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

Journal of Computational Physics 367 (2018) 347–373

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

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

René Beltman [a,*], Jan ten Thije Boonkkamp [a], Wilbert IJzerman [b]

[a] Dept. of Mathematics and Computer Science, Eindhoven University of Technology, PO Box 513 5600 MB Eindhoven, The Netherlands
[b] Philips Lighting, High Tech Campus 7, 5656 AE Eindhoven, The Netherlands

ARTICLE INFO

Article history:
Received 21 February 2018
Received in revised form 13 April 2018
Accepted 20 April 2018
Available online 27 April 2018

Keywords:
Monge–Ampère equation
Inverse reflector problem
Least-squares method

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 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, viz., LEDs integrated in optical systems for illumination, are used in illumination optics ever more frequently.

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 [3] 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, for given light source and desired light output, a partial differential equation can be derived relating these to the geometry of the optical system. The solution of this partial differential equation then 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 forward 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

* Corresponding author.
E-mail address: r.beltman@tue.nl (R. Beltman).

https://doi.org/10.1016/j.jcp.2018.04.041
0021-9991/

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 [4].

In this paper we consider an optical system consisting of an incoming parallel bundle of light and a reflecting surface. This optical system is relevant, because parallel bundles can be easily formed in LED lighting systems by placing a converging lens above 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 is an equation of the Monge–Ampère type.

Monge–Ampère type equations also arise in the context of optimal mass transport (OMT). Inverse reflector problems and OMT problems 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 a 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 and Monge–Ampère type equations 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 Prins [12] and the aforementioned article by Brix et al.

In a recent publication, Prins et al. [13] introduced a least-squares method (LS method) to solve the OMT problem related to an inverse reflector problem. This method is based on a least-squares method for the Monge–Ampère equation with Dirichlet boundary conditions by Caboussat et al. [14]. 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. It 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 the Dutch painter Johannes 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 troublesome 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 derive the Monge–Ampère equation in coordinate-free form and generalize the LS method to arbitrary orthogonal coordinate systems. The GLS method like the LS method is an iterative method in which each iteration consists of three minimization steps. In one of these minimization steps a pair of boundary value problems is solved. For the LS method, these problems are decoupled because it uses Cartesian coordinates. However, in the GLS method 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 three test cases. In all 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 the light source is mapped to a square gradient set, in the second test a Gaussian source emittance is used and in the third test we consider a target distribution corresponding to a lithograph by the artist M.C. Escher (Fig. 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 pointers to classical literature on these matters. In Appendix B one can find a proof of a result used in the main text. The reading of this proof should not be necessary for understanding the rest of the paper.

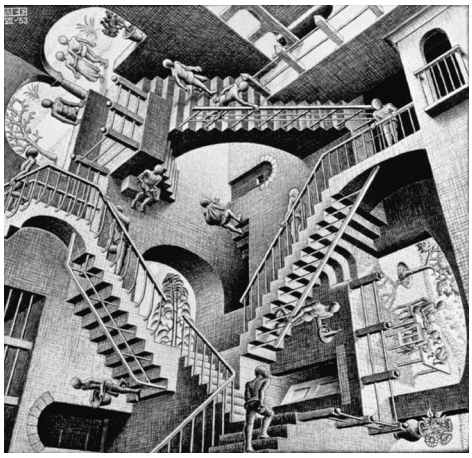

Fig. 1. Lithograph *Relativity* (1953) by the Dutch artist M.C. Escher who was frequently inspired by mathematics [15]. This lithograph, with its great detail, will serve as the ultimate test.

2. Monge–Ampère equation and inverse reflector problem

We introduce in this section the optical system and derive the corresponding Monge–Ampère equation. This formulation is not new (cf. the overview in [11] for earlier derivations), but it allows us to introduce our notation and reformulate the problem in the specific form that we will solve in Section 3.

The optical 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 $\mathscr{E} \subset \mathbb{R}^2 \times \{-a\}$, a subset of a plane below and parallel to the x-y plane at a distance $a > 0$. For now, we assume a Cartesian coordinate system on $\mathscr{E}$ with coordinates x and y, and, corresponding basis vectors $\boldsymbol{e}_x$ and $\boldsymbol{e}_y$. Let $(\cdot,\cdot)$ denote the Euclidean inner product on the ambient space $\mathbb{R}^3$ and let $\|\cdot\|$ be the corresponding norm. 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 $\boldsymbol{x} = (x, y) \in \mathscr{E}$ is given in luminous flux per area by $E(x, y)$ [lm/m^2], where $E : \mathscr{E} \to (0, \infty)$ is the emittance function, which we assume to be continuous. $E(x, y)\mathrm{d}x\mathrm{d}y$ expresses the light flux through an infinitesimal area element on $\mathscr{E}$. For details on photometric quantities, see for example [16]. The light rays leaving the source will all hit upon the reflector surface. We describe the reflector surface by a function $u : \mathscr{E} \to (-a, \infty)$. A ray leaving from the point $\boldsymbol{x} \in \mathscr{E}$ will travel a distance $a + u(\boldsymbol{x})$ in the z-direction before hitting upon the reflector surface. The function u is the Monge parameterization of the reflector surface [17]. 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, as a consequence of upcoming Lemma 1 and Lemma 2, that strict convexity of u implies that a pair of rays leaving $\mathscr{E}$ 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 [3, p. 132]

$$\boldsymbol{r} = \boldsymbol{i} - 2(\boldsymbol{i}, \boldsymbol{n})\boldsymbol{n}, \tag{1}$$

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 $\boldsymbol{x} \in \mathscr{E}$ 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 $\boldsymbol{e}_z$. This vector complements the two-dimensional bases on $\mathscr{E}$ 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 $\boldsymbol{e}_z$ and when we substitute this in (1), we obtain

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

For all $\boldsymbol{x} \in \mathscr{E}$ 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, $\mathscr{S}^2$. 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 $r : \mathscr{E} \to \mathscr{S}^2$. The optical system is depicted in Fig. 2.

The reflected light will shine in a set of directions $\mathscr{G} \subset \mathscr{S}^2$. We assume a spherical coordinate system on $\mathscr{G}$ with $\phi \in [0, \pi]$ the zenith angle between $\boldsymbol{r}$ and $\boldsymbol{e}_z$, and, $\theta \in [0, 2\pi)$ the azimuth angle between $\boldsymbol{e}_x$ and the projection of $\boldsymbol{r}$ on

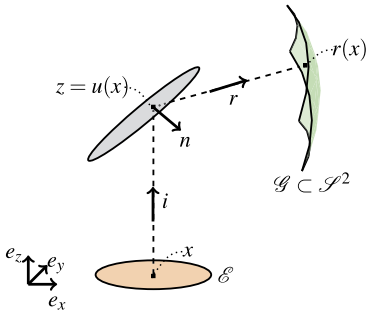

Fig. 2. The set-up of the reflector problem. A light ray from a point $x \in \mathscr{E}$ is emitted in the direction $\boldsymbol{i} = \boldsymbol{e}_z$. Subsequently, the ray is reflected at the point $u(x)\boldsymbol{e}_z$ according to the law of reflection (1) in the direction $\boldsymbol{r}$. This direction corresponds to a point $r(x) \in \mathscr{G}$. The set $\mathscr{G}$ (in green) is a subset of the unit sphere $\mathscr{S}^2$. Note the unit vectors and unit sphere are not drawn in proportion. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

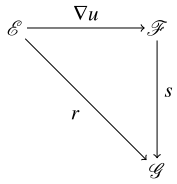

Fig. 3. The mappings and sets involved in the inverse reflector problem.

the x-y plane. Let us describe the luminous intensity in the directions $\mathscr{G}$ by a continuous function $G : \mathscr{G} \to (0, \infty)$. The luminous intensity is the luminous flux per solid angle [lm/sr]. The luminous flux through an infinitesimal surface area element on $\mathscr{G}$ is given by $G(\phi, \theta)\sin(\phi)\mathrm{d}\phi\mathrm{d}\theta$. In practice the couple $\mathscr{G}$ 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 $\mathscr{G}$ is confined to one half of $\mathscr{S}^2$ there is a one-to-one correspondence between the couple $\mathscr{G}$ and G on the one hand and the intensity pattern in the far-field on the other hand. Details can be found in [12]. We call $\mathscr{G}$ the target set.

The problem we want to solve is informally stated as follows. *Given a light source $\mathscr{E}$ 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 $\mathscr{G}$ 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 $u \in C^2(\mathscr{E})$, derive a partial differential equation from the principle of conservation of luminous flux. The luminous flux through $U \subset \mathscr{E}$ results in a luminous flux through the set $r(U) \subset \mathscr{S}^2$. By conservation of luminous flux these two fluxes must be equal and therefore we have

$$\int_U E(x, y)\mathrm{d}x\mathrm{d}y = \int_{r(U)} G(\phi, \theta)\sin(\phi)\mathrm{d}\phi\mathrm{d}\theta, \tag{3}$$

for every Lebesgue measurable set $U \subset \mathscr{E}$. We can use (3) to derive the partial differential equation. To see this we must closely examine the map $r : \mathscr{E} \to \mathscr{S}^2$. From (2) it can be seen that $\boldsymbol{r}(\boldsymbol{x})$ only depends on the gradient of u at 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. The relation between ∇u, s and r is depicted in Fig. 3. 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 $\mathscr{E}$ and because it has its initial point in the origin it lies in the plane $\mathbb{R}^2 \times \{0\}$. From relation (2) we see that s maps a vector $\boldsymbol{v}$ in this plane to the unit sphere according to

$$\boldsymbol{v} \mapsto \boldsymbol{e}_z + 2\frac{\boldsymbol{v} - \boldsymbol{e}_z}{\|\boldsymbol{v} - \boldsymbol{e}_z\|^2}. \tag{4}$$

Closer inspection reveals that this map is the inverse of the stereographic projection pictured in Fig. 4 [18, p. 26]. It is the bijection between $\mathscr{S}^2 \backslash \boldsymbol{e}_z$, i.e., the unit-sphere without its north pole, and $\mathbb{R}^2 \times \{0\}$, the plane intersecting the equator of $\mathscr{S}^2 \backslash \boldsymbol{e}_z$. To substantiate this claim, consider a point on the unit sphere given by $\xi\boldsymbol{e}_x + \eta\boldsymbol{e}_y + \zeta\boldsymbol{e}_z$, $\xi^2 + \eta^2 + \zeta^2 = 1$. The stereographic projection maps $\xi\boldsymbol{e}_x + \eta\boldsymbol{e}_y + \zeta\boldsymbol{e}_z$ to the point $p(\xi, \eta, \zeta)\boldsymbol{e}_x + q(\xi, \eta, \zeta)\boldsymbol{e}_y$, with

$$p(\xi, \eta, \zeta) = \frac{\xi}{1 - \zeta}, \qquad q(\xi, \eta, \zeta) = \frac{\eta}{1 - \zeta}. \tag{5}$$

We can easily verify that the inversion of these equations is given by

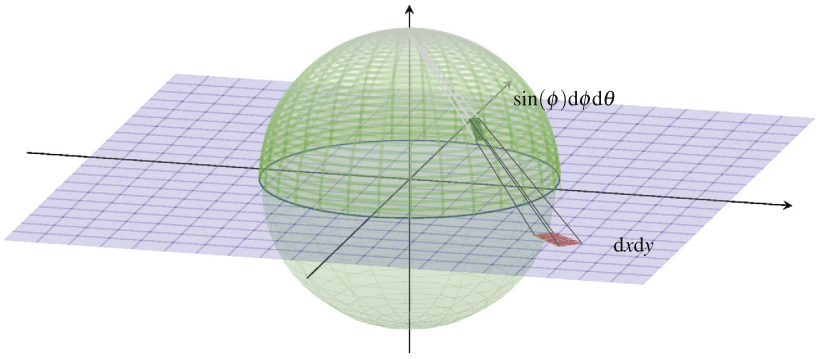

Fig. 4. The inverse stereographic projection maps the surface element $\mathrm{d}x\mathrm{d}y$ on $\mathbb{R}^2 \times \{0\}$ to the surface element $\sin(\phi)\mathrm{d}\phi\mathrm{d}\theta$ on $\mathscr{S}^2 \backslash \boldsymbol{e}_z$.

$$\xi(p,q) = \frac{2p}{p^2+q^2+1}, \qquad \eta(p,q) = \frac{2q}{p^2+q^2+1}, \qquad \zeta(p,q) = \frac{p^2+q^2-1}{p^2+q^2+1}.$$

We now see from (4) that $s : p\boldsymbol{e}_x + q\boldsymbol{e}_y \mapsto \xi(p,q)\boldsymbol{e}_x + \eta(p,q)\boldsymbol{e}_y + \zeta(p,q)\boldsymbol{e}_z$, which implies that s is indeed the inverse of the stereographic projection as defined in (5).

We proceed with examining the map $r : \mathscr{E} \to \mathscr{S}^2$ in order to derive the sought partial differential equation. When we identify the vector $\nabla u(\boldsymbol{x})$ with its endpoint, we can interpret $\nabla u(\mathscr{E})$ as a subset of $\mathbb{R}^2 \times \{0\}$. We will present bijectivity of ∇u in the following lemma.

Lemma 1. *Let $u \in C^2(\mathscr{E})$ be strictly convex and let us define*

$$p := \frac{\partial u}{\partial x} \quad \text{and} \quad q := \frac{\partial u}{\partial y}.$$

The map $\nabla u : \mathscr{E} \to \nabla u(\mathscr{E})$ is a continuously differentiable bijection, with Jacobian

$$\frac{\partial(p,q)}{\partial(x,y)} = \det \begin{pmatrix} \dfrac{\partial^2 u}{\partial x^2} & \dfrac{\partial^2 u}{\partial x \partial y} \\ \dfrac{\partial^2 u}{\partial x \partial y} & \dfrac{\partial^2 u}{\partial y^2} \end{pmatrix}, \tag{6}$$

i.e., the Jacobian of the map ∇u is the determinant of the Hessian tensor.

Proof. $u \in C^2(\mathscr{E})$ 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 $\boldsymbol{x}, \boldsymbol{x}' \in \mathscr{E}$, 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. The fact that the Jacobian is the determinant of the Hessian follows directly from the definition of the Cartesian coordinates p and q on $\nabla u(\mathscr{E})$. □

Note that the fact that u is convex implies that the Hessian is positive semi-definite and this again implies that the determinant of the Hesssian is nonnegative. For the inverse of the stereographic projection, s, we have the following similar result.

Lemma 2. *The inverse of the stereographic projection $s : \mathbb{R}^2 \times \{0\} \to \mathscr{S}^2 \backslash \boldsymbol{e}_z$ is continuously differentiable and hence $s : \nabla u(\mathscr{E}) \to s(\nabla u(\mathscr{E}))$ is a continuously differentiable bijection. Let us denote by the Cartesian coordinates (p,q) the points in $\nabla u(\mathscr{E})$ and let $(\phi,\theta) = (\phi(p,q), \theta(p,q))$ be the image of (p,q) under s represented in spherical coordinates. For the Jacobian of s we have*

$$\sin(\phi)\frac{\partial(\phi,\theta)}{\partial(p,q)} = \frac{4}{(1+p^2+q^2)^2}. \tag{7}$$

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

$$\frac{\boldsymbol{v} - \boldsymbol{e}_z}{\|\boldsymbol{v} - \boldsymbol{e}_z\|^2} = \frac{\boldsymbol{v}' - \boldsymbol{e}_z}{\|\boldsymbol{v}' - \boldsymbol{e}_z\|^2}.$$

Using the fact that $\boldsymbol{e}_z$ is orthogonal to both $\boldsymbol{v}$ and $\boldsymbol{v}'$ we find that $\|\boldsymbol{v}-\boldsymbol{e}_z\|^2 = \|\boldsymbol{v}'-\boldsymbol{e}_z\|^2$ and this in turn implies that $\boldsymbol{v}=\boldsymbol{v}'$.

The map is also surjective from $\mathbb{R}^2\times\{0\}$ to $\mathscr{S}^2\backslash\boldsymbol{e}_z$. For $(\phi,\theta)\in(0,\pi]\times[0,2\pi)$ we have [12, p. 77]

$$p(\phi,\theta)=\frac{\sin(\phi)\cos(\theta)}{1-\cos(\phi)},\qquad q(\phi,\theta)=\frac{\sin(\phi)\sin(\theta)}{1-\cos(\phi)},$$

indicating surjectivity of s. Conversely,

$$\theta(p,q)=\tan^{-1}(p,q),\qquad \phi(p,q)=\arccos\left(\frac{p^2+q^2-1}{p^2+q^2+1}\right),$$

where $\theta(p,q)=\tan^{-1}(p,q)$ is the four-quadrant variant of $\arctan(q/p)$ [12, p. 77]. By direct calculation we find

$$\frac{\partial(\phi,\theta)}{\partial(p,q)}=\frac{1}{\sqrt{p^2+q^2}}\frac{2}{p^2+q^2+1}.$$

Furthermore, we have

$$\sin(\phi)=\sin\left(\arccos\left(\frac{p^2+q^2-1}{p^2+q^2+1}\right)\right)=\frac{2\sqrt{p^2+q^2}}{p^2+q^2+1},$$

where we have used the fact that $\sin(\phi)\geq 0$ and the identity $\sin(\arccos(x))=\sqrt{1-x^2}$. Combining the two relations we find (7). □

Consequently, $s\circ\nabla u$ is a bijection implying that rays with different locations at the source are reflected in different directions.

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 1. *Assume $\mathscr{E}\subset\mathbb{R}^2\times\{-a\}$ is convex, closed and bounded, $u\in C^2(\mathscr{E})$ strictly convex and $r=s\circ\nabla u$, where s is the inverse of the stereographic projection. Let E and G be continuous, strictly positive and bounded functions. Furthermore, assume we have a coordinate system on $\mathscr{E}$ with metric e_{ij} and $e=\det(e_{ij})$, and let $h_{ij}(u)$ be the coefficients of the Hessian tensor (given in Appendix A, equation (A.6)) in the basis of this coordinate system. Then the function u satisfies the Monge–Ampère type differential equation*

$$\frac{4\det(h_{ij}(u(\boldsymbol{x})))}{e(\boldsymbol{x})(1+\|\nabla u(\boldsymbol{x})\|^2)^2}=\frac{E(\boldsymbol{x})}{G(r(\boldsymbol{x}))},\tag{8}$$

for every $\boldsymbol{x}\in\mathscr{E}$.

Proof. We first derive an equation in the Cartesian coordinate system on $\mathscr{E}$ and then generalize this equation to arbitrary coordinate systems on $\mathscr{E}$.

Both s and ∇u are continuously differentiable injections, therefore we can apply integration by substitution [19, Thm. 7.26]. For every Lebesgue measurable open subset $U\subset\mathscr{E}$ we have

$$\int_{r(U)}G(\phi,\theta)\sin(\phi)\mathrm{d}\phi\mathrm{d}\theta=\int_U G(\phi(x,y),\theta(x,y))\sin(\phi)\frac{\partial(\phi,\theta)}{\partial(p,q)}\frac{\partial(p,q)}{\partial(x,y)}\mathrm{d}x\mathrm{d}y,$$

where no absolute values appear because both Jacobian determinants are nonnegative. Using identity (3) we find

$$\int_U E(x,y)\mathrm{d}x\mathrm{d}y=\int_U G(\phi(x,y),\theta(x,y))\sin(\phi)\frac{\partial(\phi,\theta)}{\partial(p,q)}\frac{\partial(p,q)}{\partial(x,y)}\mathrm{d}x\mathrm{d}y,$$

for every Lebesgue measurable $U\subset\mathscr{E}$. The continuity of the functions E, G, r, the Jacobians and the sine function, Lemma 1 and Lemma 2 and the fact that the determinant of the Hessian of a convex function is positive imply that

$$\frac{4\det(h_{ij}(u(x,y)))}{(1+\|\nabla u(x,y)\|^2)^2}=\frac{E(x,y)}{G(\phi(x,y),\theta(x,y))}\quad\forall(x,y)\in\mathscr{E}.$$

This equation is the Cartesian coordinate expression of (8), because in the Cartesian coordinate system $e=1$ and the Hessian is just the matrix with the second order partial derivatives.

The left hand side of (8) is a scalar and hence independent of the choice of coordinate system and basis. The term $4/(1+\|\nabla u(\boldsymbol{x})\|^2)^2$ is also independent of the coordinate system and basis in use as it only involves the norm of a vector. To see that this also holds for $\det(h_{ij}(u(\boldsymbol{x})))/(e(\boldsymbol{x}))$, we consider a basis transformation $a^i_j\boldsymbol{e}_i=\bar{\boldsymbol{e}}_j$, where $\boldsymbol{e}_1,\boldsymbol{e}_2$ are the old

basis vectors and $\bar{\boldsymbol{e}}_1, \bar{\boldsymbol{e}}_2$ are the new basis vectors. Let $B = (b^i_j)$ be the inverse of $A = (a^i_j)$. The Hessian and the metric are both tensors, as such they transform according to the tensor transformation law [20, p. 204], i.e., we have $\bar{h}_{ij} = b^k_i b^l_j h_{kl}$ and $\bar{e}_{ij} = b^k_i b^l_j e_{kl}$. From this it follows that $\det(\bar{h}_{ij}) = (\det(B))^2 \det(h_{ij})$ and $\bar{e} = (\det(B))^2 e$ for the determinants of the metrics, which implies $\det(\bar{h}_{ij})/\bar{e} = \det(h_{ij})/e$. Thus we see that (8) is indeed independent of the choice of coordinate system on $\mathscr{E}$. □

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

$$\int_{\mathscr{E}} E(x, y)\mathrm{d}x\mathrm{d}y = \int_{\mathscr{G}} G(\phi, \theta)\sin(\phi)\mathrm{d}\phi\mathrm{d}\theta, \tag{9}$$

find a function $u \in C^2(\mathscr{E})$ that satisfies $r(\mathscr{E}) = \mathscr{G}$ and the Monge–Ampère type equation (8).

The condition $r(\mathscr{E}) = \mathscr{G}$ needs to be satisfied for equation (8) to have meaning. Alternatively, we can use the continuously differentiable bijection s to reformulate the problem in terms of a gradient set $\mathscr{F}$ and a function F on this set instead of the target set $\mathscr{G}$ and the luminous intensity G. Using the fact that s^{-1} exists, we define $\mathscr{F} := s^{-1}(\mathscr{G})$, and, using the differentiability of s, we define

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

for all $\boldsymbol{y} \in \mathscr{F}$. Note that $F \in C(\mathscr{F})$. Furthermore, using integration by substitution we see that (9) implies

$$\int_{\mathscr{E}} E(x, y)\mathrm{d}x\mathrm{d}y = \int_{\mathscr{F}} F(p, q)\mathrm{d}p\mathrm{d}q. \tag{10}$$

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

Inverse Reflector Problem. *Given a convex, closed and bounded light source $\mathscr{E}$ with strictly positive and bounded emittance $E \in C(\mathscr{E})$ and a closed gradient set $\mathscr{F}$ with strictly positive, bounded and bounded away from zero, function $F \in C(\mathscr{F})$ that satisfy (10), find a function $u \in C^2(\mathscr{E})$ that satisfies $\nabla u(\partial\mathscr{E}) = \partial\mathscr{F}$, $e^{ij}h_{ij}(u) > 0$ and the Monge–Ampère type equation*

$$\frac{\det(h_{ij}(u(\boldsymbol{x})))}{e(\boldsymbol{x})} = \frac{E(\boldsymbol{x})}{F(\nabla u(\boldsymbol{x}))}. \tag{11}$$

Note that we replaced the implicit boundary condition $\nabla u(\mathscr{E}) = \mathscr{F}$ with the more explicit boundary condition $\nabla u(\partial\mathscr{E}) = \partial\mathscr{F}$. The explicit boundary condition is better manageable numerically. We will show in Appendix B 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 $e^{ij}h_{ij}(u) = h^i_i(u) > 0$, demanding the trace (cf. Appendix (A.2)) 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. Thus, we conclude that a solution to the Inverse Reflector Problem as stated above and the same problem but with $\nabla u(\partial\mathscr{E}) = \partial\mathscr{F}$ replaced by $\nabla u(\mathscr{E}) = \mathscr{F}$ are truly equivalent as they both only admit strictly convex solutions and the boundary conditions are equivalent for strictly convex u.

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, pp. 96–98] 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. It is, however, not clear that, for all pairs $(\mathscr{E}, E)$ and $(\mathscr{F}, F)$, the unique weak solution of Brenier's theorem is twice continuously differentiable. Regularity of the solution for the Monge–Ampère equation (11) is a complicated matter and beyond the scope of this paper. For u to be $C^2(\mathscr{E})$, certainly continuity of E and F seems to be required (see for example [5]). However, even this requirement will not be met in all the numerical experiments we present in Section 4. Though also in these cases good numerical approximations are found.

3. Least-squares method in arbitrary orthogonal coordinates

In [13] Prins et al. proposed the LS method to solve the Inverse Reflector Problem in Cartesian coordinates. We introduce in this section the GLS method, i.e., the generalization of the LS method by Prins et al. [13] to arbitrary *orthogonal* coordinate systems. We remind the reader that a brief explanation of the notions of tensor calculus used can be found in Appendix A.

We assume an arbitrary coordinate system on the source $\mathscr{E}$ with orthogonal coordinates x^1, x^2, local orthogonal basis vectors $\boldsymbol{e}_1, \boldsymbol{e}_2$ and a metric $e_{ij} = (\boldsymbol{e}_i, \boldsymbol{e}_j)$ with corresponding norm $\|\cdot\|$. 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 $\boldsymbol{m} : \mathscr{E} \to \mathscr{F}$ representing ∇u such that:

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

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

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 $u \in C^2(\mathscr{E})$ 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.3), the Hessian of some function and therefore needs to be symmetric (Appendix A.3). 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 [21, p. 170]

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

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 [23, p. 551]. Thus we conclude that $\boldsymbol{m}$ equals the gradient of some function $u \in C^2(\mathscr{E})$ 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 $u \in C^2(\mathscr{E})$ is convex if and only if $\mathscr{E}$ is convex and the Hessian tensor $\boldsymbol{H}(u)$ is *positive semi-definite*, see for example [24, 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

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

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

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

Unfortunately, we can not demand positive definiteness, because, although every $u \in C^2(\mathscr{E})$ with positive definite Hessian tensor is strictly convex, not every strictly convex $u \in C^2(\mathscr{E})$ 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 map $\boldsymbol{m} : \mathscr{E} \to \mathscr{F}$ that satisfies*

$$\frac{\det(\nabla \hat{\boldsymbol{m}}(\boldsymbol{x}))}{e(\boldsymbol{x})} = \frac{E(\boldsymbol{x})}{F(\boldsymbol{m}(\boldsymbol{x}))}, \quad \boldsymbol{x} \in \mathscr{E}, \tag{12a}$$
$$\boldsymbol{m}(\partial\mathscr{E}) = \partial\mathscr{F}, \tag{12b}$$

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 $\mathscr{E}$ and $\mathscr{F}$, respectively, such that (10) is satisfied and F is bounded away from zero.

[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.

If u is a solution to the Inverse Reflector Problem, then u is strictly convex and $\boldsymbol{m} = \nabla u$ will be a solution to the Transport BVP. The reverse statement is not necessarily true because a solution $\boldsymbol{m}$ of the Transport BVP may be such that u satisfying $\boldsymbol{m} = \nabla u$ is convex but not strictly convex. The Transport BVP thus admits for a slightly larger solution class.

We numerically solve the Transport BVP by starting with an initial guess $\boldsymbol{m}^0$ and improve this initial guess in an iterative manner. One iteration consists of three stages. These three stages together give an improved mapping $\boldsymbol{m}^{n+1}$ from the current mapping $\boldsymbol{m}^n$. We now explain these three stages.

First, we approximate $\nabla\hat{\boldsymbol{m}}$ by a symmetric positive semi-definite tensor P by minimizing the functional

$$J_1(\boldsymbol{m}, \boldsymbol{P}) := \frac{1}{2}\int_{\mathscr{E}} \|\nabla\hat{\boldsymbol{m}} - \boldsymbol{P}\|^2 \,\mathrm{d}A, \tag{13a}$$

over the space

$$\mathscr{P}(\boldsymbol{m}) := \left\{ \boldsymbol{P} \in \boldsymbol{T}^0_2(\mathscr{E})_{C^1} \mid \det(p_{ij}(\boldsymbol{x})) = \frac{e(\boldsymbol{x})E(\boldsymbol{x})}{F(\boldsymbol{m}(\boldsymbol{x}))},\ \boldsymbol{P}(\boldsymbol{x}) \text{ is spsd} \right\}, \tag{13b}$$

where "spsd" stands for symmetric positive semi-definite and $\boldsymbol{P} = p_{ij}\boldsymbol{e}^i \otimes \boldsymbol{e}^j$. Furthermore, we use $\boldsymbol{T}^0_2(\mathscr{E})_{C^1}$ to denote the space of continuously differentiable tensor fields of contravariant rank 0 and covariant rank 2 (Appendix A.2). It seems as if we demand more smoothness than necessary, because Transport BVP and (13a) suggest that $\boldsymbol{P}$ only needs to be continuous for continuous E and F. However, in one of the minimization procedures that follows we need $\nabla\hat{\boldsymbol{m}}$ to be continuously differentiable and therefore we also need $\boldsymbol{P}$ to be continuously differentiable.

The norm in functional (13a) is defined in the following way. Let $\boldsymbol{A}, \boldsymbol{B} \in \boldsymbol{T}^0_2(\mathscr{E})$, 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

$$\boldsymbol{A} : \boldsymbol{B} := e^{ik}e^{jl}a_{ij}b_{kl} = a^{ij}b_{ij}, \tag{14}$$

defines an inner product on $\boldsymbol{T}^0_2(\mathscr{E})$. This inner product on $\boldsymbol{T}^0_2(\mathscr{E})$ 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 $J_1 = 0$, $\boldsymbol{m}$ will satisfy (12a) and $\nabla\hat{\boldsymbol{m}}$ will be symmetric positive semi-definite.

Secondly, we approximate the boundary value of $\boldsymbol{m}$ at $\partial\mathscr{E}$ by a vector field $\boldsymbol{b}$ that exactly satisfies the boundary condition (12b). This is done by minimizing

$$J_B(\boldsymbol{m}, \boldsymbol{b}) := \frac{1}{2}\int_{\partial\mathscr{E}} \|\boldsymbol{m} - \boldsymbol{b}\|^2 \mathrm{d}s, \tag{15a}$$

over the space

$$\mathscr{B} := \left\{ \boldsymbol{b} = \bar{\boldsymbol{b}}\big|_{\partial\mathscr{E}} \mid \bar{\boldsymbol{b}} \in \boldsymbol{T}^1_0(\mathbb{R}^2)_C, \boldsymbol{b}(x) \in \partial\mathscr{F} \right\}, \tag{15b}$$

for an arc-length parameterization of the boundary. Analogously to the functional J_1 we notice that if $J_B = 0$, $\boldsymbol{m}$ satisfies (12b).

Finally, to find the improved mapping $\boldsymbol{m}^{n+1}$ we minimize J_1 and J_B simultaneously. To do so we define a third functional:

$$J(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b}) := \alpha J_1(\boldsymbol{m}, \boldsymbol{P}) + (1-\alpha) J_B(\boldsymbol{m}, \boldsymbol{b}) \tag{16}$$

with $\alpha \in (0, 1)$. We minimize this functional for $\boldsymbol{m}$ over the space $\mathscr{M} := \boldsymbol{T}^1_0(\mathscr{E})_{C^2}$. Thus, to determine $\boldsymbol{m}^{n+1}$ three stages are performed subsequently:

$$\boldsymbol{b}^{n+1} = \operatorname*{argmin}_{\boldsymbol{b}\in\mathscr{B}} J_B(\boldsymbol{m}^n, \boldsymbol{b}), \tag{17a}$$

$$\boldsymbol{P}^{n+1} = \operatorname*{argmin}_{\boldsymbol{P}\in\mathscr{P}(\boldsymbol{m}^n)} J_1(\boldsymbol{m}^n, \boldsymbol{P}), \tag{17b}$$

$$\boldsymbol{m}^{n+1} = \operatorname*{argmin}_{\boldsymbol{m}\in\mathscr{M}} J(\boldsymbol{m}, \boldsymbol{P}^{n+1}, \boldsymbol{b}^{n+1}). \tag{17c}$$

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. The minimization problems (17c) will then be translated to discrete problems on this grid.

The first minimization step, step (17a), can be performed in an efficient point-wise 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 (17b) 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 (17c) is covered in great detail, because this minimization problem becomes substantially more involved for arbitrary coordinate systems. We start with minimization problem (17b).

3.1. Minimization of J_1

The integrand of J_1 does not contain derivatives of $\boldsymbol{P}$, therefore we can carry out the minimization for each grid point $\boldsymbol{x} \in \mathscr{E}$ individually. For each grid point $\boldsymbol{x} \in \mathscr{E}$ 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 $\nabla_{\boldsymbol{e}_i} m_j$, i.e., the difference of the value in two neighboring grid points in the $\boldsymbol{e}_i$ direction divided by the distance between those two points. 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

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

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 $\big(e^{ij} p_{jk}\big)$ is positive semi-definite. Recall that symmetric 2×2 matrices are positive semi-definite if and only if their trace and determinant are both nonnegative. However, the matrix $(e^{ij} p_{jk})$ is not symmetric, because

$$\big(e^{ij} p_{jk}\big) = \begin{pmatrix} e^{11} p_{11} & e^{11} p_{12} \\ e^{22} p_{12} & e^{22} p_{22} \end{pmatrix}, \tag{18}$$

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

$$T\big(e^{ij} p_{jk}\big)T^{-1} = \begin{pmatrix} e^{11} p_{11} & p_{12}/\sqrt{e} \\ p_{12}/\sqrt{e} & e^{22} p_{22} \end{pmatrix}. \tag{19}$$

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

$$\mu_\pm = \frac{1}{2e}\left(e_{22} p_{11} + e_{11} p_{22} \pm \sqrt{\big(e_{22} p_{11} + e_{11} p_{22}\big)^2 - 4e \det\big(p_{ij}\big)} \right), \tag{20}$$

which are both real since the matrix $T\big(e^{ij} p_{jk}\big)T^{-1}$ is symmetric. It is a familiar result that a matrix is positive semi-definite if and only if its eigenvalues are nonnegative. The matrix $\big(e^{ij} p_{jk}\big)$ is positive semi-definite if and only if the matrix in (19) is positive semi-definite. The matrix in (19) 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 (19) 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 therefore 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

$$\bar{p}_{11} := e^{11} p_{11}, \qquad \bar{p}_{12} := p_{12}/\sqrt{e}, \qquad \bar{p}_{22} := e^{22} p_{22}, \tag{21a}$$

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

We can give a more convenient reformulation of the minimization problem in terms of these variables. Moreover, we also drop the 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 then reads 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 (21b), 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

$$H(\bar{P}) := \frac{1}{2}\|\bar{D} - \bar{P}\|_F^2, \tag{22}$$

under the constraints $\det(\bar{P}) = E/F$ *and* $\mathrm{tr}(\bar{P}) \geq 0$, *where the norm used in (22) is the Frobenius norm for matrices, defined as* $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$ *for a matrix* $A = (a_{ij})$.

From the relations (21a) 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*

$$\Lambda(\bar{P}; \lambda) = H(\bar{P}) + \lambda\left(\det \bar{P} - \frac{E}{F}\right). \tag{23}$$

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

$$\bar{p}_{11} + \lambda\bar{p}_{22} = \bar{d}_{11}, \tag{24a}$$
$$\lambda\bar{p}_{11} + \bar{p}_{22} = \bar{d}_{22}, \tag{24b}$$
$$(1-\lambda)\bar{p}_{12} = \bar{d}_{12}, \tag{24c}$$
$$\bar{p}_{11}\bar{p}_{22} - \bar{p}_{12}^2 = E/F. \tag{24d}$$

In the Lagrange function (23) the condition $\mathrm{tr}(\bar{P}) \geq 0$ has not been taken into account, hence a solution of (24a)–(24d) might have $\mathrm{tr}(\bar{P}) < 0$. In what follows, we will show that there always exists a solution to (24a)–(24d) such that $\mathrm{tr}(\bar{P}) \geq 0$.

Let us now give a geometric interpretation of the minimizers of the Lagrange function. The minimizers correspond to a joint tangent plane of a hyperboloid and an ellipsoid. We introduce the iso-surfaces $H(\bar{P}) = C_H$. 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

$$\left(\frac{\bar{p}_{11} - \bar{d}_{11}}{\sqrt{2C_H}}\right)^2 + \left(\frac{\bar{p}_{12} - \bar{d}_{12}}{\sqrt{C_H}}\right)^2 + \left(\frac{\bar{p}_{22} - \bar{d}_{22}}{\sqrt{2C_H}}\right)^2 = 1. \tag{25}$$

Equation (25) 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 the new coordinate axes. We perform the rotation given by

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

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

$$\left(\frac{p_1}{\sqrt{2E/F}}\right)^2 - \left(\frac{p_2}{\sqrt{2E/F}}\right)^2 + \left(\frac{p_3}{\sqrt{E/F}}\right)^2 = -1. \tag{26}$$

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 $\mathrm{tr}(\bar{P}) > 0$ and the extremum of the sheet with $\mathrm{tr}(\bar{P}) < 0$ is both $\sqrt{2E/F}$. Equation (25) 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 (Fig. 5) 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.

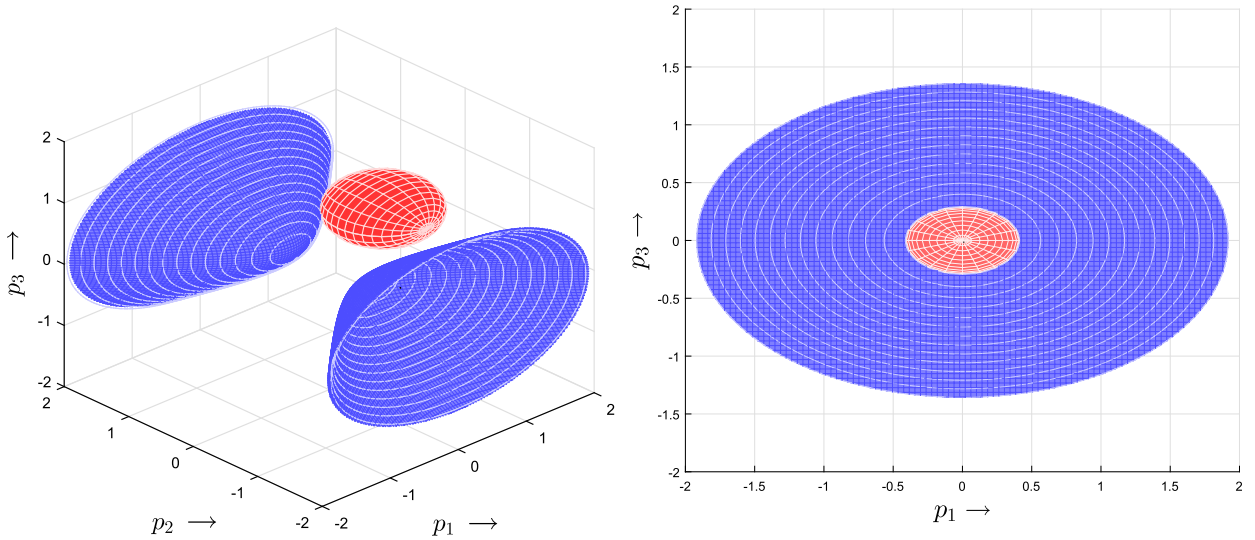

Fig. 5. An example of an ellipsoidal iso-surface of H and a hyperboloid are shown from two different perspectives for two instances of $\bar{D}$. We see that the principal p_1- and p_3-axis have the same proportion for the hyperboloid and the ellipsoid.

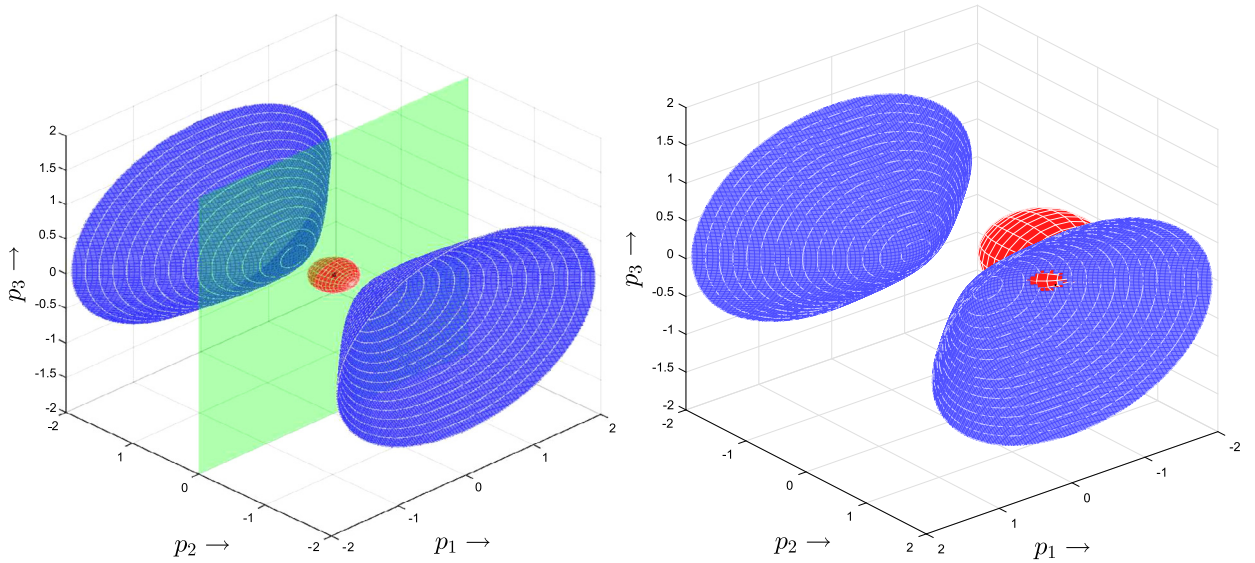

Fig. 6. On the left side the two sheets of the hyperboloid and the dividing plane $p_2 = \mathrm{tr}(\bar{P})/\sqrt{2} = 0$ are shown. On the right side an example of an ellipsoid which is tangent to the hyperboloid with $\mathrm{tr}(\bar{P}) > 0$ is shown. Some red of the ellipsoid can be seen through the hyperboloid. This point is the minimizer.

The local minimizers of the Lagrange function (23) are exactly the points where an iso-surface of H is tangent to the hyperboloid. The plane $p_2 = \mathrm{tr}(\bar{P})/\sqrt{2} = 0$ 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 $\mathrm{tr}(\bar{P}) > 0$ are actual minimizers of Minimization Problem. In Fig. 6 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 (24). 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 (24a)–(24d).

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 (24a)–(24c). Doing this we obtain

$$\bar{p}_{11} = \frac{\lambda \bar{d}_{22} - \bar{d}_{11}}{\lambda^2 - 1}, \qquad \bar{p}_{12} = \frac{\bar{d}_{12}}{1 - \lambda}, \qquad \bar{p}_{22} = \frac{\lambda \bar{d}_{11} - \bar{d}_{22}}{\lambda^2 - 1}. \tag{27}$$

However, these equations only hold if $\lambda \neq \pm 1$. From (24a)–(24c) 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 (27) holds. We will now treat the three different cases in turn, starting out with the general case.

Lemma 3. *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 (27), where λ is given by one of the following expressions:*

$$\begin{aligned} \lambda_i &= -\sqrt{\frac{y}{2}} + (-1)^i \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} + \frac{a_1}{2a_4\sqrt{2y}}}, \qquad i = 1, 2, \\ \lambda_i &= \sqrt{\frac{y}{2}} + (-1)^i \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} - \frac{a_1}{2a_4\sqrt{2y}}}, \qquad i = 3, 4. \end{aligned} \tag{28}$$

In (28) y is given by the following two sets of equations:

$$\begin{aligned} y &= A + \frac{Q}{A} - \frac{b_2}{3}, \qquad A = -\operatorname{sgn}(R)\left(|A| + \sqrt{R^2 - Q^3}\right)^{1/3}, \\ R &= \frac{2b_2^3 - 9b_1b_2 + 27b_0}{54}, \qquad Q = \frac{b_2^3 - 3b_1}{9}, \end{aligned} \tag{29}$$

and

$$\begin{aligned} a_4 &= \frac{E}{F}, \qquad a_2 = -2a_4 - \det(\bar{D}), \qquad a_1 = \|\bar{D}\|^2, a_0 = a_4 - \det(\bar{D}), \\ b_0 &= -\frac{a_1^2}{8a_4^2}, \qquad b_1 = \frac{a_2^2 - 4a_0a_4}{4a_4^2}, \qquad b_2 = \frac{a_2}{a_4}. \end{aligned} \tag{30}$$

At least one of the four choices for λ is such that the requirement $\operatorname{tr}(\bar{P}) > 0$ is satisfied by (27).

Proof. Substituting the expressions (27) in (24d) we obtain the following quartic equation $\Pi(\lambda) := a_4\lambda^4 + a_2\lambda^2 + a_1\lambda + a_0 = 0$, where the coefficients are as given in (30). In [13] it is shown that this polynomial admits the four solutions (28). 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 (24) it follows that $\operatorname{tr}(\bar{P}) = \operatorname{tr}(\bar{D})/(1 + \lambda)$. This shows that for one of the two real roots $\operatorname{tr}(\bar{P}) > 0$, while for the other real root $\operatorname{tr}(\bar{P}) < 0$.

We now have established the fact that one of the four roots λ in (28) gives the 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 (27), with λ given by one of the real roots of (28). 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 4. *When $\bar{d}_{11} = -\bar{d}_{22}$, the global minimizer to Minimization Problem is given by*

$$\bar{p}_{11} = \frac{1}{2}\left(\bar{d}_{11} + \sqrt{\bar{d}_{11}^2 + 4E/F + \bar{d}_{12}^2}\right), \qquad \bar{p}_{12} = \frac{\bar{d}_{12}}{2}, \qquad \bar{p}_{22} = \bar{p}_{11} - \bar{d}_{11}. \tag{31}$$

Proof. When $\bar{d}_{11} = -\bar{d}_{22}$, the Lagrange conditions (24a) and (24b) 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 (24d) 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 (24a)–(24d) now simplify to

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

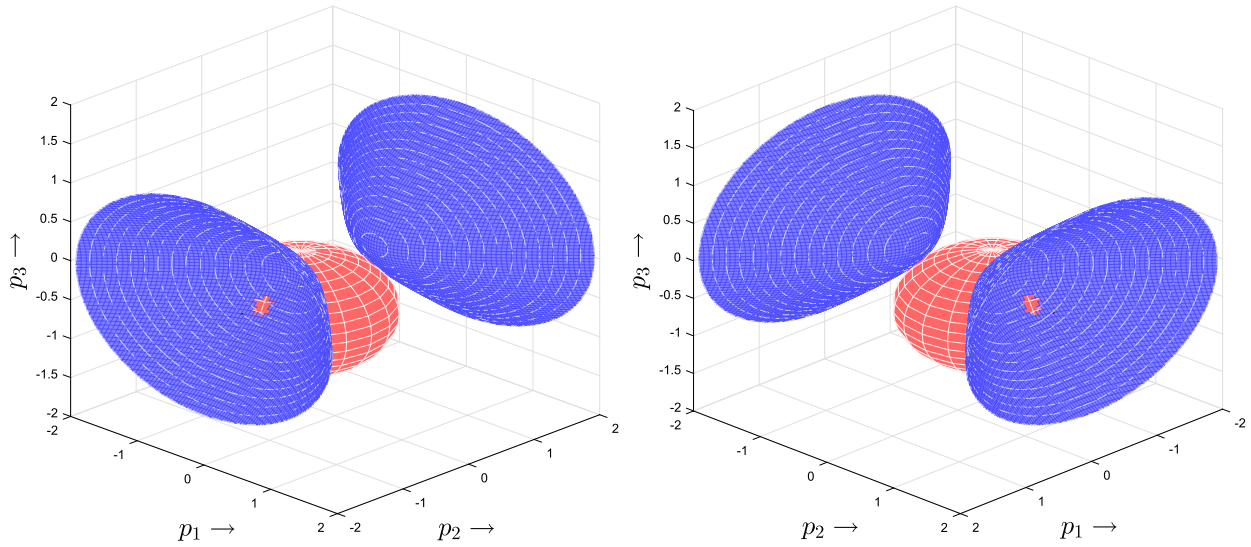

Fig. 7. These figures correspond to Lemma 4. The ellipsoid is centered around the same point $(d_1 = (\bar{d}_{11} - \bar{d}_{22})/\sqrt{2} = \sqrt{2}\bar{d}_{11}, d_2 = (\bar{d}_{11} + \bar{d}_{22})/\sqrt{2} = 0, d_3 = \bar{d}_{12})$ in both figures. This results in two local minima with the same function value for H. In the figure on the left we see the minimum on the hyperboloid sheet with $\mathrm{tr}(\bar{P}) < 0$, which has to be discarded, and in the figure on the right we see the minimum on the sheet with $\mathrm{tr}(\bar{P}) > 0$.

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 (31). In Fig. 7 these findings are illustrated. □

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

Lemma 5. *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*

$$\bar{p}_{11} = \sqrt{E/F}, \quad \bar{p}_{12} = 0, \quad \bar{p}_{22} = \sqrt{E/F}, \tag{32}$$

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

$$\bar{p}_{11} \in \left[\frac{\bar{d}_{11} - l}{2}, \frac{\bar{d}_{11} + l}{2}\right], \quad \bar{p}_{12} = \pm\sqrt{\bar{d}_{11}\bar{p}_{11} - \bar{p}_{11}^2 - \frac{E}{F}}, \quad \bar{p}_{22} = \bar{d}_{11} - \bar{p}_{11}, \tag{33}$$

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

Proof. In the case that $\bar{d}_{11} = \bar{d}_{22}$ and $\bar{d}_{12} = 0$, Lagrange conditions (24a) and (24b) 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 $\bar{p}_{11} = \bar{p}_{22}$. Let us first deal with the case $\lambda \neq 1$. When $\lambda \neq 1$, the Lagrange conditions (24c) and (24d) 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 $\mathrm{tr}(\bar{P}) > 0$ is only satisfied when the plus sign holds, hence we find the minimizer given by (32).

Now suppose that $\lambda = 1$. Lagrange condition (24a) implies $\bar{p}_{22} = \bar{d}_{11} - \bar{p}_{11}$ and from Lagrange condition (24d) 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 $\bar{p}_{11} \in [(\bar{d}_{11} - l)/2, (\bar{d}_{11} + l)/2]$, where $l = \sqrt{\bar{d}_{11}^2 - 4E/F}$. This gives us

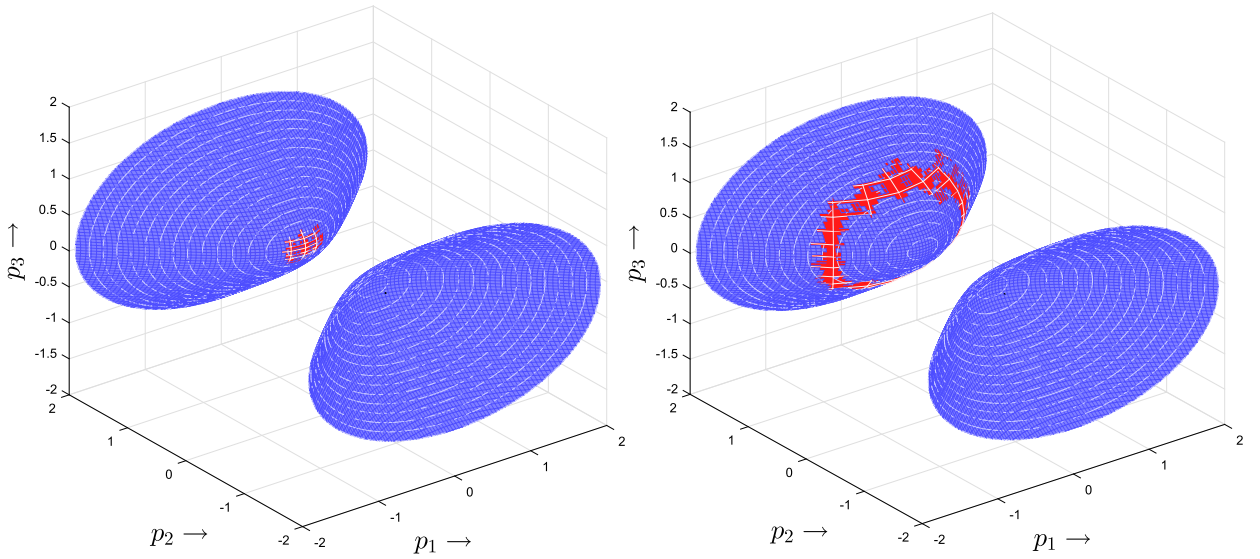

Fig. 8. These figures correspond to the minimizers in Lemma 5. The ellipsoid is located behind the hyperboloid. We see the sheet of the hyperboloid with $\mathrm{tr}(\bar{P}) > 0$ on the left. The ellipsoid is centered around the point $(d_1 = (\bar{d}_{11} - \bar{d}_{22})/\sqrt{2} = 0, d_2 = (\bar{d}_{11} + \bar{d}_{22})/\sqrt{2} = \sqrt{2}\bar{d}_{11}, d_3 = \bar{d}_{12} = 0)$. 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.

the continuum of minimizers (33). However, $\bar{p}_{11}$ is only real if $|\bar{d}_{11}| \geq 2\sqrt{E/F}$. Moreover, because $\mathrm{tr}(\bar{P}) = \bar{d}_{11}$, we see that $\mathrm{tr}(\bar{P}) > 0$ 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 (32). To decide for $\bar{d}_{11} \geq 2\sqrt{E/F}$ whether the global minimizer is given by (32) or by an element of the continuum (33), 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

$$H_{\mathrm{cont}} = \frac{1}{2}\left\| \begin{pmatrix} \bar{d}_{11} - \bar{p}_{11} & \sqrt{\bar{d}_{11}\bar{p}_{11} - \bar{p}_{11}^2 - E/F} \\ \sqrt{\bar{d}_{11}\bar{p}_{11} - \bar{p}_{11}^2 - E/F} & \bar{p}_{11} \end{pmatrix} \right\|^2 = \frac{\bar{d}_{11}^2}{2} - \frac{E}{F}.$$

On the other hand, for the local minimizer at the extremum of the hyperboloid, given by (32), we have

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

This implies that $H_{\mathrm{cont}} - H_{\mathrm{ext}} = -\bar{d}_{11}^2/2 + 2\bar{d}_{11}\sqrt{E/F} - 2E/F$. This polynomial in $\bar{d}_{11}$ has its maximal value in $\bar{d}_{11} = 2\sqrt{E/F}$ 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 (33). □

In Fig. 8 examples of the results from Lemma 5 are shown. Recall that the extrema of the two sheets of the hyperboloid are located at

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

Thus Lemma 5 implies the following.

- The point $(p_1, p_2, p_3) = (0, \sqrt{2E/F}, 0)$ is the global minimizer when $\bar{d}_2 < 2\sqrt{2E/F}$, $\bar{d}_{11} = \bar{d}_{22}$ and $\bar{d}_{12} = 0$, i.e., when the center of the ellipsoid is located in $(0, p_2, 0)$, where $\bar{p}_2 = \sqrt{2}\bar{d}_{11} < 2\sqrt{2E/F}$, in other words, 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 $\mathrm{tr}(\bar{P}) > 0$, or if the center of the ellipsoid is located beneath the plane $p_2 = \mathrm{tr}(\bar{P})/\sqrt{2} = 0$, then the global minimizer is given by the extremum of the upper sheet of the hyperboloid. This is depicted in the graph on the left in Fig. 8.
- If $\bar{d}_{11} = \bar{d}_{22}$, $\bar{d}_{12} = 0$, the center of the ellipsoid is located above the plane given by $p_2 = \mathrm{tr}(\bar{P})/\sqrt{2} = 0$ 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 Fig. 8.

Summarizing, we have proved the following theorem.

Theorem 2. *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 (27), with* λ *given by one of the four possibilities in (28). At least two of the* λ*'s in (28) 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 given by (31). 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 (32). If* $\bar{d}_{11} \geq 2\sqrt{E/F}$, *there is a whole continuum of solutions to Minimization Problem given by (33).*

3.2. Minimization of J

In this section we focus on the last step of the least-squares method, i.e. (17c). We will minimize the functional J, defined in (16), for $\boldsymbol{m} \in \mathscr{M}$, while keeping $\boldsymbol{P}$ and $\boldsymbol{b}$ constant. Again we do this for an arbitrary coordinate system on $\mathscr{E}$ 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, pp. 142–144].

3.2.1. Derivation of the 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 $\boldsymbol{\eta} \in \mathscr{M}$. 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 derivative we find

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

We will now need the following convenient property of inner product on $\boldsymbol{T}^0_2(\mathscr{E})$ as defined in (14). Let $\boldsymbol{A}, \boldsymbol{B} \in \boldsymbol{T}^0_2(\mathscr{E})$, then we have

$$\|\boldsymbol{A} + \boldsymbol{B}\|^2 = e^{ik}e^{jl}(a_{ij} + b_{ij})(a_{kl} + b_{kl}) = \|\boldsymbol{A}\|^2 + 2\boldsymbol{A} : \boldsymbol{B} + \|\boldsymbol{B}\|^2.$$

Applying this property to $\|\varepsilon \nabla \hat{\boldsymbol{\eta}} + \nabla \hat{\boldsymbol{m}} - \boldsymbol{P}\|^2$, with $\boldsymbol{A} = \varepsilon \nabla \hat{\boldsymbol{\eta}}$ and $\boldsymbol{B} = \nabla \hat{\boldsymbol{m}} - \boldsymbol{P}$, we obtain

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

In the same fashion, using the fact that

$$\|\varepsilon\boldsymbol{\eta} + \boldsymbol{m} - \boldsymbol{b}\|^2 - \|\boldsymbol{m} - \boldsymbol{b}\|^2 = \varepsilon^2 \|\boldsymbol{\eta}\|^2 + 2\varepsilon(\boldsymbol{\eta}, \boldsymbol{m} - \boldsymbol{b}), \tag{34}$$

we find the Gâteaux derivative of J_{B} to be

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

Combining the results for J_{I} and J_{B} we find that

$$\forall \boldsymbol{\eta} \in \mathscr{M} : \ \alpha \int_{\mathscr{E}} \nabla \hat{\boldsymbol{\eta}} : (\nabla \hat{\boldsymbol{m}} - \boldsymbol{P}) \mathrm{d}A + (1-\alpha) \int_{\partial \mathscr{E}} (\boldsymbol{\eta}, \boldsymbol{m} - \boldsymbol{b}) \mathrm{d}s = 0. \tag{35}$$

In order to proceed we will rewrite the integrands in terms of their components. For the first integrand in (35) we have by (14)

$$\nabla \hat{\boldsymbol{\eta}} : (\nabla \hat{\boldsymbol{m}} - \boldsymbol{P}) = D_j \eta_i (D^j m^i - p^{ij}),$$

where $D_j \eta_i$ are the components of the covariant derivative of $\hat{\boldsymbol{\eta}}$ (Appendix A.3), $D^j = e^{ij} D_i$ and we used bijection (A.1). The product rule implies

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

If we integrate the first term and apply the *Green's Theorem* [22, p. 135] we find

$$\int_{\mathscr{E}} D_j\big(\eta_i(D^jm^i - p^{ij})\big)\mathrm{d}A = \int_{\partial\mathscr{E}} (D^jm^i - p^{ij})\eta_i n_j \mathrm{d}s,$$

where n_j are the covariant components of the outward unit normal vector on the boundary $\partial\mathscr{E}$ and the orientation on $\partial\mathscr{E}$ is the one induced by $\mathscr{E}$ [22, p. 119]. It follows that

$$\int_{\mathscr{E}} \nabla\hat{\boldsymbol{\eta}} : (\nabla\hat{\boldsymbol{m}} - \boldsymbol{P})\mathrm{d}A = \int_{\partial\mathscr{E}} (D^jm^i - p^{ij})\eta_i n_j \mathrm{d}s - \int_{\mathscr{E}} D_j(D^jm^i - p^{ij})\eta_i \mathrm{d}A.$$

Combining this result with identity (35) we obtain

$$0 = \int_{\partial\mathscr{E}} \big[\alpha(D^jm^i - p^{ij})n_j + (1-\alpha)(m^i - b^i)\big]\eta_i \,\mathrm{d}s - \alpha\int_{\mathscr{E}} D_j(D^jm^i - p^{ij})\eta_i \mathrm{d}A,$$

for all $\boldsymbol{\eta} \in \mathscr{M}$. Invoking the *Fundamental Lemma of Calculus of Variations* [25, p. 185] we find from this the boundary value problem

$$D_jD^jm^i = D_jp^{ij} \qquad \text{in } \mathscr{E}, \tag{36a}$$

$$\alpha(D^jm^i)n_j + (1-\alpha)m^i = \alpha p^{ij}n_j + (1-\alpha)b^i \qquad \text{on } \partial\mathscr{E}. \tag{36b}$$

The solution of boundary value problem (36b) will minimize J for given $\boldsymbol{P}$ and $\boldsymbol{b}$. Note that (36b) are vector equations. The term $D_jD^jm^i$ is the so-called *vector Laplacian* [26, p. 91], which can be written in terms of vector calculus operators as $\nabla(\nabla\cdot\boldsymbol{m}) - \nabla\times(\nabla\times\boldsymbol{m})$. In Cartesian coordinates $D_jD^jm^i = (\partial_1^2 + \partial_2^2)m^i$, 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 $[D_jD^jm^i]_{i=1,2}$ 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 (36b) decouple. Let us denote the standard Cartesian basis vectors by $\boldsymbol{e}_x$ and $\boldsymbol{e}_y$, define

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

and write $\boldsymbol{m} = m^x\boldsymbol{e}_x + m^y\boldsymbol{e}_y$. With the use of this definition we can rewrite $(D_jp^{ij})_{i=1}$ as $\nabla\cdot\boldsymbol{p}_x$ and $(D_jp^{ij})_{i=2}$ as $\nabla\cdot\boldsymbol{p}_y$. From this we see that in Cartesian coordinates (36b) reduces to the decoupled set of equations

$$\begin{aligned} \Delta m^x &= \nabla\cdot\boldsymbol{p}_x && \text{in } \mathscr{E}, \\ \alpha(\nabla m^x, \boldsymbol{n}) + (1-\alpha)m^x &= \alpha(\boldsymbol{p}_x, \boldsymbol{n}) + (1-\alpha)b^x && \text{on } \partial\mathscr{E}, \end{aligned} \tag{37a}$$

$$\begin{aligned} \Delta m^y &= \nabla\cdot\boldsymbol{p}_y && \text{in } \mathscr{E}, \\ \alpha(\nabla m^y, \boldsymbol{n}) + (1-\alpha)m^y &= \alpha(\boldsymbol{p}_y, \boldsymbol{n}) + (1-\alpha)b^y && \text{on } \partial\mathscr{E}. \end{aligned} \tag{37b}$$

The boundary value problems (37a) and (37b) 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 (36b) does depend on the choice of basis for polar coordinates, because (36b) is a vector equation. Thus, we find for an anholonomic basis different expressions than for a holonomic basis (Appendix A.2).

To derive the boundary value problem in polar coordinates, let us first elaborate the components of the covariant derivatives appearing in (36b). We start with the vector Laplacian. By the fact that $D_i(e^{jk}) = 0$ (Appendix A.3) it follows by (A.5c) that $D_jD^jm^i = e^{jk}\big(\nabla_{\boldsymbol{e}_j}(D_km^i) - \Gamma^l_{kj}D_lm^i + \Gamma^i_{lj}D_km^l\big)$ and by (A.4a) that $D_km^i = \nabla_{\boldsymbol{e}_k}m^i + \Gamma^i_{lk}m^l$, hence

$$D_jD^jm^i = e^{jk}\big(\nabla_{\boldsymbol{e}_j}\nabla_{\boldsymbol{e}_k}m^i + \nabla_{\boldsymbol{e}_j}(\Gamma^i_{lk})m^l + \Gamma^i_{lk}\nabla_{\boldsymbol{e}_j}m^l - \Gamma^l_{kj}\nabla_{\boldsymbol{e}_l}m^i - \Gamma^l_{kj}\Gamma^i_{sl}m^s + \Gamma^i_{lj}\nabla_{\boldsymbol{e}_k}m^l + \Gamma^i_{lj}\Gamma^l_{sk}m^s\big). \tag{38}$$

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

[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.

$$D_j p^{ij} = \nabla_{\boldsymbol{e}_j} p^{ij} + \Gamma^i_{lj} p^{lj} + \Gamma^j_{lj} p^{il}. \tag{39}$$

Similarly, we find for the normal derivative of $\boldsymbol{m}$ in (36b)

$$(D^j m^i) n_j = e^{jk} (D_k m^i) n_j = e^{jk} \left(\nabla_{\boldsymbol{e}_k} m^i + \Gamma^i_{lk} m^l \right) n_j. \tag{40}$$

We use (38)–(40) to expand the boundary value problem (36b) in polar coordinates. We consider the anholonomic orthonormal basis ($e_{ij} = \delta_{ij}$), because this is the basis we used in the implementation. The only nonzero Christoffel symbols in the anholonomic basis are $\Gamma^r_{\theta\theta} = -r^{-1}$ and $\Gamma^\theta_{r\theta} = r^{-1}$ [20, p. 218]. After doing the tedious calculations of determining the coordinate system specific expressions of the various terms in (38) we find

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

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 expressions for the normal derivative of $\boldsymbol{m}$ from (40) and find

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

We define

$$\boldsymbol{p}_r = \begin{pmatrix} p^{rr} \\ p^{r\theta} \end{pmatrix} \quad \text{and} \quad \boldsymbol{p}_\theta = \begin{pmatrix} p^{r\theta} \\ p^{\theta\theta} \end{pmatrix}, \tag{41}$$

and collect all the different terms and find that for polar coordinates with an orthonormal basis (36b) is given by

$$\begin{aligned}
\Delta m^r - \frac{1}{r^2}\left(m^r + 2\frac{\partial m^\theta}{\partial \theta} \right) &= \nabla \cdot \boldsymbol{p}_r - \frac{1}{r} p^{\theta\theta} && \text{in } \mathscr{E}, \\
\alpha(\nabla m^r, \boldsymbol{n}) - \alpha\frac{m^\theta n^\theta}{r} + (1-\alpha) m^r &= \alpha(\boldsymbol{p}_r, \boldsymbol{n}) + (1-\alpha) b^r && \text{on } \partial\mathscr{E},
\end{aligned} \tag{42}$$

and

$$\begin{aligned}
\Delta m^\theta - \frac{1}{r^2}\left(m^\theta - 2\frac{\partial m^r}{\partial \theta} \right) &= \nabla \cdot \boldsymbol{p}_\theta + \frac{1}{r} p^{r\theta} && \text{in } \mathscr{E}, \\
\alpha(\nabla m^\theta, \boldsymbol{n}) + \alpha\frac{m^r n^\theta}{r} + (1-\alpha) m^\theta &= \alpha(\boldsymbol{p}_\theta, \boldsymbol{n}) + (1-\alpha) b^\theta && \text{on } \partial\mathscr{E},
\end{aligned} \tag{43}$$

where Δ, $\nabla\cdot$ and ∇ are the familiar Laplace, divergence and gradient operator in polar coordinates with orthonormal basis [23, pp. 542–543]. The equations (42) and (43) are coupled.

In the implementation of the GLS method we solve these two boundary value problems by using the standard second order central difference discretization. This provides us with a linear system, the matrix of which does not change during an entire run of the GLS method. This allows us to determine the LU-decomposition before the start of the algorithm and we can use this to solve the linear system when necessary.

To deal with the fact that, in polar coordinates, (42) and (43) are coupled, we will iterate between the two. Starting with (42) we keep m^θ fixed and solve for m^r. Next we keep m^r fixed and solve (43) for m^θ. We stop this iterative procedure when $J^{(n+1,i)} < c J^{(n)}$, where n is the outer iteration count of (17c) and i is the inner iteration count, or, when the number of inner iterations 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 (17c), the outer iterative procedure might converge faster. However, demanding far more precision in (17c) 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 $J^{(n+1,i)} < c J^{(n)}$ is a too severe requirement. This will come into play in the final part of the iteration sequence. We have no proof of convergence, but in practice this procedure always converged in a few iterations, because the mapping $\boldsymbol{m}^n$ provides a very good initial guess for (17c). 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. Alternatively, the complete coupled system can be solved at once. However, this is likely to be more expensive.

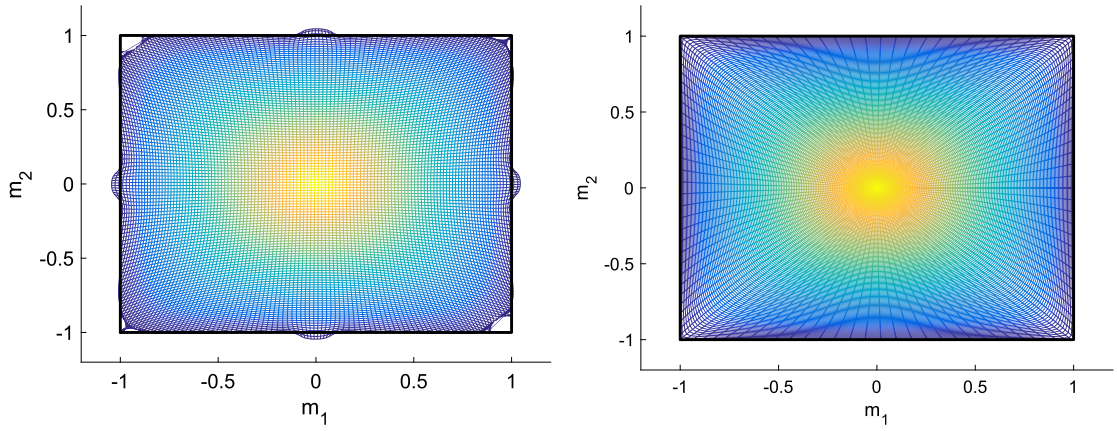

Fig. 9. The resulting mapping: on the left for Cartesian coordinates and on the right for polar coordinates. For both a 500×500 grid is used and $\alpha = 0.2$. We see the grid as it is mapped on $\mathscr{F}_S$. Grid points that initially had the same distance to the center of $\mathscr{E}$ have the same color. Bright yellow corresponds to points in the center of $\mathscr{E}$ and dark blue corresponds to points on $\partial\mathscr{E}$.

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 Prins et al. [13] to arbitrary coordinate systems. We remarked earlier 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 $u : \mathscr{E} \to (0, \infty)$ with gradient equal to $\boldsymbol{m}$ in an L^2-sense, hence we will search for u that minimizes

$$I(u) := \int_{\mathscr{E}} \|\nabla u - \boldsymbol{m}\|^2 \mathrm{d}A.$$

After a derivation very similar to the one by which we arrived at boundary value problem (36b), which we leave out for brevity, we obtain the Poisson problem

$$\begin{aligned} \Delta u &= \nabla \cdot \boldsymbol{m} && \text{in } \mathscr{E}, \\ (\nabla u, \boldsymbol{n}) &= (\boldsymbol{m}, \boldsymbol{n}) && \text{on } \partial\mathscr{E}. \end{aligned}$$

It is this problem that we solve to find the reflector surface for the problems presented in the next section. We discretize this Poisson problem 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 three test cases. In the first test case we compare the method in polar coordinates with the method in Cartesian coordinates as presented in [13], in the second test case we test the performance of the method for a source with non-uniform emittance and in the third test case we investigate the performance of the method in polar coordinates for a complex problem with a desired light output with a lot of contrast.

For all test cases we take as source domain $\mathscr{E}$ the unit disk. In the first and last test case we use a source with uniform emittance E_U and in the second test case we use a Gaussian emittance E_G. We choose these sources, because they frequently occur in lighting systems. For the first and second test case we take as the target $\mathscr{F}_S = [-1, 1] \times [-1, 1]$ with a uniform intensity function F_S. For the third test we have determined the pair $(\mathscr{F}_E, F_E)$ such that an intensity pattern corresponding to the lithograph by M. C. Escher (Fig. 1) is projected on a screen in the far-field. We normalize the source emittances and target intensities such that (10) holds.

4.1. From a circle to a square: uniform emittance

For the first test the mapping is presented in Fig. 9. In this figure we see that near the boundary $\partial\mathscr{F}_S$ the method in Cartesian coordinates has great difficulties. This results from the implementation where as actual source the smallest bounding box of $\mathscr{E}$ is used with emittance zero in the points outside $\mathscr{E}$ (see [13] for details). In the polar coordinate method the grid perfectly aligns with $\partial\mathscr{E}$. In Fig. 9 it can be seen that now all the difficulties at $\partial\mathscr{F}_S$ are resolved.

Fig. 10 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 (11),

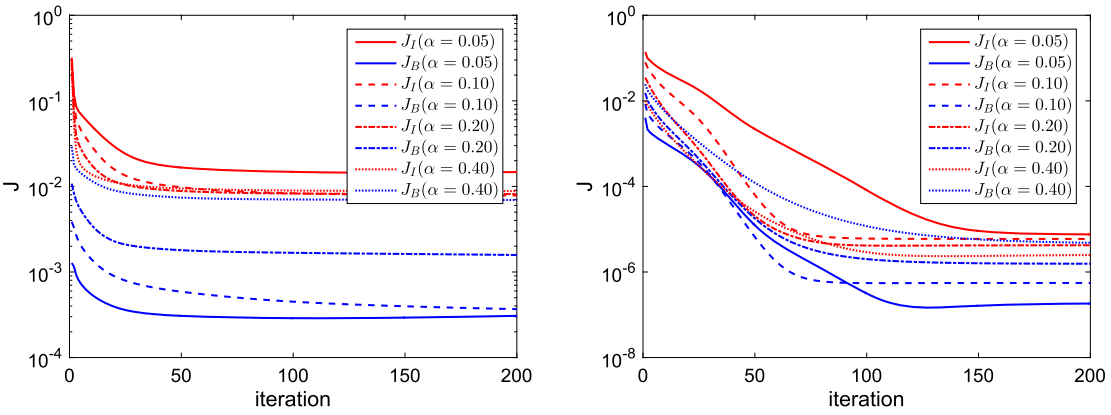

Fig. 10. For the first test case the interior functional J_I and the boundary functional J_B are shown as function of the number of iterations for different α's, on the left for the method in Cartesian coordinates and on the right for the method in polar coordinates. In both cases a 100×100 grid was used.

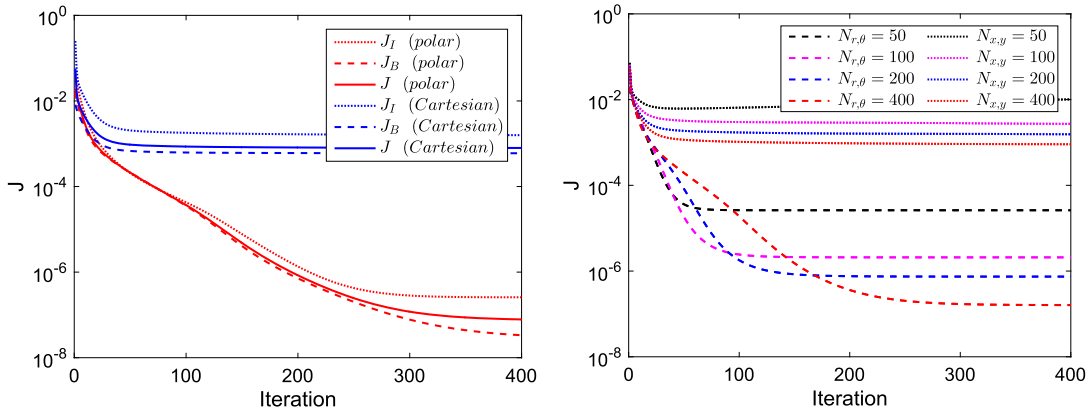

Fig. 11. The convergence history for Cartesian and polar coordinates is compared for the first test case. In the left plot a 500×500 grid is used and the different error components are shown. In the right plot $N_{r,\theta}$ refers tot the number of grid points that is used in both the radial and angular directions and $N_{x,y}$ has an analogous meaning. 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 $\alpha = 0.2$.

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 $(\mathscr{E}, E)$ and $(\mathscr{F}, F)$ the solution found by the method seems less dependable on the choice of α. However, in cases where for example $\partial\mathscr{F}$ is not differentiable, as in the current test case, the boundary condition and (11) seem to be conflicting goals. For such problems they cannot be satisfied exactly and simultaneously. In order to clarify and quantify this alleged conflict further study has to be done. Nonetheless, the freedom in α provides the user of the GLS method with an opportunity to choose the best balance of relative weight given to the boundary condition and (11) for the specific application at hand. Moreover, it can be seen that due to better handling of the boundary $\partial\mathscr{E}_\mathrm{U}$ the gap between J_I and J_B is far smaller for the method in polar coordinates for all choices of α.

Fig. 11 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 as 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, the final value for J_I seems not to convergence to zero when ever finer grids are used.

In Fig. 12 the convergence of the method as a function of the number of variables is analyzed. It is seen that the convergence rate is approximately second order for the method in polar coordinates. In Cartesian coordinates only a convergence rate of approximately first order is achieved. This is probably a result of the boundary issue alluded to above.

4.2. From a circle to a square: Gaussian emittance

As a second test we analyze the performance of the method for a source with non-uniform intensity. We again take $\mathscr{E}$ to be the unit disk, but now take the intensity function to be a (radial-symmetric) Gaussian:

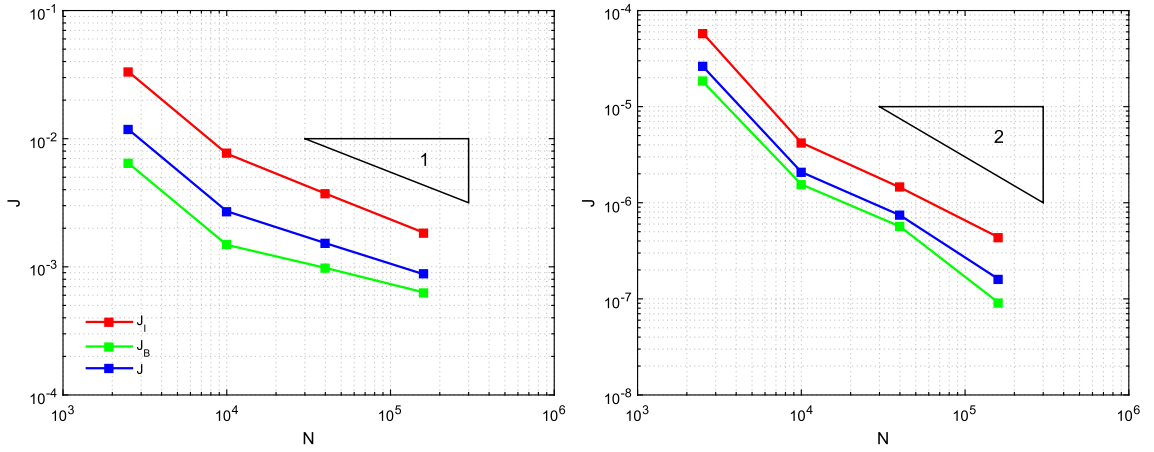

Fig. 12. The final errors J_{I}, J_{B} and J are shown as a function of the total number of variables N for the method applied to first test case. Left: Cartesian coordinates with $\sqrt{N} = N_x = N_y$. Right: polar coordinates with $\sqrt{N} = N_r = N_\theta$.

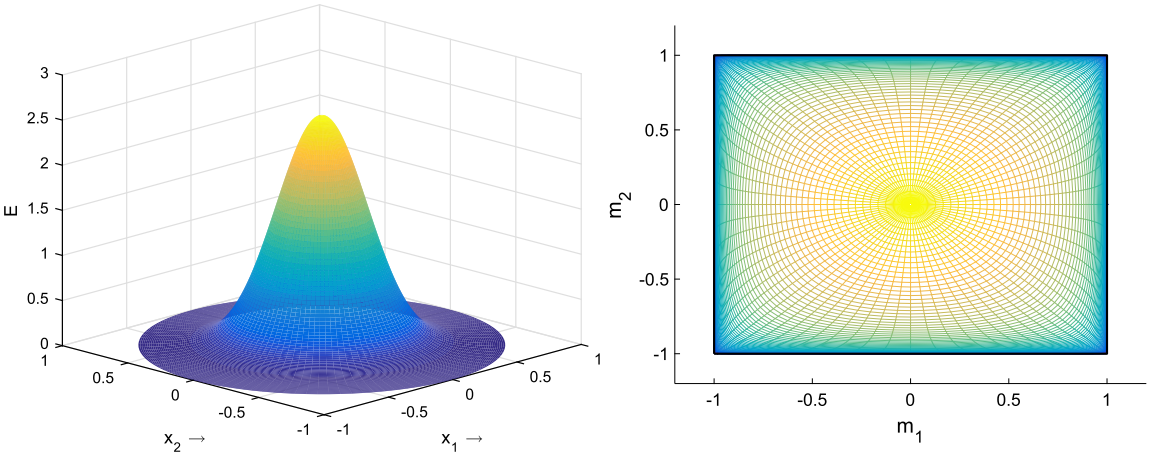

Fig. 13. Left: Emittance function E_{G}. Right: mapping for $N_r = N_\theta = 400$ and $\alpha = 0.2$.

$$E_{\mathrm{G}}(r) = a \exp\left(-\frac{r^2}{2\sigma^2}\right),$$

where a is chosen such that the integral of E_{G} over $\mathscr{E}$ is 1 and we taken the standard deviation $\sigma = 0.25$. The source intensity at $\partial\mathscr{E}$ is equal to

$$E_{\mathrm{G}}(1) = \frac{1}{2\pi\sigma^2}\left(\exp\left(\frac{1}{2\sigma^2}\right) - 1\right)^{-1} \approx 8.5454 \cdot 10^{-4}.$$

As a result the ratio between the maximum and minimum intensity approximately equals $E_{\mathrm{G}}(0)/E_{\mathrm{G}}(1) \approx 2.9786 \cdot 10^3$.

In Fig. 13 the emittance function E_{G} and the resulting mapping are shown for the method in polar coordinates. For this test case the method in Cartesian coordinates does not even converge. We again expect this to be a result of the boundary issue discussed before in the first paragraph of Section 4.1. Moreover, for the current test E_{G} is close to 0 near $\partial\mathscr{E}$ which the Cartesian method cannot deal with.

The convergence of the spatial discretization for this test case as a function of the number of variables is shown in Fig. 14. The order of convergence for this problem is at least 4. It is not entirely clear to us why the convergence for this test case is more than fourth order while it was only approximately second order for the test case with uniform emittance.

4.3. From a circle to an Escher lithograph in the far-field

In the final test case we calculate the reflector for the target intensity $(\mathscr{F}_{\mathrm{E}}, F_{\mathrm{E}})$ corresponding to the lithograph by Escher. We take the projection screen at a distance 100 times the radius of the source and we take $(\mathscr{F}_{\mathrm{E}}, F_{\mathrm{E}})$ such that width and height of the projection are 4.3 times the radius of the source. To avoid division by zero in (12a), we increased the minimum

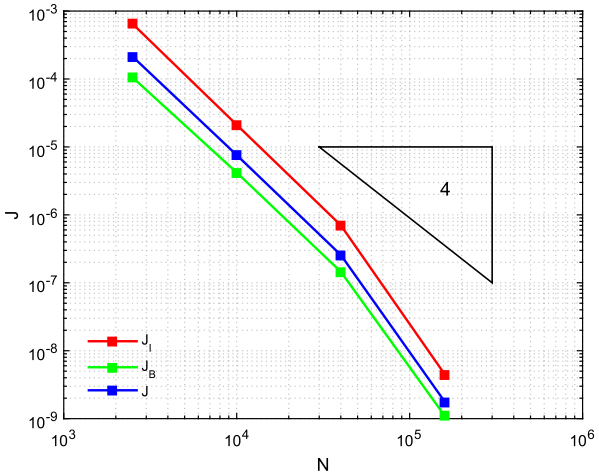

Fig. 14. The final errors J_{I}, J_{B} and J are shown as a function of the total number of variables N for the method applied to second test case with $\sqrt{N} = N_r = N_\theta$.

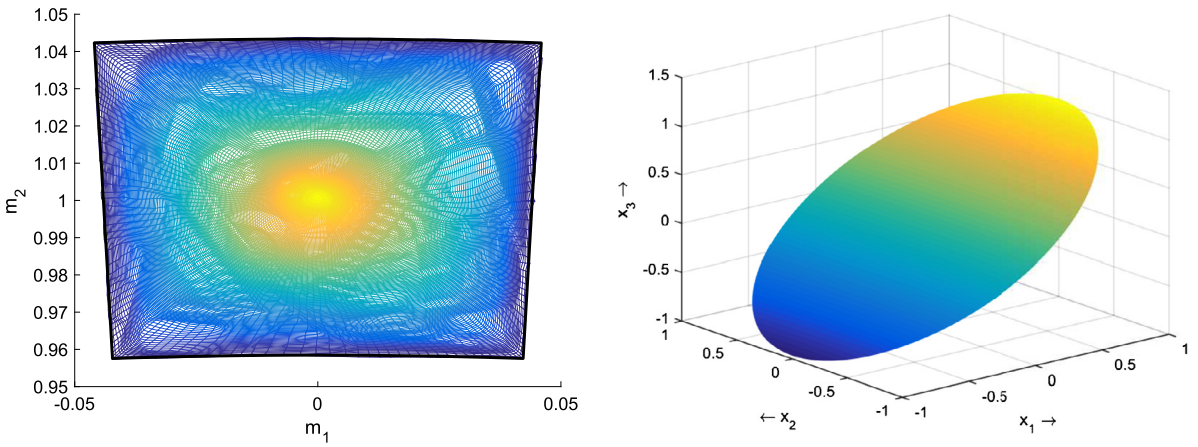

Fig. 15. The mapping and corresponding reflector are shown for the Escher test case.

intensity of the lithograph to be 5% of the maximum intensity. For this very demanding test case a 1400×1400 grid was used and $\alpha = 0.1$.

Fig. 15 shows the results of applying the method in polar coordinates. The elliptic shaped reflector is globally close to flat but locally contains great detail. The integrands of the final errors J_{I} and J_{B} are shown in Fig. 16. It is seen that the errors are small for this fine mesh.

Subsequently, the reflector was simulated by using a ray trace method [3]. The result can be seen in Fig. 17. The ray trace result closely resembles 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 function of the original picture. Nonetheless the resolution obtained is high enough to carry over all the minute details of the original picture.

5. Summary and concluding 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 $\mathscr{E}$ to the direction of the reflected 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 $(\mathscr{E}, E)$ and the gradient set and intensity function $(\mathscr{F}, F)$.

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.

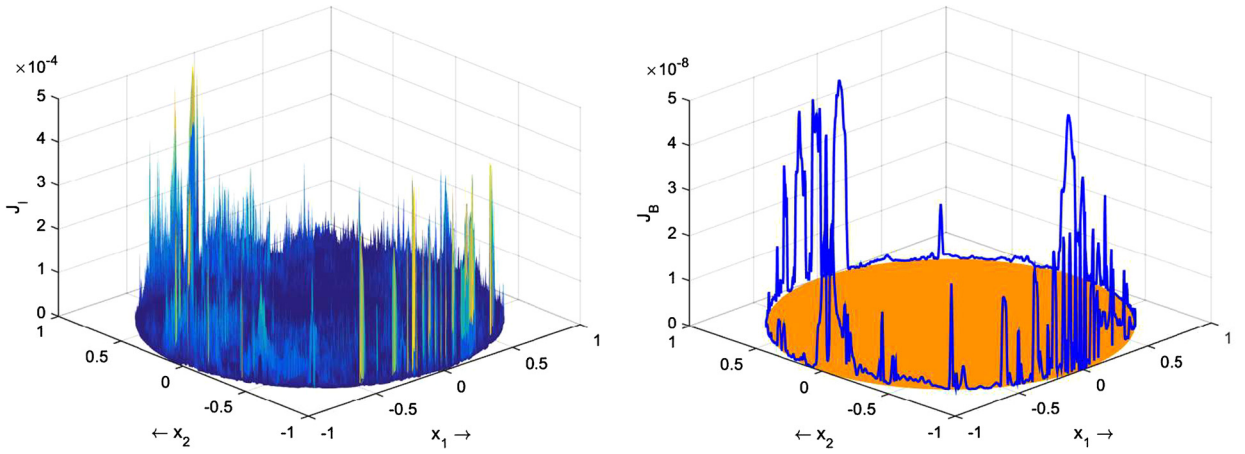

Fig. 16. Left: The interior error $\|\nabla\hat{\boldsymbol{m}} - \boldsymbol{P}\|^2/2$ as a function of $x \in \mathscr{E}$. Right: The boundary error $\|\boldsymbol{m} - \boldsymbol{b}\|^2/2$ as a function of $x \in \partial\mathscr{E}$.

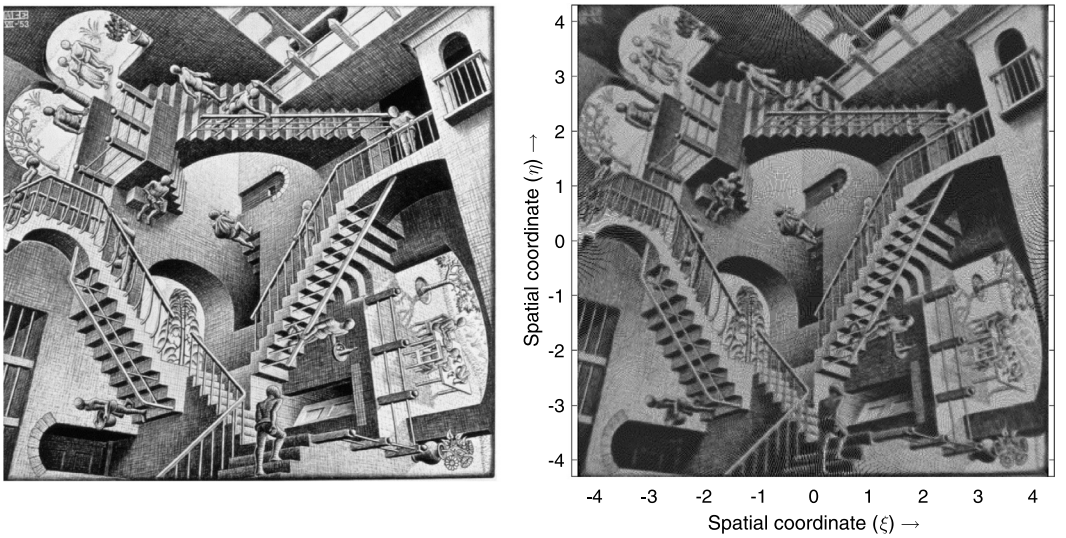

Fig. 17. The image projected on the screen by the reflector of Fig. 15 is determined by ray tracing about 4 million rays, with uniform fixed spacing, that leave $\mathscr{E}$. The original is shown on the left and the ray trace result is shown on the right.

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 method 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 (16). 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 $(\mathscr{E}, E)$ and target pair $(\mathscr{F}, F)$ 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 on $\mathbb{R}^n$ in a nutshell

A.1. Tensors

Consider the space of linear functionals from the n-dimensional vector space V to the real numbers: $\mathscr{L}(V;\mathbb{R})$. This space is called the dual space of V and we will denote it by $V^* := \mathscr{L}(V;\mathbb{R})$. Suppose $\boldsymbol{v} \in V$ and $\boldsymbol{v} = v^i \boldsymbol{e}_i$, where v^i are the components with respect to a basis $(\boldsymbol{e}_i)_{i=1}^n$ of the vector space V and we employ the Einstein summation convention. Furthermore, let $\hat{\boldsymbol{w}} \in V^*$, where we use a hat on $\boldsymbol{w}$ to indicate that its from V^* and not from V. By linearity of $\hat{\boldsymbol{w}}$ it follows that $\hat{\boldsymbol{w}}(\boldsymbol{v}) = v^i \hat{\boldsymbol{w}}(\boldsymbol{e}_i)$. This shows that a linear mapping $\hat{\boldsymbol{v}}$ is completely determined by specifying its action on the

basis elements of V. We can define in this way a set of elements of V^*, denoted by $(\boldsymbol{e}^i)_{i=1}^n$, according to $\boldsymbol{e}^i(\boldsymbol{e}_j) = \delta^i_j$. The elements $(\boldsymbol{e}^i)_{i=1}^n$ form a basis for V^*. They span V^*, because we have

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

where $w_i := \hat{\boldsymbol{w}}(\boldsymbol{e}_i)$. Furthermore, suppose $\hat{\boldsymbol{w}} = w_i \boldsymbol{e}^i = 0$ then

$$0 = \hat{\boldsymbol{w}}(\boldsymbol{e}_j) = w_i \boldsymbol{e}^i(\boldsymbol{e}_j) = w_i \delta^i_j = w_j,$$

which indicates that $w_j = 0$ for $1 \leq j \leq n$, implying that $(\boldsymbol{e}^i)_{i=1}^n$ are linearly independent. The basis $(\boldsymbol{e}^i)_{i=1}^n$ is called the dual basis and elements of V^* are also called covectors.

The double dual $V^{**} := \mathscr{L}(V^*; \mathbb{R})$ can be identified with V. This follows because when we define the double dual basis $(\boldsymbol{f}_i)_{i=1}^n$ according to $\boldsymbol{f}_j(\boldsymbol{e}^i) = \delta^i_j$, the identification of $\boldsymbol{f}_i$ with $\boldsymbol{e}_i$ gives an isomorphism between V and V^{**}. This shows that we can equally well consider V to be the dual to V^* and write $\boldsymbol{v}(\hat{\boldsymbol{w}}) := \hat{\boldsymbol{w}}(\boldsymbol{v})$. Similar to $w_i := \hat{\boldsymbol{w}}(\boldsymbol{e}_i) = \boldsymbol{e}_i(\hat{\boldsymbol{w}})$, we define $v^i := \boldsymbol{e}^i(\boldsymbol{v}) = \boldsymbol{v}(\boldsymbol{e}^i)$.

The space $\mathscr{L}(V, V^*; \mathbb{R})$ is the space of multilinear functionals (mappings linear in each argument) from $V \times V^*$ to $\mathbb{R}$. We will denote this space by $\boldsymbol{T}^1_1(V)$. An example of a functional in $V \times V^*$ to $\mathbb{R}$ is $\boldsymbol{e}^i \otimes \boldsymbol{e}_j$ which is defined as

$$\boldsymbol{e}^i \otimes \boldsymbol{e}_j(\boldsymbol{v}, \hat{\boldsymbol{w}}) = \boldsymbol{e}^i(\boldsymbol{v})\boldsymbol{e}_j(\hat{\boldsymbol{w}}) = v^i w_j.$$

This element of $\boldsymbol{T}^1_1(V)$ is called the tensor product of $\boldsymbol{e}^i$ and $\boldsymbol{e}_j$. It can be shown that $(\boldsymbol{e}^i \otimes \boldsymbol{e}_j)_{i,j=1}^n$ is actually a basis for $\boldsymbol{T}^1_1(V)$ [27, p. 25]. For an element $\boldsymbol{T} \in \boldsymbol{T}^1_1(V)$ we define its coefficients as $T^i_j = \boldsymbol{T}(\boldsymbol{e}_j, \boldsymbol{e}^i)$ and have $\boldsymbol{T} = T^i_j \boldsymbol{e}^j \otimes \boldsymbol{e}_i$. By linearity it follows that for a vector $\boldsymbol{v}$ and covector $\hat{\boldsymbol{w}}$ we have

$$\boldsymbol{T}(\boldsymbol{v}, \hat{\boldsymbol{w}}) = T^i_j v^j w_i.$$

Analogously we can define the spaces $\boldsymbol{T}^2_0(V)$ and $\boldsymbol{T}^0_2(V)$, with bases $(\boldsymbol{e}_i \otimes \boldsymbol{e}_j)_{i,j=1}^n$ and $(\boldsymbol{e}^i \otimes \boldsymbol{e}^j)_{i,j=1}^n$, respectively. Or, more generally, we can define the space $\boldsymbol{T}^p_q(V)$ and its corresponding basis, formed by taking consecutive tensor products between basis elements of V and V^*. The elements of the space $\boldsymbol{T}^p_q(V)$ are called tensors of contravariant rank p and covariant rank q. The spaces V and V^* are denoted by $\boldsymbol{T}^1_0(V)$ and $\boldsymbol{T}^0_1(V)$, respectively.

An example of a tensor of covariant rank 2 is given by the inner product $(\cdot, \cdot) : V \times V \to \mathbb{R}$. This tensor is known as the metric tensor or simply the metric. We denote the coefficients of the metric by $e_{ij} = (\boldsymbol{e}_i, \boldsymbol{e}_j)$. It follows that the metric can also be written as $\boldsymbol{e} = e_{ij}\boldsymbol{e}^i \otimes \boldsymbol{e}^j$. The inner product also provides us with a bijection between V and V^*. This bijection is given by the map

$$\boldsymbol{v} \mapsto (\boldsymbol{v}, \cdot). \tag{A.1}$$

In terms of the components of $\boldsymbol{v}$ this is given by $v^i \boldsymbol{e}_i \mapsto v^i e_{ij} \boldsymbol{e}^j$. The inverse of this bijection is given by the inverse of the matrix (e_{ij}), whose components we denote by (e^{ij}). Using this matrix we can express the inverse of (A.1) as $v_i \boldsymbol{e}^i \mapsto v_j e^{ij} \boldsymbol{e}_i$. It is conventional to write $v_j = v^i e_{ij}$ and $v^i = v_j e^{ij}$. We will follow this convention in this paper. The symmetry of the inner product implies that the matrices (e_{ij}) and (e^{ij}) are symmetric as well.

For tensors $\boldsymbol{T} \in \boldsymbol{T}^2_0(V)$, $\boldsymbol{S} \in \boldsymbol{T}^1_1(V)$ and $\boldsymbol{R} \in \boldsymbol{T}^0_2(V)$ we can define the trace as

$$\operatorname{tr}(\boldsymbol{T}) := T^{ij} e_{ij}, \quad \operatorname{tr}(\boldsymbol{S}) := S^i_i, \quad \operatorname{tr}(\boldsymbol{R}) := R_{ij} e^{ij}. \tag{A.2}$$

This definition as a contraction involves only the tensor itself and possibly the metric tensor and therefore independent of the choice of coordinate system [20, p. 82]. For an orthonormal coordinate system the traces of all three are just the ordinary traces of their coefficient matrix.

A.2. Tensor fields and coordinate systems on $\mathbb{R}^n$

We restrict the discussion now to $n = 2$, the dimension of the source $\mathscr{E}$. Just as we have vector fields on $\mathbb{R}^2$, which assign to each point $\boldsymbol{x} \in \mathbb{R}^2$ a vector in $V = \mathbb{R}^2$, we can define a tensor field, which assigns to each point $\boldsymbol{x} \in \mathbb{R}^n$ a tensor in $\boldsymbol{T}^p_q(V) = \boldsymbol{T}^p_q(\mathbb{R}^2)$. It is important to note that the basis, with respect to which a tensor in this tensor field can be expressed in a certain point $\boldsymbol{x} \in \mathbb{R}^2$, in general depends on this point. For the basis vectors of a Cartesian coordinate system this is not the case, but for polar coordinates the basis vectors do depend on position. We will denote the space of tensor fields of contravariant rank p and covariant rank q on some subset $\mathscr{U}$ of $\mathbb{R}^2$ by $\boldsymbol{T}^p_q(\mathscr{U})$.

In principle the basis in each point $\boldsymbol{x}$ can be chosen independently of the coordinate system used to represent the points in $\mathbb{R}^2$. However, a logical choice is to use the vectors found by taking the tangent vectors to the coordinate lines. For Cartesian coordinates this gives the usual Cartesian basis $\{\boldsymbol{e}_x, \boldsymbol{e}_y\}$ and for polar coordinates we find

$$\boldsymbol{e}_r := \frac{\partial}{\partial r}\big(r\cos(\theta)\boldsymbol{e}_x + r\sin(\theta)\boldsymbol{e}_y\big) = \cos(\theta)\boldsymbol{e}_x + \sin(\theta)\boldsymbol{e}_y, \tag{A.3a}$$

$$\boldsymbol{e}_\theta := \frac{\partial}{\partial \theta}\big(r\cos(\theta)\boldsymbol{e}_x + r\sin(\theta)\boldsymbol{e}_y\big) = -r\sin(\theta)\boldsymbol{e}_x + r\cos(\theta)\boldsymbol{e}_y. \tag{A.3b}$$

A basis defined by taking the tangent vectors to the coordinate lines is called a coordinate basis or holonomic basis. It is clear that not for every choice of basis vector fields there exists a coordinate system to which these basis vector fields are the tangent vector fields. An often used example of such a choice of basis is the orthonormal polar coordinate basis, which is found by rescaling (A.3b) to unit length. Such a type of basis is called anholonomic.

A.3. Covariant derivative of tensor fields

We now introduce the covariant derivative on $\mathbb{R}^2$. On Euclidean spaces the covariant derivative is nothing more than the directional derivative. Suppose $f \in C^1(\mathbb{R}^2)$ then the directional derivative of f in the direction of the vector $\boldsymbol{v} \in \boldsymbol{T}^1_0(\mathbb{R}^2)$ at $\boldsymbol{x} \in \mathbb{R}^2$ is defined as

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

where $\gamma_{\boldsymbol{v}}$ is a curve such that $\gamma_{\boldsymbol{v}}(0) = \boldsymbol{x}$ and $\mathrm{d}\gamma_{\boldsymbol{v}}(t)/\mathrm{d}t|_{t=0} = \boldsymbol{v}$. When $\boldsymbol{e}_i$ is a coordinate basis vector then we have $\nabla_{\boldsymbol{e}_i} f = \frac{\partial f}{\partial x^i}$, where x^i is the coordinate corresponding to $\boldsymbol{e}_i$. The differential of f is the covector given by $\mathrm{d}f := (\nabla_{\boldsymbol{e}_i} f)\boldsymbol{e}^i$. The vector corresponding to this covector by bijection (A.1) is $\nabla f = e^{ij}(\nabla_{\boldsymbol{e}_i} f)\boldsymbol{e}_j$ and called the gradient of f.

When we take the directional derivatives of vectors, covectors and general tensors we have to take care of the fact that the basis vectors might also change in the direction of the derivative. The directional derivatives of vectors in $\mathbb{R}^2$ are itself vectors in $\mathbb{R}^2$ and therefore there exist coefficients Γ^k_{ji} such that $\nabla_{\boldsymbol{e}_i}(\boldsymbol{e}_j) = \Gamma^k_{ji}\boldsymbol{e}_k$. These coefficients are called Christoffel symbols. Using these we can determine the directional derivative of contravariant tensors. For $\boldsymbol{v} \in T^1_0(\mathbb{R}^2)$, $\boldsymbol{T} \in T^2_0(\mathbb{R}^2)$, we have by the product rule

$$\nabla_{\boldsymbol{e}_i}\boldsymbol{v} = \nabla_{\boldsymbol{e}_i}(v^j)\boldsymbol{e}_j + v^j\nabla_{\boldsymbol{e}_i}(\boldsymbol{e}_j) = \big(\nabla_{\boldsymbol{e}_i}(v^j) + v^k\Gamma^j_{ki}\big)\boldsymbol{e}_j, \tag{A.4a}$$

$$\nabla_{\boldsymbol{e}_i}\boldsymbol{T} = \nabla_{\boldsymbol{e}_i}(T^{jk})\boldsymbol{e}_j\otimes\boldsymbol{e}_k + T^{jk}(\nabla_{\boldsymbol{e}_i}\boldsymbol{e}_j)\otimes\boldsymbol{e}_k + T^{jk}\boldsymbol{e}_j\otimes(\nabla_{\boldsymbol{e}_i}\boldsymbol{e}_k) = \big(\nabla_{\boldsymbol{e}_i}(T^{jk}) + T^{lk}\Gamma^j_{li} + T^{jl}\Gamma^k_{li}\big)\boldsymbol{e}_j\otimes\boldsymbol{e}_k. \tag{A.4b}$$

We will sometimes use the notation $D_i v^j = \nabla_{\boldsymbol{e}_i}(v^j) + v^k\Gamma^j_{ki}$ and $D_i T^{jk} = \nabla_{\boldsymbol{e}_i}(T^{jk}) + T^{lk}\Gamma^j_{li} + T^{jl}\Gamma^k_{li}$. Furthermore, we will use the notational convention $D^i = e^{ij}D_j$. Using the fact that the inner product and hence the corresponding metric tensor $\boldsymbol{e}$ does not change with position [20, p. 215] we have $D_k(e^{ij}) = 0$ from which it follows that $\nabla_{\boldsymbol{e}_k}(e^{ij}) = -\Gamma^j_{lk}e^{il} - \Gamma^i_{lk}e^{lj}$ by (A.4b). Using this it follows from calculating $D_k(v^i) = D_k(e^{ij}v_j)$ and $D_m(S^{ij}) = D_m(e^{ik}e^{jl}S_{kl})$ that we have for $\hat{\boldsymbol{v}} \in T^0_1(\mathbb{R}^2)$, $\boldsymbol{S} \in T^0_2(\mathbb{R}^2)$, $\boldsymbol{R} \in T^1_1(\mathbb{R}^2)$ that

$$\nabla_{\boldsymbol{e}_i}\hat{\boldsymbol{v}} = D_i(v_j)\boldsymbol{e}^j = \big(\nabla_{\boldsymbol{e}_i}(v_j) - v_k\Gamma^k_{ji}\big)\boldsymbol{e}^j, \tag{A.5a}$$

$$\nabla_{\boldsymbol{e}_i}\boldsymbol{S} = D_i(S_{jk})\boldsymbol{e}^j\otimes\boldsymbol{e}^k = \big(\nabla_{\boldsymbol{e}_i}(S_{jk}) - S_{lk}\Gamma^l_{ji} - S_{jl}\Gamma^l_{ki}\big)\boldsymbol{e}^j\otimes\boldsymbol{e}^k. \tag{A.5b}$$

$$\nabla_{\boldsymbol{e}_i}\boldsymbol{R} = D_i(R^j_k)\boldsymbol{e}_j\otimes\boldsymbol{e}^k = \big(\nabla_{\boldsymbol{e}_i}(R^j_k) + R^l_k\Gamma^j_{li} - R^j_l\Gamma^l_{ki}\big)\boldsymbol{e}_j\otimes\boldsymbol{e}^k. \tag{A.5c}$$

A tensor of special interest in this paper is found by again differentiating the differential of a function. This tensor is the Hessian tensor and defined as [26, p. 172]

$$\boldsymbol{H}(v) := \nabla_{\boldsymbol{e}_j}(\mathrm{d}v)\otimes\boldsymbol{e}^j = D_j(\nabla_{\boldsymbol{e}_i}v)\boldsymbol{e}^i\otimes\boldsymbol{e}^j = \big(\nabla_{\boldsymbol{e}_j}(\nabla_{\boldsymbol{e}_i}v) - \Gamma^k_{ij}\nabla_{\boldsymbol{e}_k}v\big)\boldsymbol{e}^i\otimes\boldsymbol{e}^j. \tag{A.6}$$

The covariant directional derivative as introduced above can be generalized to Riemannian manifolds and is in that context called a Levi-Civita connection [26, p. 160]. For a Levi-Civita connection the Hessian matrix is symmetric [28, p. 4], hence the Hessian matrix will always be symmetric in this paper. Note that in Cartesian coordinates $(h_{ij}(v))$, the coefficient matrix of $\boldsymbol{H}(v)$, is the matrix with second derivatives of v.

Appendix B. Equivalence of boundary conditions for a strictly convex reflector surface

In this appendix we show that the Inverse Reflector Problem as stated on page 353 is equivalent to this same problem but with the boundary condition $\nabla u(\partial\mathscr{E}) = \partial\mathscr{F}$ replaced by $\nabla u(\mathscr{E}) = \mathscr{F}$. 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 6. *Suppose that $u \in C^2(\mathscr{E})$ is the strictly convex solution to the Inverse Reflector Problem with either $\nabla u(\mathscr{E}) = \mathscr{F}$ or $\nabla u(\partial\mathscr{E}) = \partial\mathscr{F}$ as boundary condition. Then the map ∇u is open, i.e., for each open subset $A \subset \mathscr{E}$ the image $\nabla u(A)$ is an open subset of $\mathscr{F}$.*

Proof. In Lemma 1 we saw that for the strictly convex solution $u \in C^2(\mathscr{E})$, ∇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 [29] are satisfied. The inverse function theorem states, among other things, that for every open subset A of $\mathscr{E}$ and $\boldsymbol{x} \in A$, there exists an open set $U_{\boldsymbol{x}}$ in A containing $\boldsymbol{x}$, and an open set $V_{\boldsymbol{x}}$ in $\mathscr{F}$ containing $\nabla u(\boldsymbol{x})$ such that ∇u is a bijection from $U_{\boldsymbol{x}}$ to $V_{\boldsymbol{x}}$ and the inverse $(\nabla u)^{-1}$ is continuously differentiable on $V_{\boldsymbol{x}}$.

From this it follows that ∇u is open. To see this, suppose A is some open set in $\mathscr{E}$. By the inverse function theorem there exists for every $\boldsymbol{x} \in \mathscr{E}$ open sets $U_{\boldsymbol{x}}$ and $V_{\boldsymbol{x}}$ such that $\boldsymbol{x} \in U_{\boldsymbol{x}}$, $\nabla u(\boldsymbol{x}) \in V_{\boldsymbol{x}}$ and $U_{\boldsymbol{x}} \subset A$. $\nabla u(U_{\boldsymbol{x}}) = V_{\boldsymbol{x}}$ is open for every $\boldsymbol{x} \in A$. Notice that $\cup_{\boldsymbol{x}\in A} U_{\boldsymbol{x}} = A$ and that $\nabla u(A) = \cup_{\boldsymbol{x}\in A} \nabla u(U_{\boldsymbol{x}}) = \cup_{\boldsymbol{x}\in A} V_{\boldsymbol{x}}$. Thus ∇u is an open map. □

The map ∇u is a homeomorphism from $\mathscr{E}$ to $\mathscr{F}$, 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 7. *Let $u \in C^2(\mathscr{E})$ be the strictly convex solution to the Inverse Reflector Problem with $\nabla u(\mathscr{E}) = \mathscr{F}$ instead of $\nabla u(\partial\mathscr{E}) = \partial\mathscr{F}$. Then also $\nabla u(\partial\mathscr{E}) = \partial\mathscr{F}$.*

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

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

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

Thus the strictly convex solution of the Inverse Reflector Problem with boundary condition $\nabla u(\mathscr{E}) = \mathscr{F}$ is also a solution to the Inverse Reflector Problem with boundary condition $\nabla u(\partial\mathscr{E}) = \partial\mathscr{F}$. Now the following lemma states the converse.

Lemma 8. *Let $u \in C^2(\mathscr{E})$ be a strictly convex solution to Inverse Reflector Problem. Then $\nabla u(\mathscr{E}) = \mathscr{F}$.*

Proof. The map ∇u is a homeomorphism from $\mathscr{E}$ to $\nabla u(\mathscr{E}) \subset \mathbb{R}^2$. The set $\mathscr{E}$ is convex and hence simply connected. The set $\partial\mathscr{E}$ is a simple and closed curve, i.e., a Jordan curve. The map ∇u is continuous and injective and hence $\nabla u(\partial\mathscr{E}) = \partial\mathscr{F}$ is also a Jordan curve. Now the Jordan curve theorem [30, p. 198] states that the complement $\mathbb{R}^2 \backslash \partial\mathscr{F}$ 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 $\partial\mathscr{F}$. The set $\mathscr{E}$ is simply connected, therefore $\nabla u(\mathscr{E})$ is simply connected also. The interior and exterior to the curve $\nabla u(\partial\mathscr{E}) = \partial\mathscr{F}$ are the only two subsets of $\mathbb{R}^2$ with $\partial\mathscr{F}$ as boundary. The fact that ∇u is a homeomorphism implies $\nabla u(\partial\mathscr{E}) = \partial(\nabla u(\mathscr{E}))$, because $\nabla u(\mathrm{int}(\mathscr{E})) = \mathrm{int}(\nabla u(\mathscr{E}))$ as we showed in the proof of Lemma 7. The fact that $\partial\mathscr{F} = \nabla u(\partial\mathscr{E}) = \partial(\nabla u(\mathscr{E}))$ implies that $\mathrm{int}(\nabla u(\mathscr{E}))$ is one of two sets of the Jordan curve theorem. The exterior set is clearly not simply connected, while $\nabla u(\mathscr{E})$ is, therefore $\mathrm{int}(\nabla u(\mathscr{E}))$ is the interior set in the Jordan curve theorem. The fact that $\mathscr{E}$ is bounded, that the functions E and F are strictly positive and bounded and that F is bounded away from zero implies by (10) that the set $\mathscr{F}$ is bounded also. This implies that $\mathrm{int}(\mathscr{F})$ needs to be the interior set also and hence we find that $\nabla u(\mathscr{E}) = \mathscr{F}$. □

We have established that u is a strictly convex solution to the Inverse Reflector Problem with boundary conditions $\nabla u(\mathscr{E}) = \mathscr{F}$ if and only if it is a strictly convex solution to the Inverse Reflector Problem with boundary condition $\nabla u(\partial\mathscr{E}) = \partial\mathscr{F}$. 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, Phys. Status Solidi A 208 (1) (2011) 17–29.
[2] T. Taguchi, Present status of energy saving technologies and future prospects in white LED lighting, IEEJ Trans. Electr. Electron. Eng. 3 (2008) 21–26.
[3] A.S. Glassner, An introduction to Ray Tracing, Academic Press, 1991.
[4] F.Z. Fang, X.D. Zhang, A. Weckenmann, G.X. Zhang, C. Evans, Manufacturing and measurement of freeform optics, CIRP Ann. – Manuf. Technol. 62 (2) (2013) 823–846.
[5] C. Villani, Topics in Optimal Transportation, American Mathematical Society, Providence, 2003.
[6] V. Oliker, Designing freeform lenses for intensity and phase control of coherent light with help from geometry and mass transport, Arch. Ration. Mech. Anal. 201 (3) (2011).
[7] J.D. Benamou, Y. Brenier, A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem, Numer. Math. 84 (2000) 375–393.

[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 (2010) 197–211.
[9] B.D. Froese, A.M. Oberman, Fast finite difference solvers for singular solutions of the elliptic Monge–Ampère equation, J. Comput. Phys. 230 (3) (2011) 818–834.
[10] J.D. Benamou, B.D. Froese, A.M. Oberman, Numerical solution of the optimal transportation problem using the Monge–Ampère equation, J. Comput. Phys. 260 (2014) 107–126.
[11] K. Brix, Y. Hafizogullari, A. Platen, Solving the Monge–Ampère equations for the inverse reflector problem, Math. Models Methods Appl. Sci. 25 (2015) 803–837.
[12] C.R. Prins, Inverse Methods for Illumination Optics, Ph.D. Thesis, 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) (2015) B937–B961.
[14] A. Caboussat, R. Glowinski, D.C. Sorensen, A least-squares method for the numerical solution of the Dirichlet problem for the elliptic Monge–Ampère equation in dimension two, ESAIM Control Optim. Calc. Var. 19 (2013) 780–810.
[15] Official website on Escher: http://www.mcescher.com/about/.
[16] M. Bass (Ed.), Handbook of Optics Volume II – Devices, Measurements and Properties, 2nd ed, McGraw-Hill, 1995.
[17] G. Monge, Application de l'analyse a la geometrie, a l'usage de l'Ecole imperiale polytechnique, Bernard, Paris, 1807.
[18] H. Cohn, Conformal Mapping on Riemann Surfaces, McGraw-Hill, 1967.
[19] W. Rudin, Real and Complex Analysis, 3rd ed., McGraw-Hill, London, 1987.
[20] C.W. Misner, K.S. Thorne, J.A. Wheeler, Gravitation, W. H. Freeman, San Francisco, 1973.
[21] R. Aris, Vectors, Tensors, and the Basic Equations of Fluid Mechanics, Dover Publications, 1989.
[22] M. Spivak, Calculus on Manifolds: a Modern Approach to Classical Theorems of Advanced Calculus, Benjamin, Amsterdam, 1965.
[23] J.E. Marsden, A.J. Tromba, Vector Calculus, 5th ed., W.H. Freeman and Company, 2003.
[24] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.
[25] R. Courant, D. Hilbert, Methods of Mathematical Physics, vol. 1, 1989 ed., John Wiley & Sons, 1937.
[26] J. Jost, Riemannian Geometry and Geometric Analysis, 6th ed., Springer-Verlag, Berlin Heidelberg, 2011.
[27] S. Kobayashi, K. Nomizu, Foundations of Differential Geometry, Interscience, S.I., 1963.
[28] J. Morgan, G. Tian, Ricci Flow and the Poincaré Conjecture, American Mathematical Society, Providence RI, 2007.
[29] T. Tao, Analysis II, 2nd ed., Hindustan Book Agency, 2009.
[30] E.H. Spanier, Algebraic Topology, McGraw-Hill, London, 1966.