boldsymbol{m} + \varepsilon\boldsymbol{\eta}, \boldsymbol{P}, \boldsymbol{b}) - J(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b}) \right) = 0,$$

for every direction $\boldsymbol{\eta} \in \mathcal{M}$. δJ_I and δJ_B are defined analogously. By linearity of the Gâteaux derivative we have

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

We first determine $\delta J_{\mathrm{I}}(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b}; \boldsymbol{\eta})$. By linearity of the covariant derivative we find

$$\delta J_{\mathrm{I}}(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b}; \boldsymbol{\eta}) = \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \int_{\mathcal{E}} \left(\|\varepsilon \nabla \hat{\boldsymbol{\eta}} + \nabla \hat{\boldsymbol{m}} - \boldsymbol{P}\|^2 - \|\nabla \hat{\boldsymbol{m}} - \boldsymbol{P}\|^2 \right) \sqrt{e} \boldsymbol{e}^1 \wedge \boldsymbol{e}^2.$$

We will now need the following convenient property of inner product on $\boldsymbol{T}^2_0(T_x\mathcal{E})$ as defined on page 10. Let $\boldsymbol{A}, \boldsymbol{B} \in \boldsymbol{T}^2_0(T_x\mathcal{E})$, then we have

$$\|\boldsymbol{A} + \boldsymbol{B}\|^2 = (A_{ij} + B_{ij})(A^{ij} + B^{ij}) = \|\boldsymbol{A}\|^2 + 2\boldsymbol{A} : \boldsymbol{B} + \|\boldsymbol{B}\|^2.$$

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

$$\begin{aligned} \delta J_{\mathrm{I}}(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b}; \boldsymbol{\eta}) &= \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \int_{\mathcal{E}} \left(\varepsilon^2 \|\nabla \hat{\boldsymbol{\eta}}\|^2 + 2\varepsilon \nabla \hat{\boldsymbol{\eta}} : (\nabla \hat{\boldsymbol{m}} - \boldsymbol{P}) \right) \sqrt{e} \boldsymbol{e}^1 \wedge \boldsymbol{e}^2 \\ &= \int_{\mathcal{E}} \nabla \hat{\boldsymbol{\eta}} : (\nabla \hat{\boldsymbol{m}} - \boldsymbol{P}) \, \sqrt{e} \boldsymbol{e}^1 \wedge \boldsymbol{e}^2. \end{aligned}$$

In the same fashion, using the fact that

$$\|\boldsymbol{m} + \varepsilon\boldsymbol{\eta} - \boldsymbol{b}\|^2 - \|\boldsymbol{m} - \boldsymbol{b}\|^2 = \varepsilon^2\|\boldsymbol{\eta}\|^2 + 2\varepsilon(\boldsymbol{\eta}, \boldsymbol{m} - \boldsymbol{b}), \tag{3.24}$$

we find the Gâteaux derivative of J_B to be

$$\delta J_\mathrm{B}(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b}; \boldsymbol{\eta}) = \oint_{\partial\mathcal{E}} (\boldsymbol{\eta}, \boldsymbol{m} - \boldsymbol{b})\ \mathrm{d}s.$$

Combining the results for J_I and J_B we find that

$$\forall\boldsymbol{\eta} \in \mathcal{M}: \quad \alpha \int_{\mathcal{E}} \nabla\hat{\boldsymbol{\eta}} : (\nabla\hat{\boldsymbol{m}} - \boldsymbol{P})\ \sqrt{e}\boldsymbol{e}^1 \wedge \boldsymbol{e}^2 + (1-\alpha) \oint_{\partial\mathcal{E}} (\boldsymbol{\eta}, \boldsymbol{m} - \boldsymbol{b})\ \mathrm{d}s = 0. \tag{3.25}$$

In order to proceed we will rewrite the integrands in terms of their components. For the first integral in (3.25) we have

$$\int_{\mathcal{E}} \nabla\hat{\boldsymbol{\eta}} : (\nabla\hat{\boldsymbol{m}} - \boldsymbol{P})\sqrt{e}\boldsymbol{e}^1 \wedge \boldsymbol{e}^2 = \int_{\mathcal{E}} D_j\eta_i(D^j m^i - p^{ij})\sqrt{e}\boldsymbol{e}^1 \wedge \boldsymbol{e}^2,$$

where $D_j\eta_i$ are the components of the covariant derivative of $\hat{\boldsymbol{\eta}}$ (Appendix A) and $D^j = e^{ij}D_i$. The product rule implies

$$D_j\eta_i(D^j m^i - p^{ij}) = D_j(\eta_i(D^j m^i - p^{ij})) - \eta_i D_j(D^j m^i - p^{ij}).$$

If we integrate the first term and apply *Green's theorem* [26, p.134] we find

$$\int_{\mathcal{E}} D_j(\eta_i(D^j m^i - p^{ij}))\sqrt{e}\boldsymbol{e}^1 \wedge \boldsymbol{e}^2 = \oint_{\partial\mathcal{E}} (D^j m^i - p^{ij})\eta_i n_j\ \mathrm{d}s,$$

where n_j are the covariant components of the outward unit normal vector on the boundary $\partial\mathcal{E}$ and the orientation on $\partial\mathcal{E}$ is the one induced by $\mathcal{E}$ [26, p.119]. It follows that

$$\begin{aligned}\int_{\mathcal{E}} \nabla\hat{\boldsymbol{\eta}} : (\nabla\hat{\boldsymbol{m}} - \boldsymbol{P})\sqrt{e}\boldsymbol{e}^1 \wedge \boldsymbol{e}^2 = &\oint_{\partial\mathcal{E}} (D^j m^i - p^{ij})\eta_i n_j\ \mathrm{d}s \\ &- \int_{\mathcal{E}} D_j(D^j m^i - p^{ij})\eta_i\sqrt{e}\boldsymbol{e}^1 \wedge \boldsymbol{e}^2.\end{aligned}$$

Combining this result with equation (3.25) we obtain

$$\begin{aligned}0 = &\oint_{\partial\mathcal{E}} \left[\alpha(D^j m^i - p^{ij})n_j + (1-\alpha)(m^i - b^i)\right]\eta_i\ \mathrm{d}s \\ &- \alpha\int_{\mathcal{E}} D_j(D^j m^i - p^{ij})\eta_i\sqrt{e}\boldsymbol{e}^1 \wedge \boldsymbol{e}^2,\end{aligned}$$

for all $\boldsymbol{\eta} \in \mathcal{M}$. Invoking the *Fundamental Lemma of Calculus of Variations* [29, p.185] we find from this the boundary value problem

$$D_j D^j m^i = D_j p^{ij} \qquad \text{in } \mathcal{E}, \tag{3.26a}$$

$$\alpha(D^j m^i)n_j + (1-\alpha)m^i = \alpha p^{ij}n_j + (1-\alpha)b^i \qquad \text{on } \partial\mathcal{E}. \tag{3.26b}$$

The solution of boundary value problem (3.26) will minimize J for constant $\boldsymbol{P}$ and $\boldsymbol{b}$. Note that (3.26) are vector equations. The term $D_j D^j m^i$ is the so-called *vector Laplacian* [28, p.91]. In Cartesian coordinates $D_j D^j m^i = \partial_j \partial^j m^i$, 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 $[D_j D^j m^i]_{i=1,2}$ depend both on both m^1 and m^2. This results for an arbitrary coordinate system in two coupled equations, while for a Cartesian coordinate system these two decouple. This will become more apparent when we derive the coordinate specific boundary value problem for Cartesian and polar coordinates.

3.2.2. The boundary value problem in specific coordinate systems. In Cartesian coordinates the partial differential equations in (3.26) decouple. Let us denote the standard Cartesian basis vectors by $\boldsymbol{e}_x$ and $\boldsymbol{e}_y$, define

$$\boldsymbol{p}_x = \begin{pmatrix} p^{xx} \\ p^{xy} \end{pmatrix} = \begin{pmatrix} p^{11} \\ p^{12} \end{pmatrix} \quad \text{and} \quad \boldsymbol{p}_y = \begin{pmatrix} p^{yx} \\ p^{yy} \end{pmatrix} = \begin{pmatrix} p^{12} \\ p^{22} \end{pmatrix},$$

and write $\boldsymbol{m} = m^x \boldsymbol{e}_x + m^y \boldsymbol{e}_y$. With the use of this definition we can rewrite $(D_j p^{ij})_{i=1}$ as $\operatorname{div} \boldsymbol{p}_x$ and $(D_j p^{ij})_{i=2}$ as $\operatorname{div} \boldsymbol{p}_y$. From this we see that in Cartesian coordinates (3.26) reduces to the decoupled set of equations

$$\begin{aligned} \Delta m^x &= \operatorname{div} \boldsymbol{p}_x && \text{in } \mathcal{E}, \\ \alpha(\nabla m^x, \boldsymbol{n}) + (1-\alpha) m^x &= \alpha(\boldsymbol{p}_x, \boldsymbol{n}) + (1-\alpha) b^x && \text{on } \partial\mathcal{E}, \end{aligned} \tag{3.27a}$$

$$\begin{aligned} \Delta m^y &= \operatorname{div} \boldsymbol{p}_y && \text{in } \mathcal{E}, \\ \alpha(\nabla m^y, \boldsymbol{n}) + (1-\alpha) m^y &= \alpha(\boldsymbol{p}_y, \boldsymbol{n}) + (1-\alpha) b^y && \text{on } \partial\mathcal{E}. \end{aligned} \tag{3.27b}$$

The boundary value problems (3.27a) and (3.27b) are exactly the same as in [12, p.143].

In polar coordinates the equations do not decouple. Notice that the coordinate specific boundary value problem that we deduce from (3.26) does depend on the choice of basis for polar coordinates, because (3.26) is a vector equation. Thus, we find for an anholonomic basis (Appendix A) different expressions than for a holonomic basis.

To derive the boundary value problem in polar coordinates, let us first write out the components of the covariant derivatives appearing in (3.26) in terms of Christoffel symbols and derivatives. We start out with the vector Laplacian. By the definition of D_j (Appendix A) it follows that $D_j D^j m^i = e^{jk}(\nabla_{\boldsymbol{e}_j}(D_k m^i) - \Gamma^l_{kj} D_l m^i + \Gamma^i_{lj} D_k m^l)$ and $D_k m^i = \nabla_{\boldsymbol{e}_k} m^i + \Gamma^i_{lk} m^l$, hence

$$\begin{aligned} D_j D^j m^i = e^{jk}\big(&\nabla_{\boldsymbol{e}_j}\nabla_{\boldsymbol{e}_k} m^i + \nabla_{\boldsymbol{e}_j}(\Gamma^i_{lk}) m^l + \Gamma^i_{lk}\nabla_{\boldsymbol{e}_j} m^l - \Gamma^l_{kj}\nabla_{\boldsymbol{e}_l} m^i \\ &- \Gamma^l_{kj}\Gamma^i_{sl} m^s + \Gamma^i_{lj}\nabla_{\boldsymbol{e}_k} m^l + \Gamma^i_{lj}\Gamma^l_{sk} m^s\big). \end{aligned} \tag{3.28}$$

Doing the same derivation for the divergence of $\boldsymbol{P}$ we obtain[3]

$$D_j p^{ij} = \nabla_{\boldsymbol{e}_j} p^{ij} + \Gamma^i_{lj} p^{lj} + \Gamma^j_{lj} p^{il}. \tag{3.29}$$

Similarly, we find for the normal derivative of $\boldsymbol{m}$ in equation (3.26b)

$$(D^j m^i) n_j = e^{jk}(D_k m^i) n_j = e^{jk}\big(\nabla_{\boldsymbol{e}_k} m^i + \Gamma^i_{lk} m^l\big) n_j. \tag{3.30}$$

[3] Note, that due to the symmetry of $\boldsymbol{P}$ it is clear what we 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.

We use (3.28) - (3.30) to determine the boundary value problem (3.26) in polar coordinates. We first consider the anholonomic basis, because this is the basis we used in the implementation. The Christoffel symbols in the anholonomic basis are $\Gamma^r_{\theta\theta} = -r^{-1}$ and $\Gamma^\theta_{r\theta} = r^{-1}$ [18, p.218]. After doing the tedious calculations of determining the coordinate system specific expressions of the various terms in (3.28) we find

$$\begin{aligned}
(D_j D^j m^i)_{i=r} &= \frac{\partial^2 m^r}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 m^r}{\partial \theta^2} - \frac{2}{r^2}\frac{\partial m^\theta}{\partial \theta} + \frac{1}{r}\frac{\partial m^r}{\partial r} - \frac{m^r}{r^2},\\
(D_j D^j m^i)_{i=\theta} &= \frac{\partial^2 m^\theta}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 m^\theta}{\partial \theta^2} + \frac{1}{r}\frac{\partial m^\theta}{\partial r} + \frac{2}{r^2}\frac{\partial m^r}{\partial \theta} - \frac{m^\theta}{r^2}.
\end{aligned}$$

In the same way we calculate the expressions for the divergence of $\boldsymbol{P}$ and obtain

$$\begin{aligned}
(D_j p^{ij})_{i=r} &= \frac{\partial p^{rr}}{\partial r} + \frac{1}{r}\frac{\partial p^{r\theta}}{\partial \theta} + \frac{p^{rr} - p^{\theta\theta}}{r},\\
(D_j p^{ij})_{i=\theta} &= \frac{\partial p^{r\theta}}{\partial r} + \frac{1}{r}\frac{\partial p^{\theta\theta}}{\partial \theta} + \frac{2p^{r\theta}}{r}.
\end{aligned}$$

Finally, we determine the expression for the normal derivative of $\boldsymbol{m}$ from (3.30) and find

$$\begin{aligned}
((D^j m^i) n_j)_{i=r} &= \frac{\partial m^r}{\partial r} n^r + \frac{\partial m^r}{\partial \theta}\frac{n^\theta}{r} - \frac{m^\theta n^\theta}{r},\\
((D^j m^i) n_j)_{i=\theta} &= \frac{\partial m^r}{\partial r} n^r + \frac{\partial m^r}{\partial \theta}\frac{n^\theta}{r} + \frac{m^r n^\theta}{r}.
\end{aligned}$$

We define

$$\boldsymbol{p}_r = \begin{pmatrix} p^{rr} \\ p^{r\theta} \end{pmatrix} \quad \text{and} \quad \boldsymbol{p}_\theta = \begin{pmatrix} p^{r\theta} \\ p^{\theta\theta} \end{pmatrix}, \tag{3.31}$$

and collect all the different terms and find that the polar coordinate with anholonomic basis variant of (3.26) is given by

$$\begin{aligned}
\Delta m^r - \frac{1}{r^2}\left(m^r + 2\frac{\partial m^\theta}{\partial \theta}\right) &= \operatorname{div} \boldsymbol{p}_r - \frac{p^{\theta\theta}}{r} && \text{in } \mathcal{E},\\
\alpha(\nabla m^r, \boldsymbol{n}) - \alpha\frac{m^\theta n^\theta}{r} + (1-\alpha)m^r &= \alpha(\boldsymbol{p}_r, \boldsymbol{n}) + (1-\alpha)b^r && \text{on } \partial\mathcal{E},
\end{aligned} \tag{3.32}$$

and

$$\begin{aligned}
\Delta m^\theta - \frac{1}{r^2}\left(m^\theta - 2\frac{\partial m^r}{\partial \theta}\right) &= \operatorname{div} \boldsymbol{p}_\theta + \frac{p^{r\theta}}{r} && \text{in } \mathcal{E},\\
\alpha(\nabla m^\theta, \boldsymbol{n}) + \alpha\frac{m^r n^\theta}{r} + (1-\alpha)m^\theta &= \alpha(\boldsymbol{p}_\theta, \boldsymbol{n}) + (1-\alpha)b^\theta && \text{on } \partial\mathcal{E},
\end{aligned} \tag{3.33}$$

where Δ, div and ∇ are the familiar Laplace, divergence and gradient operator in polar coordinates with anholonomic basis [24, p.542-543]. The equations (3.32) and (3.33) are coupled.

In the implementation of the GLS method we solve this boundary value problem by using a standard second order central finite difference method. This provides us with a linear system of the form $A\boldsymbol{x} = \boldsymbol{b}$, where A does not change between time-steps.

We use this fact by determining an LU-decomposition, before the first iteration and we use this to solve linear system for each time-step. To deal with the fact that, in for example polar coordinates, (3.32) and (3.33) are two coupled equations, we will solve them by iterating between the two. Starting with (3.32) we keep m^θ fixed and solve for m^r. Next we keep m^r fixed and solve (3.33) for m^θ. In this way we iterate between (3.32) and (3.33). We stop this iterative procedure when $J^{(n+1,i)} < cJ^{(n)}$, where n is the outer iteration count of (3.6) and i is the inner iteration count, or, when the number of inner iteration is larger than a specified value d, i.e., $i > d$. The optimal choice for these values are problem specific and if the number of inner iterations is increased by demanding more precision in (3.6c), the outer iterative procedure might converge faster. However, demanding far more precision in (3.6c) than is achieved by the outer iterative procedure up to that point is a waste of time. A maximum on the number of iterations is introduced to make sure that the method does not stall when $J^{(n+1,i)} < cJ^{(n)}$ is a too severe requirement. This will come in to play in the final iterations. We found that only few inner iterations in (3.6c) are sufficient, because the mapping $\boldsymbol{m}^n$ provides a very good initial guess for (3.6c). In practice we took $c = 0.9$ and $d = 5$ and these values seem to be a good choice for the problems tested so far.

In the next section we will present the results for polar coordinates with anholonomic basis. However, before proceeding to the next section, we first have to clear up how to determine function $u : \mathcal{E} \to (0, \infty)$ once we have found a mapping $\boldsymbol{m} \in T\mathcal{E}_{C^2}$ that is a solution to Transport BVP.

3.3. Determining the reflector surface from the mapping. To determine the reflector surface from the mapping $\boldsymbol{m}$ we generalize the derivation given by C. Prins et al. [12, p.144] to arbitrary coordinate systems. We earlier remarked that $\boldsymbol{m}$ equals the gradient of u if and only if $\nabla\hat{\boldsymbol{m}}$ is symmetric. However, in the GLS method J_I is minimized in the L^2-norm and hence $\nabla\hat{\boldsymbol{m}}$ is not exactly symmetric. We can therefore only search for a function $u : \mathcal{E} \to (0, \infty)$ with gradient equal to $\boldsymbol{m}$ in an L^2-sense, hence we will search for u that minimizes

$$I(u) := \int_{\mathcal{E}} \|\nabla u - \boldsymbol{m}\|^2 \sqrt{e}\boldsymbol{e}^1 \wedge \boldsymbol{e}^2.$$

After a derivation very similar to the one by which we arrived at boundary value problem (3.26), which we leave out for brevity, we obtain the Poisson problem

$$\begin{aligned} D_i D^i u &= D_i m^i && \text{in } \mathcal{E}, \\ D^i(u) n_i &= m^i n_i && \text{on } \partial\mathcal{E}, \end{aligned}$$

where n_i are again the covariant components of the normal covector. In Cartesian coordinates this problem is the one previously given in [13]. For polar coordinates with an anholonomic basis it is given by

$$\begin{aligned} \Delta u &= \operatorname{div} \boldsymbol{m} && \text{in } \mathcal{E}, \\ (\nabla u, \boldsymbol{n}) &= (\boldsymbol{m}, \boldsymbol{n}) && \text{on } \partial\mathcal{E}, \end{aligned}$$

where Δ, div and ∇ are again the familiar polar coordinate differentiation operators. It is this problem that we solve to find the reflector surface for the problems presented in next section. We then discretize this Poisson problem by using second order central differences, giving us a linear system. The solution of this linear system gives us the reflector surface.

4. Numerical results. We show the performance of the least-squares method in polar coordinates on the basis of two test cases. In the first test case we compare the method in polar coordinates with the method in Cartesian coordinates as presented in [13] and in the second test case we investigate the performance of the method in polar coordinates for a complex problem with a discontinuous desired light output.

For both test cases we take for the source pair $(\mathcal{E}, E)$ a unit disk with uniform emittance. We choose this source, because it frequently occurs in lighting systems and is the natural environment to apply the polar-coordinate least-squares method. For the first test case we take as the target $\mathcal{F}_1 = [-1, 1] \times [-1, 1]$ with a uniform intensity function F_1. For the second test case we have determined the pair $(\mathcal{F}_2, F_2)$ such that an intensity pattern corresponding to the sketch by M. C. Escher (Figure 1.1) is projected on a screen in the far-field. We take the projection screen at a distance 100 times the radius of the source and we take $(\mathcal{F}_2, F_2)$ such that width and height of the projection are 4.3 times the radius of the source. We will normalize F_1 and F_2 such that (2.7) holds.

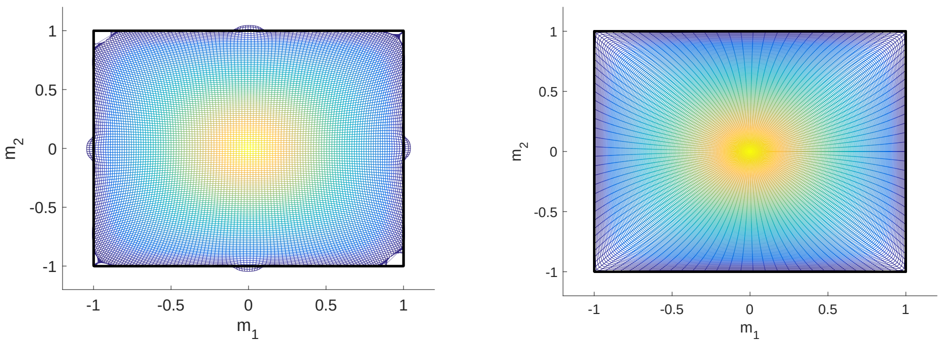

Figure 4.1: The resulting mapping is shown: on the left for Cartesian coordinates and on the right for polar coordinates. For both a 500×500 grid is used and $\alpha = 0.2$. We see the grid as it is mapped on $\mathcal{F}_1$. Grid points that initially had the same distance to the center of $\mathcal{E}$ have the same colour. Bright yellow corresponds to points in the center of $\mathcal{E}$ and dark blue corresponds to points on $\partial\mathcal{E}$.

4.1. From a circle to a square. In Figure 4.1 we see that near the boundary $\partial\mathcal{F}_1$ the method in Cartesian coordinates has great difficulties. This results from the implementation where as actual source the smallest bounding box of $\mathcal{E}$ is used with emittence zero in the points outside $\mathcal{E}$; see [13] for details. In the polar coordinate method the grid perfectly aligns with $\partial\mathcal{E}$. In Figure 4.1 it can be seen that now all the difficulties at $\partial\mathcal{F}_1$ are resolved.

Figure 4.2 shows the convergence history of the method for different values of α. The value of α determines approximately the ratio between J_{I} and J_{B}. In general, for values of α close to 1 the method finds a reflector that closely satisfies (2.8), but might possibly be less accurate concerning the boundary condition of the Inverse reflector problem, and vice versa for α close to 0. For smoother $(\mathcal{E}, E)$ and $(\mathcal{F}, F)$ the solution found by the method seems less dependable on the choice of α. However, in cases where for example $\partial\mathcal{F}$ is not differentiable, as in the current test case, there seems to be a pay-off. For such problems the boundary condition and (2.8) cannot be satisfied

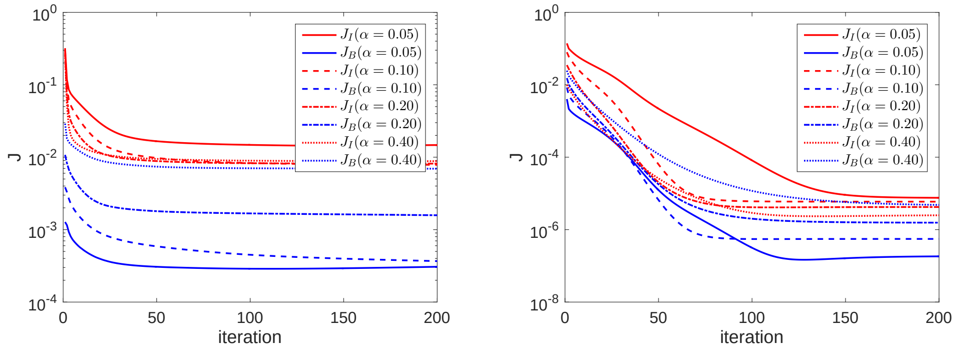

Figure 4.2: For the first test case the interior error J_{I} and the boundary error J_{B} are shown as function of the number of iterations for different α's, on the left for the method in Cartesian coordinates and on the right for the method in polar coordinates. In both cases a 100×100 grid was used.

exactly and simultaneously. In order to clarify this alleged pay-off further study has to be done. Nonetheless, the freedom in α provides the user of the GLS method with an opportunity to select the most fitting pay-off between interior and boundary for the specific application at hand. Moreover, it can be seen that due to better handling of the boundary $\partial\mathcal{E}$ the gap between J_{I} and J_{B} is far smaller for the method in polar coordinates for all choice of α.

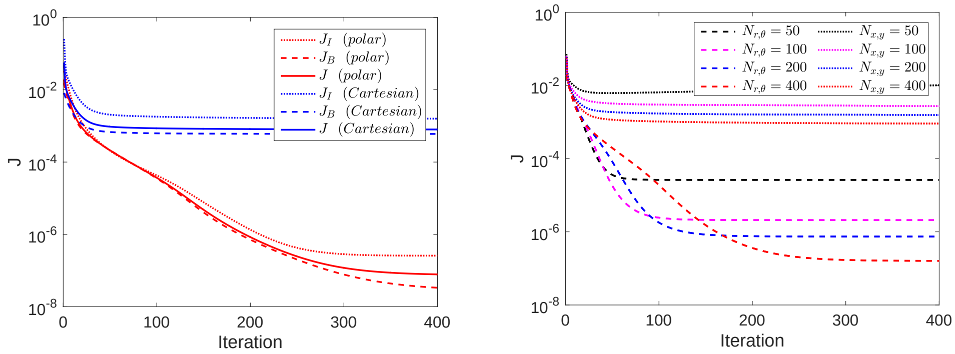

Figure 4.3: The convergence history for Cartesian and polar coordinates is compared for the first test case. In the left plot a 500×500 grid is used and the different error components are shown. An improvement by a factor 10^4 is observed when using polar instead of Cartesian coordinates. On the right J is shown for different grids. For both plots we used $\alpha = 0.2$

Figure 4.3 shows that the method in polar coordinates significantly outperforms its Cartesian counterpart. In the figure on the left it is seen that for a 500×500 grid the convergence of the Cartesian method stalls after approximately 75 iterations. The convergence of the polar method proceeds for another 300 iterations and this eventually leads to a value for J_{I} that is 10^4 times as small is the J_{I} found with the

Cartesian method. In the figure on the right it is seen that the use of increasingly finer grids has more effect for the polar method. However, even for the polar method final value for J_{I} seems not to convergence to zero when ever finer grids are used.

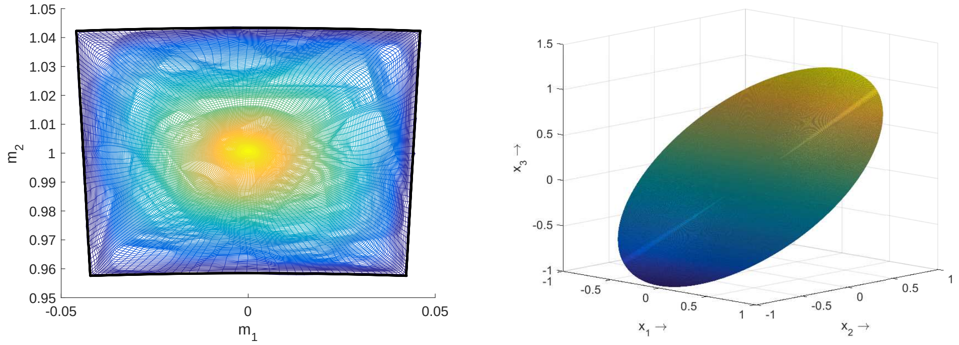

Figure 4.4: The mapping and corresponding reflector are shown for the second test case.

4.2. From a circle to an Escher-sketch in the far-field. Figure 4.4 shows the results of applying the method in polar coordinates to the second test case. For this very demanding test case a 1400 × 1400 grid was used. The elliptic shaped reflector is globally close to flat but locally contains great detail. Subsequently the reflector was simulated by using ray trace methods [4]. The result can be seen in Figure 4.5. The ray trace results closely resemble the original picture, although there is some decrease in contrast. In the algorithm the reflector surface is in the class of twice continuously differentiable functions. This naturally results in smoothing of the, often discontinuous, intensity in the original picture. Nonetheless the resolution obtained is high enough to carry over all the minute details of the original picture.

5. Summary and final remarks. In Section 2 we derived the Monge-Ampère equation, describing the reflector surface, for an arbitrary coordinate system. We found the map r, which maps a point on the source $\mathcal{E}$, from which a light ray leaves, to the direction of reflection of this ray, to be the composition of the gradient of the reflector surface, ∇u, and the inverse of the stereographic projection, s. Furthermore, we formulated the Inverse reflector problem in terms of the source and emittance $(\mathcal{E}, E)$ and the gradient set and intensity function $(\mathcal{F}, F)$.

In Section 3 we introduced the GLS method by generalizing the LS method, earlier introduced in [13], to general coordinate systems. Moreover, we gave a new geometric interpretation to the minimization problem for the functional J_{I} and found that the minimization problem for the total functional J consists of two Poisson problems which, contrary to the Cartesian case, are coupled in general coordinate systems.

In Section 4 we showed that the GLS method has far wider applicability than the LS method. We showed that for a disk-shaped source the GLS in polar coordinates gave a significant improvement over the LS method, decreasing the error by four orders of magnitude. It was seen that for problems with non-smooth desired output intensity the final ratio between J_I and J_B depends on the value of α in equation (3.5). Further research and literature study into this relation should be done. It would be for example important to know for which combination of source pair $(\mathcal{E}, E)$

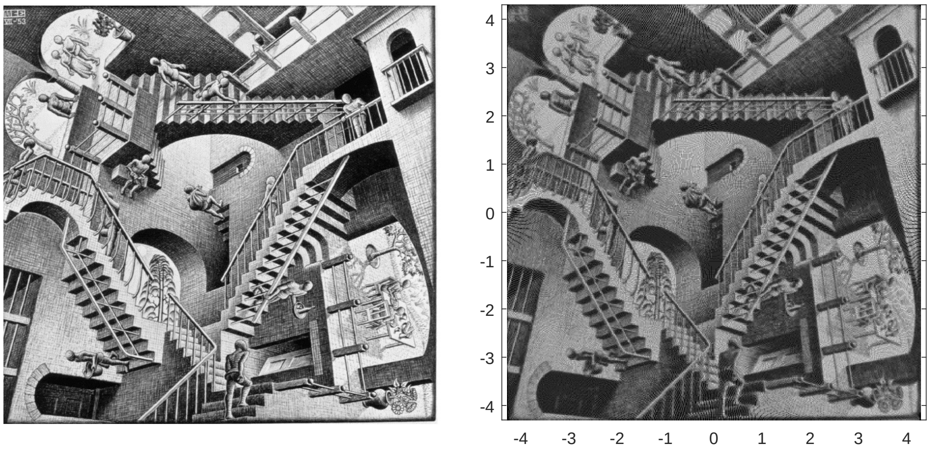

Figure 4.5: The image projected on the screen by the reflector of Figure 4.4 is determined by ray tracing about 4 million rays, with uniform fixed spacing, that leave $\mathcal{E}$. The original is shown on the left and the ray trace result is shown on the right.

and target pair $(\mathcal{F}, F)$ the solution of the method depends on α and to quantify to what extend.

Lastly, the method was applied to a very challenging problem concerning a detailed piece of art and still the method obtained a high resolution preserving the details of the original picture. This gives confidence in the wide applicability of the method in an industrial context.

Appendix A. Tensor Calculus. In this appendix we introduce the concepts from tensor calculus used in this paper. We will be very brief but make precise references to the literature where detailed explanations are given.

Let us consider $\mathcal{U} \subset \mathbb{R}^n$ with a coordinate system $\boldsymbol{x} = (x^i)_{i=1}^n$ [26, p.111] and a local basis $(\boldsymbol{e}_i(\boldsymbol{x}))_{i=1}^n$ at each point of $\boldsymbol{x} \in \mathcal{U}$. The vector space spanned by $(\boldsymbol{e}_i(\boldsymbol{x}))_{i=1}^n$ is called the tangent space to $\mathcal{U}$ at $\boldsymbol{x}$ and denoted by $T_{\boldsymbol{x}}\mathcal{U}$ [26, p.115]. The union $\cup_{\boldsymbol{x}\in\mathcal{U}} T_{\boldsymbol{x}}\mathcal{U}$ is called the tangent bundle to $\mathcal{U}$ and denoted by $T\mathcal{U}$. The linear mappings on $T_{\boldsymbol{x}}\mathcal{U}$ form its dual space $T^*_{\boldsymbol{x}}\mathcal{U}$ and the elements of this are called covectors. Similarly, vectors can be seen as linear mappings on $T^*_{\boldsymbol{x}}\mathcal{U}$. Let $(\boldsymbol{e}^i)_{i=1}^n$ denote the dual basis to $(\boldsymbol{e}_i)_{i=1}^n$, i.e. , $\boldsymbol{e}^i(\boldsymbol{e}_j) = \delta^i_j$. Taking the tensor product between vectors and covectors we can construct general linear mappings on products of $T_{\boldsymbol{x}}\mathcal{U}$ and $T^*_{\boldsymbol{x}}\mathcal{U}$ called tensors [26, p.75]. A general linear map from $(T_{\boldsymbol{x}}\mathcal{U})^k \times (T^*_{\boldsymbol{x}}\mathcal{U})^l$ to $\mathbb{R}$ (order of these $k+l$ spaces may be different and does matter) is called a tensor of contravariant rank k and covariant rank l. The space of such tensors we denote by $\boldsymbol{T}^k_l(T_{\boldsymbol{x}}\mathcal{U})$ [23, p.20]. From every pair of tensors we can define a third one through the tensor product. Suppose we have two covectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in T^*_{\boldsymbol{x}}\mathcal{U}$ then we can define the tensor product $\boldsymbol{v}_1 \otimes \boldsymbol{v}_2$ of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ by $\boldsymbol{v}_1 \otimes \boldsymbol{v}_2(\boldsymbol{w}_1, \boldsymbol{w}_2) = \boldsymbol{v}_1(\boldsymbol{w}_1)\boldsymbol{v}_2(\boldsymbol{w}_2)$ for every $\boldsymbol{w}_1, \boldsymbol{w}_2 \in T_{\boldsymbol{x}}\mathcal{U}$ [26, p.75]. The tensor product is defined analogously for arbitrary pairs of tensors. An example of a tensor of covariant rank 2 is the metric. The tensors $\boldsymbol{e}_{i_1} \otimes \cdots \otimes \boldsymbol{e}_{i_k} \otimes \boldsymbol{e}^{j_1} \otimes \cdots \otimes \boldsymbol{e}^{j_l}$ with $1 \leq i_1, \ldots, i_k, j_1, \ldots, j_l \leq n$ form a basis for the space $\boldsymbol{T}^k_l(T_{\boldsymbol{x}}\mathcal{U})$ [23, p.21] and hence for every $T \in \boldsymbol{T}^k_l(T_{\boldsymbol{x}}\mathcal{U})$ there exist coefficients $T^{i_1\dots i_k}_{j_1\dots j_l}$ such that [23, p.21]

$$\boldsymbol{T} = T^{i_1\dots i_k}_{j_1\dots j_l}\boldsymbol{e}_{i_1} \otimes \cdots \otimes \boldsymbol{e}_{i_k} \otimes \boldsymbol{e}^{j_1} \otimes \cdots \otimes \boldsymbol{e}^{j_l},$$

where the Einstein summation rule applies. The coefficients $T^{i_1\dots i_k}_{j_1\dots j_l}$ are called the components of $\boldsymbol{T}$. A tensor $\boldsymbol{T} \in \boldsymbol{T}^k(T_{\boldsymbol{x}}\mathcal{U})$ is called alternating [26, p.78] when $\boldsymbol{T}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_i, \dots, \boldsymbol{v}_j, \dots, \boldsymbol{v}_k) = -\boldsymbol{T}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_j, \dots, \boldsymbol{v}_i, \dots, \boldsymbol{v}_k)$ for all $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ in $T_{\boldsymbol{x}}\mathcal{U}$. Every $\boldsymbol{T} \in \boldsymbol{T}^k(T_{\boldsymbol{x}}\mathcal{U})$ can be made alternating by the operation [26, p.78]

$$\mathrm{Alt}(\boldsymbol{T})(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k) = \frac{1}{k!}\sum_{\sigma \in S_k} \mathrm{sgn}(\sigma)\boldsymbol{T}(\boldsymbol{v}_{\sigma(1)}, \dots, \boldsymbol{v}_{\sigma(k)}),$$

where S_k is the set of all permutations of $1, \dots, k$. The space of alternating $\boldsymbol{T} \in \boldsymbol{T}^k(T_{\boldsymbol{x}}\mathcal{U})$ is denoted by $\boldsymbol{\Lambda}^k(T_{\boldsymbol{x}}\mathcal{U})$. $\mathrm{Alt}(\boldsymbol{T}) \in \boldsymbol{\Lambda}^k(T_{\boldsymbol{x}}\mathcal{U})$ for every $\boldsymbol{T} \in \boldsymbol{T}^k(T_{\boldsymbol{x}}\mathcal{U})$ and $\mathrm{Alt}(\boldsymbol{T}) = \boldsymbol{T}$ for every $\boldsymbol{T} \in \boldsymbol{\Lambda}^k(T_{\boldsymbol{x}}\mathcal{U})$ [26, p.78]. For $\boldsymbol{S} \in \boldsymbol{\Lambda}^k(T_{\boldsymbol{x}}\mathcal{U})$ and $\boldsymbol{T} \in \boldsymbol{\Lambda}^l(T_{\boldsymbol{x}}\mathcal{U})$ their tensor product $\boldsymbol{S} \otimes \boldsymbol{T}$ is usually not alternating, but their wedge product, defined by [26, p.79]

$$\boldsymbol{S} \wedge \boldsymbol{T} = \frac{(k+l)!}{k!l!}\mathrm{Alt}(\boldsymbol{S} \otimes \boldsymbol{T})$$

is alternating, i.e., $\boldsymbol{S} \wedge \boldsymbol{T} \in \boldsymbol{\Lambda}^{k+l}(T_{\boldsymbol{x}}\mathcal{U})$. A basis for $\boldsymbol{\Lambda}^k(T_{\boldsymbol{x}}\mathcal{U})$ is formed by the set of all $\boldsymbol{e}^{i_1} \wedge \cdots \wedge \boldsymbol{e}^{i_k}$ with $1 \leq i_1 < i_2 < \cdots < i_k \leq n$ [26, p.81]. An example of an n-form on $\mathcal{U} \subset \mathbb{R}^n$ is the volume form [28, p.105] given by $\boldsymbol{e}^1 \wedge \cdots \wedge \boldsymbol{e}^n$ for an orthonormal (dual) basis $\boldsymbol{e}^1, \dots, \boldsymbol{e}^n$. In classical notation this volume form is often denoted by $\mathrm{d}V$ or $\mathrm{d}A$, depending on the dimension.

A vector field on $\mathcal{U}$ is a function that assigns to each $\boldsymbol{x} \in \mathcal{U}$ a vector in $T_{\boldsymbol{x}}\mathcal{U}$ and k-form on $\mathcal{U}$ is a function that assigns to each $\boldsymbol{x} \in \mathcal{U}$ a tensor in $\boldsymbol{\Lambda}^k(T_{\boldsymbol{x}}\mathcal{U})$. Similarly we can define arbitrary tensor fields. A tensor field is differentiable if its components are differentiable. An important tensor field on $\mathcal{U}$ is the metric $\boldsymbol{e} = e_{ij}\boldsymbol{e}^i \otimes \boldsymbol{e}^j$ whose components are defined as $e_{ij} = (\boldsymbol{e}_i, \boldsymbol{e}_j)$. The symmetry of the inner product implies that e_{ij} is also symmetric. The components of the inverse of the matrix (e_{ij}) are denoted by e^{ij}, i.e., we have $e^{ij}e_{jk} = \delta^i_k$. Furthermore, we write $e = \det(e_{ij})$. The metric for example allows us to express the volume form in any basis. It is given by $\sqrt{e}\boldsymbol{e}^1 \wedge \cdots \wedge \boldsymbol{e}^n$ [28, p.105].

In this paper we use two notions of derivatives. We first introduce the covariant derivative. Suppose $v \in C^1(\mathcal{U})$, then the directional derivative of v in the direction $\boldsymbol{e}_i$ is denoted by $\nabla_{\boldsymbol{e}_i}v$. If the basis vectors $\boldsymbol{e}_i$ are the tangents to the coordinate lines of x^i, i.e., the lines of varying x^i while keeping x^j $(J \neq i)$ constant, then the basis is called a coordinate basis or a holonomic basis and we have $\nabla_{\boldsymbol{e}_i}v = \partial v/\partial x^i$. However, in general a basis is anholonomic and this does not apply. The differential of $v \in C^1(\mathcal{U})$ is given by $\mathrm{d}v = (\nabla_{\boldsymbol{e}_i}v)\boldsymbol{e}^i$, which is a 1-form, and the gradient of v is the corresponding vector $\nabla v = e^{ij}(\nabla_{\boldsymbol{e}_i}v)\boldsymbol{e}_j$. If we take the directional derivative of a vector field $\boldsymbol{v} = v^j\boldsymbol{e}_j$ on $\mathcal{U}$ we get, applying the product rule,

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

The derivatives $\nabla_{\boldsymbol{e}_i}(\boldsymbol{e}_j)$ are nonzero in a general coordinate system, however, as they are directional derivatives of vectors in $\mathcal{U}$, they are itself vectors in $\mathcal{U}$. As such they can be written as linear combination of basis vectors as $\nabla_{\boldsymbol{e}_i}(\boldsymbol{e}_j) = \Gamma^k_{ji}\boldsymbol{e}_k$. Using this we have

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

The coefficients Γ^k_{ji} are called the Christoffel symbols. We can equivalently write $\nabla_{\boldsymbol{e}_i}\boldsymbol{v} = D_i(v^k)\boldsymbol{e}_k$, where $D_i(v^k) = \nabla_{\boldsymbol{e}_i}(v^k) + v^j\Gamma^k_{ji}$ are the components of the covariant derivative of $\boldsymbol{v}$ with respect to the i-th basis vector. Extending this reasoning to general tensors one finds that for $\boldsymbol{T} \in \boldsymbol{T}^k_l(T_{\boldsymbol{x}}\mathcal{U})$ the covariant derivative in the direction $\boldsymbol{e}_i$ is given by $\nabla_{\boldsymbol{e}_i}\boldsymbol{T} = D_i\left(T^{i_1\dots i_k}_{j_1\dots j_l}\right)\boldsymbol{e}_{i_1}\otimes\cdots\otimes\boldsymbol{e}_{i_k}\otimes\boldsymbol{e}^{j_1}\otimes\cdots\otimes\boldsymbol{e}^{j_l}$, with

$$\begin{aligned} D_i\left(T^{i_1\dots i_k}_{j_1\dots j_l}\right) = \nabla_{\boldsymbol{e}_i}\left(T^{i_1\dots i_k}_{j_1\dots j_l}\right) + T^{k\dots i_k}_{j_1\dots j_l}\Gamma^{i_1}_{ki} + \cdots + T^{i_1\dots k}_{j_1\dots j_l}\Gamma^{i_k}_{ki} \\ - T^{i_1\dots i_k}_{k\dots j_l}\Gamma^k_{j_1 i} - \cdots - T^{i_1\dots i_k}_{j_1\dots k}\Gamma^k_{j_l i}. \end{aligned} \tag{A.1}$$

In (A.1) we get for every contravariant index a Christoffel symbol with a plus sign and for every covariant index a Christoffel symbol with a minus sign [18, p.209].

The exterior derivative is an operation that takes a k-form into a $(k+1)$-form. Suppose we have a holonomic basis $(\boldsymbol{e}_i)^n_{i=1}$ and a differentiable k-form $\boldsymbol{T} = T_{i_1\dots i_k}\boldsymbol{e}^{i_1}\wedge\cdots\wedge\boldsymbol{e}^{i_k}$, then the exterior derivative of $\boldsymbol{T}$ is given by [26, p.91]

$$\mathrm{d}\boldsymbol{T} = \mathrm{d}T_{i_1\dots i_k}\boldsymbol{e}^{i_1}\wedge\cdots\wedge\boldsymbol{e}^{i_k}, \tag{A.2}$$

where $\mathrm{d}T_{i_1\dots i_k}$ is the differential of $T_{i_1\dots i_k}$.

A tensor that will be of special interest in this paper is the Hessian tensor. For a twice differentiable function v it is given by

$$\boldsymbol{H}(v) = \nabla\mathrm{d}v = \Big(\nabla_{\boldsymbol{e}_j}(\nabla_{\boldsymbol{e}_i}v) - \Gamma^k_{ij}\nabla_{\boldsymbol{e}_k}v\Big)\boldsymbol{e}^i\otimes\boldsymbol{e}^j.$$

The covariant directional derivative as introduced above can be generalized to Riemannian manifolds and is in that context called a Levi-Civita connection [28, p.160]. For a Levi-Civita connection the Hessian matrix is symmetric [27, p.4], hence the Hessian matrix will always be symmetric in this paper. Note that in Cartesian coordinates $(H_{ij}(v))$ is the matrix with second derivatives of v.

Appendix B. Proofs of Lemmas 2.1 and 2.2.

Proof of Lemma 2.1. $u \in C^2(\mathcal{E})$ implies that ∇u is continuously differentiable. The bijectivity of ∇u follows from the strict convexity of u. ∇u is surjective by definition. To show injectivity, we argue by contradiction and will use a reasoning presented in [12, p.93]. Suppose $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{E}$, such that $\boldsymbol{x} \neq \boldsymbol{x}'$ and $\nabla u(\boldsymbol{x}) = \nabla u(\boldsymbol{x}')$. Due to strict convexity u lies above its tangent planes, i.e. $u(\boldsymbol{x}') > u(\boldsymbol{x}) + (\nabla u(\boldsymbol{x}), \boldsymbol{x}' - \boldsymbol{x})$ and similarly $u(\boldsymbol{x}) > u(\boldsymbol{x}') + (\nabla u(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}')$. Adding these two inequalities and subtracting $u(\boldsymbol{x}) + u(\boldsymbol{x}')$ from both sides we obtain $0 > (\nabla u(\boldsymbol{x}') - \nabla u(\boldsymbol{x}), \boldsymbol{x}' - \boldsymbol{x})$, which is contradicting the assumption $\nabla u(\boldsymbol{x}') = \nabla u(\boldsymbol{x})$. We have shown that ∇u is a continuously differentiable bijection and now proceed with the Jacobian.

We assume a Cartesian coordinate system on $\mathcal{F}$, with corresponding orthonormal basis vectors $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, and an arbitrary coordinate system and holonomic basis on $\mathcal{E}$. Let for each point in $\boldsymbol{x} \in \mathcal{E}$ $A(\boldsymbol{x}) = (a^i_j(\boldsymbol{x}))$ be the matrix that transforms that basis $\boldsymbol{e}_1(\boldsymbol{x}), \boldsymbol{e}_2(\boldsymbol{x})$ into $\boldsymbol{f}_1, \boldsymbol{f}_2$, i.e., $\boldsymbol{f}_i = a^j_i\boldsymbol{e}_j$. Let $B = (b^i_j)$ be the inverse of A. From $f^i(\boldsymbol{x})\boldsymbol{f}_i = \nabla u(\boldsymbol{x}) = e^{kj}\nabla_{\boldsymbol{e}_k}u(\boldsymbol{x})\boldsymbol{e}_j$ it follows that $f^i = b^i_j e^{kj}\nabla_{\boldsymbol{e}_k}u$. The area form on $\mathcal{F}$ is given by $\sqrt{f}\boldsymbol{f}^1\wedge\boldsymbol{f}^2 = \mathrm{d}f^1\wedge\mathrm{d}f^2$ [26, p.85]. By definition we have $\mathrm{d}f^i = \mathrm{d}(b^i_j e^{kj}\nabla_{\boldsymbol{e}_k}u) = \nabla_{\boldsymbol{e}_l}(b^i_j e^{kj}\nabla_{\boldsymbol{e}_k}u)\boldsymbol{e}^l$ (Appendix A). Applying the product rule we find

$$\nabla_{\boldsymbol{e}_l}(b^i_j e^{kj}\nabla_{\boldsymbol{e}_k}u) = \nabla_{\boldsymbol{e}_l}(b^i_j)e^{kj}\nabla_{\boldsymbol{e}_k}u + b^i_j\nabla_{\boldsymbol{e}_l}(e^{kj})\nabla_{\boldsymbol{e}_k}u + b^i_j e^{kj}\nabla_{\boldsymbol{e}_l}(\nabla_{\boldsymbol{e}_k}u). \tag{B.1}$$

Note that $b^i_j = (\boldsymbol{e}_j, \boldsymbol{f}_i)$ for $i, j = 1, 2$, hence we find

$$\nabla_{\boldsymbol{e}_l}(b^i_j) = \nabla_{\boldsymbol{e}_l}\big(\boldsymbol{e}_j, \boldsymbol{f}_i\big) = \big(\nabla_{\boldsymbol{e}_l}(\boldsymbol{e}_j), \boldsymbol{f}_i\big) = \big(\Gamma^k_{jl}\boldsymbol{e}_k, \boldsymbol{f}_i\big) = \Gamma^k_{jl} b^i_k. \tag{B.2}$$

The covariant derivatives of the metric tensor and its inverse are zero [18, p.215], hence

$$\nabla_{\boldsymbol{e}_k}(e^{ij}) = -\Gamma^j_{lk} e^{il} - \Gamma^i_{lk} e^{lj}. \tag{B.3}$$

Substituting (B.2) and (B.3) into (B.1) we find

$$\nabla_{\boldsymbol{e}_l}(b^i_j e^{kj} \nabla_{\boldsymbol{e}_k} u) = b^i_j e^{kj}\big(\nabla_{\boldsymbol{e}_l}(\nabla_{\boldsymbol{e}_k} u) - \Gamma^m_{kl} \nabla_{\boldsymbol{e}_m} u\big).$$

Using this and the properties of the wedge product [26, p.79] we obtain

$$\begin{aligned}
\mathrm{d}f^1 \wedge \mathrm{d}f^2 &= \big[b^1_j e^{kj}\big(\nabla_{\boldsymbol{e}_l}(\nabla_{\boldsymbol{e}_k} u) - \Gamma^m_{kl}\nabla_{\boldsymbol{e}_m} u\big)\boldsymbol{e}^l\big] \wedge \big[b^2_q e^{rq}\big(\nabla_{\boldsymbol{e}_s}(\nabla_{\boldsymbol{e}_r} u) - \Gamma^t_{rs}\nabla_{\boldsymbol{e}_t} u\big)\boldsymbol{e}^s\big] \\
&= \big[b^1_j e^{kj}\big(\nabla_{\boldsymbol{e}_1}(\nabla_{\boldsymbol{e}_k} u) - \Gamma^m_{k1}\nabla_{\boldsymbol{e}_m} u\big) b^2_q e^{rq}\big(\nabla_{\boldsymbol{e}_2}(\nabla_{\boldsymbol{e}_r} u) - \Gamma^t_{r2}\nabla_{\boldsymbol{e}_t} u\big) \\
&\quad - b^1_j e^{kj}\big(\nabla_{\boldsymbol{e}_2}(\nabla_{\boldsymbol{e}_k} u) - \Gamma^m_{k2}\nabla_{\boldsymbol{e}_m} u\big) b^2_q e^{rq}\big(\nabla_{\boldsymbol{e}_1}(\nabla_{\boldsymbol{e}_r} u) - \Gamma^t_{r1}\nabla_{\boldsymbol{e}_t} u\big)\big]\, \boldsymbol{e}^1 \wedge \boldsymbol{e}^2 \\
&= \det\big(b^i_j e^{jk} H_{kl}\big)\boldsymbol{e}^1 \wedge \boldsymbol{e}^2,
\end{aligned}$$

where $H_{ij} = (\nabla_{\boldsymbol{e}_i}(\nabla_{\boldsymbol{e}_j} u) - \Gamma^k_{ji}\nabla_{\boldsymbol{e}_k} u)$ are the components of the Hessian. From

$$e_{ij} = \big(b^k_i \boldsymbol{f}_k, b^l_j \boldsymbol{f}_l\big) = BB^T,$$

it follows that $\det(b^i_j) = \det(B) = \sqrt{\det(e_{ij})} = \sqrt{e}$. Using this and the multiplicativity of the determinant we find

$$\mathrm{d}f^1 \wedge \mathrm{d}f^2 = \frac{\det(H_{ij})}{e}\sqrt{e}\boldsymbol{e}^1 \wedge \boldsymbol{e}^2.$$

The area form on $\mathcal{E}$ is $\sqrt{e}\boldsymbol{e}^1 \wedge \boldsymbol{e}^2$ and on $\mathcal{F}$ is $\mathrm{d}f^1 \wedge \mathrm{d}f^2$, hence $|\mathrm{D}\nabla u(\boldsymbol{x})| = |\det(H_{ij})/e|$. By the assumption of the lemma u is strictly convex and hence if we consider a Cartesian coordinate system $\det(H_{ij}) > 0$ and $e = 1$, implying $\det(H_{ij})/e > 0$. However, $\det(H_{ij})/e$ is independent of the choice of coordinate system, as can be easily checked by considering a coordinate transformation, and therefore must be positive independent of the choice of coordinate system. Thus in general we have $|\mathrm{D}\nabla u(\boldsymbol{x})| = \det(H_{ij})/e$.

Proof of Lemma 2.2. We first prove injectivity. Suppose we have two distinct $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{R}^2 \times \{0\}$ such that $\boldsymbol{s}(\boldsymbol{y}_1) = \boldsymbol{s}(\boldsymbol{y}_2)$. This implies that

$$\frac{\boldsymbol{y}_1 - \boldsymbol{e}_3}{\|\boldsymbol{y}_1 - \boldsymbol{e}_3\|^2} = \frac{\boldsymbol{y}_2 - \boldsymbol{e}_3}{\|\boldsymbol{y}_2 - \boldsymbol{e}_3\|^2}.$$

Using the fact that $\boldsymbol{e}_3$ is orthogonal to both $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ we find that $\|\boldsymbol{y}_1 - \boldsymbol{e}_3\|^2 = \|\boldsymbol{y}_2 - \boldsymbol{e}_3\|^2$ and this on its turn implies that $\boldsymbol{y}_1 = \boldsymbol{y}_2$.

The map is also surjective from $\mathbb{R}^2 \times \{0\}$ to $\mathcal{S}^2 \backslash \boldsymbol{e}_3$. Suppose we have a spherical coordinate system (g^1, g^2) on $\mathcal{S}^2 \backslash \boldsymbol{e}_3$ and a Cartesian coordinate system (f^1, f^2) on $\mathbb{R}^2 \times \{0\}$. Suppose we have defined the Cartesian coordinate system on $\mathbb{R}^2 \times \{0\}$ with basis vectors $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$. Let g^1 be the azimuthal angle with respect to $\boldsymbol{f}_1$ and let g^2 be the zenithal angle with respect to $\boldsymbol{e}_3$. For $(g^1, g^2) \in \mathcal{S}^2 \backslash \boldsymbol{e}_3$ we have [12, p.77]

$$f^1(g^1, g^2) = \frac{\sin(g^2)\cos(g^1)}{1 - \cos(g^1)}, \qquad f^2(g^1, g^2) = \frac{\sin(g^2)\sin(g^1)}{1 - \cos(g^1)},$$

indicating surjectivity of s. Conversely,

$$g^1(f^1,f^2) = \tan^{-1}\left(f^1,f^2\right), \qquad g^2(f^1,f^2) = \arccos\left(\frac{(f^1)^2+(f^2)^2-1}{(f^1)^2+(f^2)^2+1}\right),$$

where $\tan^{-1}\left(f^1,f^2\right)$ is the four-quadrant variant on $\arctan(f^2/f^1)$ [12, p.77]. In spherical coordinates the determinant of the metric $g = \det(g_{ij}) = \sin^2(g^2)$ [18, p.340], hence the area form on the $\mathcal{S}^2 \backslash \boldsymbol{e}_3$ is given by $\sqrt{g}\boldsymbol{g}^1 \wedge \boldsymbol{g}^2 = \sin(g^2)\mathrm{d}g^1 \wedge \mathrm{d}g^2$. Using the definition of the exterior derivative (Appendix A) we find $\mathrm{d}g^i = (\partial g^i/\partial f^j)\mathrm{d}f^j$ and hence $\mathrm{d}g^1 \wedge \mathrm{d}g^2 = \det(\partial g^i/\partial f^j)\mathrm{d}f^1 \wedge \mathrm{d}f^2$. By direct calculation we find

$$\det\left(\frac{\partial g^i}{\partial f^j}\right) = \frac{1}{\sqrt{(f^1)^2+(f^2)^2}}\frac{2}{(f^1)^2+(f^2)^2+1}.$$

Furthermore, we have

$$\sin(g^2) = \sin\left(\arccos\left(\frac{(f^1)^2+(f^2)^2-1}{(f^1)^2+(f^2)^2+1}\right)\right) = \frac{2\sqrt{(f^1)^2+(f^2)^2}}{(f^1)^2+(f^2)^2+1},$$

where we have used the fact that $\sin(g^2) \geq 0$ and the identity $\sin(\arccos(x)) = \sqrt{1-x^2}$. Combining the two we find

$$\begin{aligned}\sqrt{g}\mathrm{d}g^1 \wedge \mathrm{d}g^2 &= \left(\frac{2\sqrt{(f^1)^2+(f^2)^2}}{(f^1)^2+(f^2)^2+1}\frac{1}{\sqrt{(f^1)^2+(f^2)^2}}\frac{2}{(f^1)^2+(f^2)^2+1}\right)\mathrm{d}f^1 \wedge \mathrm{d}f^2 \\ &= \left(\frac{4}{((f^1)^2+(f^2)^2+1)^2}\right)\mathrm{d}f^1 \wedge \mathrm{d}f^2.\end{aligned}$$

The area form on $\mathcal{S}^2 \backslash \boldsymbol{e}_3$ is $\sqrt{g}\mathrm{d}g^1 \wedge \mathrm{d}g^2$, the area form $\mathbb{R}^2 \times \{0\}$ in Cartesian coordinates is $\sqrt{f}\boldsymbol{f}^1 \wedge \boldsymbol{f}^2 = \mathrm{d}f^1 \wedge \mathrm{d}f^2$. Thus we find that the Jacobian of s is as given in the statement of the lemma.

Appendix C. Equivalence of boundary conditions for a strictly convex reflector surface. In this appendix we show that the Inverse reflector problem as stated on page 7 is equivalent to this same problem but with the boundary condition $\nabla u(\partial\mathcal{E}) = \partial\mathcal{F}$ replaced by $\nabla u(\mathcal{E}) = \mathcal{F}$. When doing this we need to make use of the fact that ∇u is an open map, i.e. a map that maps open sets to open sets. This we show in the following lemma.

Lemma C.1. *Suppose that $u \in C^2(\mathcal{E})$ is the strictly convex solution to the Inverse reflector problem with $\nabla u(\mathcal{E}) = \mathcal{F}$ instead of $\nabla u(\partial\mathcal{E}) = \partial\mathcal{F}$. Then the map ∇u is also open, i.e., for each open subset $A \subset \mathcal{E}$ the image $\nabla u(A)$ is an open subset of $\mathcal{F}$.*

Proof. In Lemma 2.1 we saw that for the strictly convex solution $u \in C^2(\mathcal{E})$, ∇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. This implies that the conditions for the inverse function theorem [21] are satisfied. The inverse function theorem states, among other things, that for every open subset A of $\mathcal{E}$ and $x \in A$, there exists an open set U in A containing x, and an open set V in $\mathcal{F}$ containing $\nabla u(x)$ 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 A 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 $U_x \subset A$. $\nabla u(U_x) = V_x$ is open for every $x \in A$. Notice that $\cup_{x \in E} U_x = A$ and that $\nabla u(A) = \cup_{x\in A} \nabla u(U_x) = \cup_{x \in A} V_x$. Thus ∇u is an open map. □

The map ∇u is a homeomorphism from $\mathcal{E}$ to $\mathcal{F}$, because it is a continuous bijection which is open and hence also has continuous inverse. We will use this convenient property of ∇u in the following lemma.

Lemma C.2. *Let $u \in C^2(\mathcal{E})$ be the strictly convex solution to the Inverse reflector problem with $\nabla u(\mathcal{E}) = \mathcal{F}$ instead of $\nabla u(\partial\mathcal{E}) = \partial\mathcal{F}$. Then also $\nabla u(\partial\mathcal{E}) = \partial\mathcal{F}$.*

Proof. The map ∇u is a homeomorphism and therefore it links every open set in $\mathcal{E}$ with an open set in $\mathcal{F}$ and vice versa. Let us denote by $\mathrm{int}(A)$ the interior of a set A. Suppose $A \subset \mathcal{E}$. We have $\nabla u(\mathrm{int}(A)) \subset \nabla u(A)$ and because ∇u is an open map $\nabla u(\mathrm{int}(A))$ is also open. The largest open subset of $\nabla u(A)$ is the interior $\mathrm{int}(\nabla u(A))$, therefore $\nabla u(\mathrm{int}(A)) \subset \mathrm{int}(\nabla u(A))$. If $\nabla u : \mathcal{E} \to \mathcal{F}$ is a homeomorphism, then $(\nabla u)^{-1} : \mathcal{F} \to \mathcal{E}$ is a homeomorphism also, hence $(\nabla u)^{-1}(\mathrm{int}(B)) \subset \mathrm{int}((\nabla u)^{-1}(B))$ for all $B \subset \mathcal{F}$. From this it follows that we have both $\nabla u(\mathrm{int}(\mathcal{E})) \subset \mathrm{int}(\nabla u(\mathcal{E})) = \mathrm{int}(\mathcal{F})$ and $(\nabla u)^{-1}(\mathrm{int}(\mathcal{F})) \subset \mathrm{int}(\nabla u)^{-1}(\mathcal{F})) = \mathrm{int}(\mathcal{E})$. Using this we see that

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

Thus, we see that $\nabla u\left(\mathrm{int}(\mathcal{E})\right) = \mathrm{int}(\mathcal{F})$. 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 the Inverse reflector problem with boundary condition $\nabla u(\mathcal{E}) = \mathcal{F}$ is also a solution to the Inverse reflector problem with boundary condition $\nabla u(\partial\mathcal{E}) = \partial\mathcal{F}$. Now the following lemma states the converse.

Lemma C.3. *Let $u \in C^2(\mathcal{E})$ be a strictly convex solution to Inverse reflector problem. 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 also a Jordan curve. Now the Jordan curve theorem [22, p.198] 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, therefore $\nabla u(\mathcal{E})$ is simply connected also. The interior and exterior to the curve $\nabla u(\partial\mathcal{E}) = \partial\mathcal{F}$ are the only two subsets of $\mathbb{R}^2$ with $\partial\mathcal{F}$ as boundary. The fact that ∇u is a homeomorphism implies $\nabla u(\partial\mathcal{E}) = \partial(\nabla u(\mathcal{E}))$, because $\nabla u(\mathrm{int}(\mathcal{E})) = \mathrm{int}(\nabla u(\mathcal{E}))$ as we showed in the proof of Lemma C.2. The fact that $\partial\mathcal{F} = \nabla u(\partial\mathcal{E}) = \partial(\nabla u(\mathcal{E}))$ implies that $\mathrm{int}(\nabla u(\mathcal{E}))$ 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 $\mathrm{int}(\nabla u(\mathcal{E}))$ is the interior set in the Jordan curve theorem. The fact that $\mathcal{E}$ is bounded, that the functions E and F are strictly positive and bounded and that F is bounded away from zero implies by (2.7) that the set $\mathcal{F}$ is bounded also. This implies that $\mathrm{int}(\mathcal{F})$ 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 to the Inverse reflector problem with boundary conditions $\nabla u(\mathcal{E}) = \mathcal{F}$ if and only if it is a strictly convex solution to the Inverse reflector problem with boundary condition $\nabla u(\partial\mathcal{E}) = \partial\mathcal{F}$. Thus the two boundary conditions are equivalent.

REFERENCES

[1] R. Haitz, Y. T. Tsao, *Solid-state lighting: 'The Case' 10 years after and future prospects*, Physica Status Solidi A 208, N0.1, 17-29, 2011.

[2] T. Taguchi, *Present Status of Energy Saving Technologies and Future Prospects in White LED Lighting*, IEEJ Transactions On Electrical and Electronic Engineering, 3:21-26, 2008.

[3] F.Z. Fang, X.D. Zhang, A. Weckenmann, G.X. Zhang, C. Evans, *Manufacturing and measurement of freeform optics*, CIRP Annals - Manufacturing Technology Volume 62, Issue 2, 823846, 2013.

[4] A. S. Glassner, *An introduction to Ray Tracing*, Academic Press, 1991.

[5] C. Villani, *Topics in Optimal Transportation.* Providence: American Mathematical Society, 2003.

[6] V. Oliker, *Designing Freeform Lenses for Intensity and Phase Control of Coherent Light with Help from Geometry and Mass Transport.* Archive for Rational Mechanics and Analysis, Volume 201, Issue 3, 2011.

[7] J. D. Benamou, Y. Brenier, *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem.* Numer. Math, 84, pp. 375393, 2000.

[8] E. Haber, T. Rehman, A. Tannenbaum, *An efficient numerical method for the solution of the L_2 optimal mass transfer problem.* SIAM J. Sci. Comput., 32, pp. 197211, 2010.

[9] B. D. Froese, A. M. Oberman, *Fast finite difference solvers for singular solutions of the elliptic MongeAmpère equation.* Journal of Computational Physics 230 (3), 818-834, 2011.

[10] J. D. Benamou, B. D. Froese, A. M. Oberman, *Numerical solution of the optimal transportation problem using the MongeAmpère equation.* Journal of Computational Physics 260, 107-126, 2014.

[11] K. Brix, Y. Hafizogullari, A. Platen, *Designing illumination lenses and mirrors by the numerical solution of MongeAmpère equations.* Journal of the Optical Society of America Vol. 32, No. 11, 2227-2236, 2015.

[12] C. R. Prins, Ph.D. Thesis, *Inverse Methods for Illumination Optics*, Eindhoven University of Technology, 2014.

[13] C. R. Prins, R. Beltman, J. H. M. ten Thije Boonkkamp, W. L. IJzerman, T. W. Tukker, *A Least-Squares Method for Optimal Transport using the Monge-Ampère Equation.* SIAM J. Sci. Comput., 37(6), B937-B961, 2015.

[14] *Official website on Escher: http://www.mcescher.com/about/*

[15] M. Bass (ed.), *Handbook of Optics Volume II - Devices, Measurements and Properties, 2nd Ed.* McGraw-Hill, 1995.

[16] G. Monge, *Application de l'analyse a la geometrie, a l'usage de l'Ecole imperiale polytechnique.* Paris: Bernard, 1807.

[17] H. Cohn, *Conformal Mapping on Riemann Surfaces.* McGraw-Hill, 1967.

[18] C. W. Misner, K. S. Thorne, J. A. Wheeler, *Gravitation.* San Francisco: W. H. Freeman, 1973.

[19] I. N. Bronshtein, K. A. Semendyayev, G. Musiol, H. Muehlig, *Handbook of Mathematics, 4th ed.* New York: Springer-Verlag, 2004.

[20] W. Rudin, *Real and Complex Analysis 3rd ed.* London: McGraw-Hill, 1987.

[21] T. Tao, *Analysis II, 2nd ed.* Hindustan Book Agency, 2009.

[22] E.H. Spanier, *Algebraic Topology.* London: McGraw-Hill, 1966.

[23] S. Kobayashi, K. Nomizu, *Foundations of Differential Geometry.* S.I.: Interscience, 1963.

[24] J.E. Marsden, A.J. Tromba, *Vector Calculus, 5th ed.* W.H. Freeman and Company, 2003.

[25] S. Boyd, L. Vandenberghe, *Convex Optimization.* Cambridge University Press, 2004.

[26] M. Spivak *Calculus on manifolds: a modern approach to classical theorems of advanced calculus.* Amsterdam: Benjamin, 1965.

[27] J. Morgan, G. Tian, *Ricci Flow and the Poincaré Conjecture.* American Mathematical Society, Providence RI, 2007.

[28] J. Jost *Riemannian Geometry and Geometric Analysis, 6th ed.* Springer-Verlag, Berlin Heidelberg, 2011.

[29] R. Courant, D. Hilbert *Methods of Mathematical Physics, Vol. 1 (1989 ed.).* John Wiley & Sons, 1937.

PREVIOUS PUBLICATIONS IN THIS SERIES:

Number	Author(s)	Title	Month
16-13	S.W. Rienstra	Sound Propagation in Slowly Varying 2D Duct with Shear Flow	May ‘16
16-14	A.S. Tijsseling Q. Hou Z. Bozkuş	Analytical and numerical solution for a rigid liquid-column moving in a pipe with fluctuating reservoir-head and venting entrapped-gas	May ‘16
16-15	M.F.P. ten Eikelder J.H.M. ten Thije Boonkkamp B.V. Rathish Kumar	A Finite Volume-Complete Flux Scheme for the Singularly Perturbed Generalized Burgers-Huxley Equation	June ‘16
16-16	A.S. Tijsseling	An overview of fluid-structure interaction experiments in single-elbow pipe systems	July ‘16
16-17	R. Beltman J.H.M. ten Thije Boonkkamp W.L. IJzerman	A least-squares method for the inverse reflector problem in arbitrary orthogonal coordinate systems	August ‘16

Ontwerp: de Tantes,
Tobias Baanders, CWI