Inverse reflector design for a point source and far-field target

Citation for published version (APA):

Romijn, L. B., ten Thijs Boonkkamp, J. H. M., & IJzerman, W. L. (2020). Inverse reflector design for a point source and far-field target. *Journal of Computational Physics*, 408, [109283].
<https://doi.org/10.1016/j.jcp.2020.109283>

Document license:

TAVERNE

DOI:

[10.1016/j.jcp.2020.109283](https://doi.org/10.1016/j.jcp.2020.109283)

Document status and date:

Published: 01/05/2020

Document Version:

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

Please check the document version of this publication:

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

[Link to publication](#)

General rights

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

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

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

www.tue.nl/taverne

Take down policy

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

openaccess@tue.nl

providing details and we will investigate your claim.

Inverse reflector design for a point source and far-field target

Lotte B. Romijn^{a,*}, Jan H.M. ten Thije Boonkkamp^a, Wilbert L. Ijzerman^{a,b}

^a CASA, Department of Mathematics and Computer Science, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, the Netherlands

^b Signify Research, High Tech Campus 7, 5656 AE Eindhoven, the Netherlands

ARTICLE INFO

Article history:

Received 17 July 2019

Received in revised form 21 January 2020

Accepted 23 January 2020

Available online 27 January 2020

Keywords:

Geometrical optics

Optimal mass transport

Logarithmic cost function

Generalized Monge–Ampère equation

Least-squares method

Stereographic coordinates

ABSTRACT

We present a method for the design of a single freeform reflector that converts the light distribution of a point source to a desired light distribution in the far field. Using the geometrical-optics law of reflection and requiring energy conservation, this optical design problem can be represented by a generalized Monge–Ampère equation for the shape of the reflector with transport boundary condition. We use a generalized least-squares algorithm that can handle a logarithmic cost function in the corresponding optimal transport problem. The algorithm first computes the optical map and subsequently constructs the optical surface. We demonstrate that the algorithm can generate reflector surfaces for a number of complicated target distributions.

© 2020 Elsevier Inc. All rights reserved.

1. Introduction

A central problem in illumination optics is to determine an optical system that transforms a given light distribution on a source domain into a required distribution on a target domain. As an example of an application, an LED lamp contains one or multiple reflectors, lenses, diffusers and/or absorbers. These surfaces need to be designed in such a way that they redistribute the light from the LED into the required light output pattern of the lamp. Broadly, the methods for optical system design can be categorized as either *forward* or *inverse* methods.

Forward methods compute the target distribution from a known source distribution and optical system, most commonly using Monte-Carlo ray tracing techniques. The design of the optical system can be improved by making modifications to the optical elements and subsequently evaluating the output target distribution via ray tracing. Drawbacks of forward methods are that ray tracing can be slow if high precision is required and that the approach to create an improved design is often based on trial and error [1]. Filosa et al. [2,3] recently developed a new ray tracing method based on the phase space representation of the source and target domains, which improves the accuracy and reduces computation time of the classical approach.

Inverse methods directly compute the optical system converting the light from the source into the specified output. One approach for the inverse design of *freeform* (i.e., without any symmetries) optical surfaces uses the principles of geometrical optics and conservation of energy to derive a partial differential equation for the location of the optical surface. With the laws of reflection and/or refraction of geometrical optics it is possible to construct an optical mapping that connects

* Corresponding author.

E-mail address: l.b.romijn@tue.nl (L.B. Romijn).

coordinates on the source and target domains. Substituting the mapping into the relation for energy conservation leads to a fully nonlinear second order elliptic partial differential equation, which is a generalized Monge–Ampère equation [4, p. 282].

Generalized Monge–Ampère equations frequently arise in the field of optimal transport [5,4,6]. The optimal transport problem is concerned with finding a transport plan that minimizes a transport cost functional – i.e., an integral of a cost function weighted with the source distribution – under the constraint of energy conservation. More specifically, in optical design problems we seek an optimal transport plan or a mapping, such that all the light from the source is transmitted to the target and traverses the least optical path length. Each optical system corresponds to a different Monge–Ampère equation and associated cost function. For optical systems with quadratic cost functions, the optical surface can be described by the standard Monge–Ampère equation: a fully nonlinear second order partial differential equation that is linear in the determinant of the Hessian matrix [7,8]. Currently, nearly all existing algorithms only work for quadratic cost functions, for which existence and uniqueness of the solution can also be proven [9,6,10,8]. A typical example of an optical system that does not fit in this framework is a reflector that transforms the energy distribution of a point source into a required far-field intensity. This system has a logarithmic cost function in the optimal transport problem and the corresponding partial differential equation is nonlinear in the Hessian. Generalized Monge–Ampère equations with conditions for existence, uniqueness and smoothness of a solution to reflector-type problems with a point source were derived in [11–14].

Currently, there exist only a few numerical algorithms for generalized Monge–Ampère equations. Here we mention some numerical algorithms for the point source specifically, which do not assume any symmetry of the optical surface. Oliker [15] solves the problem using the method of supporting ellipsoids for a discrete target distribution and introduces an iterative optimization algorithm in [16]. Ries et al. [17] derive a set of partial differential equations for a point source using curvatures of wave fronts but do not present details of a numerical solution. Glimm and Oliker [18] develop a heuristic constructive algorithm for the dual Monge–Kantorovich problem using a linear programming approach. Fournier et al. [19] extend Oliker's method of supporting ellipsoids [16] and construct 3D reflectors that produce continuous illuminance distributions using Monte-Carlo ray tracing. Canavesi et al. [20] replace Monte-Carlo ray tracing in this algorithm by a flux estimation method which calculates the intersection points between triplets of ellipsoids. Wu et al. [21] derive the Monge–Ampère equation for a lens surface and solve the equations using standard finite differences and Newton iteration. Brix et al. [22,23] derive the Monge–Ampère equation for a point source with a near-field target and use a collocation method with a tensor-product B-spline basis to calculate reflectors and lenses capable of producing a detailed image on a near-field projection screen. Finally, we note that it is possible to use Sinkhorn's algorithm by using the logarithmic cost function in an entropic regularization framework, as illustrated in [24].

In this paper, we present a numerical algorithm for a point source with corresponding logarithmic cost function in the optimal transport problem. We show that an optical mapping can be derived both via the law of reflection and via the cost function, using a novel formulation in stereographic coordinates to represent the source and target domains. This formulation allows for a compact description of the optical mapping. The corresponding generalized Monge–Ampère equation in stereographic coordinates can be written in terms of the cost function and solved numerically using a generalization of the least-squares approach inspired by the methods presented in [7,10,25]. The method works by first computing the optical map in an iterative procedure which minimizes the defect in the energy balance. It also imposes a transport boundary condition by minimizing the deviation of the boundary of the optical map to the boundary of the target. We introduce a new method to impose the boundary condition using skew projections on line segments. Upon convergence of the iterative procedure, the location of the optical surface is calculated from the mapping also in a least-squares sense.

This paper is structured as follows. In Section 2 we present the new derivation of the Monge–Ampère equation for the freeform reflector with corresponding logarithmic cost function in stereographic coordinates. In Section 3 the generalized least-squares method is described. In Section 3.1 we introduce a variation to the method described in [7,10] to impose the transport boundary condition. In Section 4 we apply the algorithm to a few test cases. We show that our new boundary method leads to improved convergence at corners of the target domain. Notably, we challenge the algorithm to compute the reflector surface which transforms light of a point source into a complicated image on a screen in the far field and verify the result by a ray tracing method. To the best of our knowledge, such detailed reflector surfaces have not been computed for the point source before. Finally, we make some concluding remarks in Section 5.

2. Mathematical formulation

In this section we derive the partial differential equation which describes the shape of the reflector that transforms a beam of light originating from a point source into a specified output intensity distribution in the far field.

2.1. Geometrical formulation of the optical map

The geometry of the optical system transforming light from a point source into a far-field intensity is shown schematically in Fig. 1a. We assume that the light source is a point located at the origin $\mathcal{O}$ of the Cartesian coordinate system with $(x, y, z) \in \mathbb{R}^3$. In spherical coordinates the source emits light radially outward in the direction $\hat{\mathbf{s}} = \hat{\mathbf{e}}_r$. We use hats to denote unit vectors. The reflector surface is described by the parametrization $\mathbf{r}(\phi, \theta) = u(\phi, \theta) \hat{\mathbf{e}}_r$, where $u(\phi, \theta) > 0$ is the radial parameter that describes the location of the reflector surface, $0 \leq \phi \leq \pi$ is the zenith and $0 \leq \theta < 2\pi$ is the azimuth in the spherical coordinate system. The source has a given emittance $f(\phi, \theta)$ [lm/sr]. The target in the far field has

Fig. 1. Single reflector converting the emittance $f(\phi, \theta)$ of a point source into a far-field target intensity $g(\psi, \chi)$ and a diagram for the coordinate transformations of the incoming and outgoing rays.

a specified output distribution $g(\psi, \chi)$ [lm/sr], with respect to a different set of spherical coordinates (ψ, χ) , with zenith $0 \leq \psi \leq \pi$ and azimuth $0 \leq \chi < 2\pi$, taking the origin to be the reflector surface approximated as a point in space (far-field approximation). We take $f: S^2 \rightarrow [0, \infty)$ and $g: S^2 \rightarrow [0, \infty)$ to be continuous intensity functions.

2.1.1. The geometrical-optics approach

The mapping $\mathbf{m}$ can be determined by tracing a typical ray through the optical system. We consider an incident ray propagating in the direction $\hat{\mathbf{s}} = \hat{\mathbf{e}}_r$, which intercepts the reflector $\mathcal{R}$ and reflects off in direction $\hat{\mathbf{t}}$. The unit surface normal of the parametrized reflector surface $\mathbf{r}(\phi, \theta) = u(\phi, \theta) \hat{\mathbf{e}}_r$, directed towards the point source, is given by

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

with

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

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

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

Since the vectors $\hat{\mathbf{s}} = (s_1, s_2, s_3)^T$ and $\hat{\mathbf{t}} = (t_1, t_2, t_3)^T$, using a 3-tuple representation for Cartesian vectors, are both defined on the unit sphere S^2 , we can express the first two components in terms of the third component. For this reason it is convenient to perform coordinate transformations from spherical to stereographic. We use a 2-tuple representation for stereographic coordinate vectors and define

$$\mathbf{x}(\hat{\mathbf{s}}) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{1 + s_3} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \frac{1}{1 + \cos(\phi)} \begin{pmatrix} \sin(\phi) \cos(\theta) \\ \sin(\phi) \sin(\theta) \end{pmatrix}, \quad (3)$$

$$\mathbf{y}(\hat{\mathbf{t}}) = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \frac{1}{1 - t_3} \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = \frac{1}{1 - \cos(\psi)} \begin{pmatrix} \sin(\psi) \cos(\chi) \\ \sin(\psi) \sin(\chi) \end{pmatrix}, \quad (4)$$

as parametrizations of the unit sphere S^2 , with corresponding inverse projections

$$\hat{\mathbf{s}}(\mathbf{x}) = \hat{\mathbf{e}}_r = \frac{1}{1 + |\mathbf{x}|^2} \begin{pmatrix} 2x_1 \\ 2x_2 \\ 1 - |\mathbf{x}|^2 \end{pmatrix}, \quad \hat{\mathbf{t}}(\mathbf{y}) = \frac{1}{1 + |\mathbf{y}|^2} \begin{pmatrix} 2y_1 \\ 2y_2 \\ -1 + |\mathbf{y}|^2 \end{pmatrix}. \quad (5)$$

Fig. 1b schematically illustrates possible coordinate transformations for the source and target domains.

We represent the incoming rays $\hat{\mathbf{s}}$ by using a stereographic projection from the south pole $(0, 0, -1)$ onto the plane $z = 0$, as drawn schematically in Fig. 2a. The stereographic projection in (3) is undefined at the south pole, and we consider $s_3 \neq -1$ and $0 \leq \phi < \pi$. On the other hand, for the outgoing rays we use a stereographic projection from the north pole

We derive the mapping $\mathbf{y} = \mathbf{m}(\mathbf{x})$ from equation (2) and describe the steps in Appendix A. The resulting mapping $\mathbf{y} = \mathbf{m}(\mathbf{x})$ is given by

where ∇u_1 is the gradient of $u_1(\mathbf{x}) = \log(u(\mathbf{x})/(1 + |\mathbf{x}|^2))$ with respect to $\mathbf{x}$. We define our source domain $\mathcal{X}$ as the supporting domain of $\tilde{f}(\mathbf{x}) = f(\phi(\mathbf{x}), \theta(\mathbf{x}))$, and our target domain $\mathcal{Y}$ as the image under the mapping $\mathbf{m}$, i.e., $\mathcal{Y} = \mathbf{m}(\mathcal{X})$. We refer to $\mathbf{m} : \mathcal{X} \rightarrow \mathcal{Y}$ as the optical map $\mathbf{y} = \mathbf{m}(\mathbf{x})$ from the source set of stereographic coordinates $\mathcal{X}$ to the target set of stereographic coordinates $\mathcal{Y}$.

The cost function of the optical system can be derived using Fermat's principle, which states that a ray connecting the source $\mathcal{O}$ with a point on the reflected ray has stationary optical path length $L(\hat{\mathbf{s}}, \hat{\mathbf{t}})$. Let us consider an incident ray propagating from the source $\mathcal{O}$ in the direction $\hat{\mathbf{s}}$, which intercepts the reflector $\mathcal{R}$ and reflects off in the direction $\hat{\mathbf{t}}$, as shown in Fig. 3. Let P be the intersection point between the incident ray and the reflector and Q the intersection point between the reflected ray and the wave front perpendicular to it and going through $\mathcal{O}$. Then $L(\hat{\mathbf{s}}, \hat{\mathbf{t}}) = u(\hat{\mathbf{s}}) + d(P, Q)$, where we introduce $u(\hat{\mathbf{s}})$ to denote $u(\phi, \theta)$ and $d(P, Q)$ is the distance between P and Q . We see that $d(P, Q)$ is the projection of $u(\hat{\mathbf{s}})$ on $\hat{\mathbf{t}}$ and $d(P, Q) = -(\hat{\mathbf{s}} \cdot \hat{\mathbf{t}}) u(\hat{\mathbf{s}})$, $\hat{\mathbf{s}} \cdot \hat{\mathbf{t}} < 0$, and consequently $L(\hat{\mathbf{s}}, \hat{\mathbf{t}}) = u(\hat{\mathbf{s}}) (1 - \hat{\mathbf{s}} \cdot \hat{\mathbf{t}})$.

$$\tilde{u}_1(\hat{\mathbf{s}}) + \tilde{u}_2(\hat{\mathbf{t}}) = -\log(1 - \hat{\mathbf{s}} \cdot \hat{\mathbf{t}}) = \tilde{c}(\hat{\mathbf{s}}, \hat{\mathbf{t}}). \quad (7)$$

The function $\tilde{c}(\hat{\mathbf{s}}, \hat{\mathbf{t}})$ is known as the cost function in optimal transport. The derivation described above closely resembles the method used by Olikar [26], who constructs the reflector surface as an envelope of paraboloids. An alternative method to

derive the cost function is based on Hamilton's characteristic functions, measuring the optical path length between specified source and target planes [27].

Transforming the cost function in (7) to the stereographic coordinates in equation (3) and (4), defining

$$u_1(\mathbf{x}) = \tilde{u}_1(\hat{\mathbf{s}}) - \log(1 + |\mathbf{x}|^2), \quad u_2(\mathbf{y}) = \tilde{u}_2(\hat{\mathbf{t}}) + \log\left(\frac{2}{1 + |\mathbf{y}|^2}\right), \quad (8)$$

we arrive at the relation

$$u_1(\mathbf{x}) + u_2(\mathbf{y}) = -\log(\mathbf{N}(\mathbf{x}, \mathbf{y})) = c(\mathbf{x}, \mathbf{y}), \quad (9a)$$

where

$$\mathbf{N}(\mathbf{x}, \mathbf{y}) = 1 - 2\mathbf{x} \cdot \mathbf{y} + |\mathbf{x}|^2 |\mathbf{y}|^2. \quad (9b)$$

In summary, we have derived a relation of the form $u_1(\mathbf{x}) + u_2(\mathbf{y}) = c(\mathbf{x}, \mathbf{y})$ for the location of the optical surface u relative to a point source, where $u_1(\mathbf{x}) = \log(u(\mathbf{x})/(1 + |\mathbf{x}|^2))$ and $c(\mathbf{x}, \mathbf{y})$ is a logarithmic cost function in optimal transport theory.

Equation (9a) has many solutions for $u_1(\mathbf{x})$, $u_2(\mathbf{y})$, and consequently for $u(\mathbf{x})$. We can find a solution by assuming that u_1 and u_2 are either c-convex or c-concave functions [28, p. 58]. The surfaces u_1 and u_2 are c-convex if

$$u_1(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{Y}} (c(\mathbf{x}, \mathbf{y}) - u_2(\mathbf{y})), \quad \forall \mathbf{x} \in \mathcal{X}, \quad (10a)$$

$$u_2(\mathbf{y}) = \max_{\mathbf{x} \in \mathcal{X}} (c(\mathbf{x}, \mathbf{y}) - u_1(\mathbf{x})), \quad \forall \mathbf{y} \in \mathcal{Y}, \quad (10b)$$

which we call the maximum solution, or c-concave if

$$u_1(\mathbf{x}) = \min_{\mathbf{y} \in \mathcal{Y}} (c(\mathbf{x}, \mathbf{y}) - u_2(\mathbf{y})), \quad \forall \mathbf{x} \in \mathcal{X}, \quad (11a)$$

$$u_2(\mathbf{y}) = \min_{\mathbf{x} \in \mathcal{X}} (c(\mathbf{x}, \mathbf{y}) - u_1(\mathbf{x})), \quad \forall \mathbf{y} \in \mathcal{Y}, \quad (11b)$$

which we call the minimum solution. For a continuously differentiable function $c \in C^1(\mathcal{X} \times \mathcal{Y})$, the c-convex/concave functions u_1, u_2 are Lipschitz continuous [28, p. 60], and the mapping $\mathbf{y} = \mathbf{m}(\mathbf{x})$ is implicitly given by the critical point of equation (10b) or (11b), i.e.,

$$\nabla u_1(\mathbf{x}) = \nabla_{\mathbf{x}} c(\mathbf{x}, \mathbf{m}(\mathbf{x})), \quad (12)$$

where $\nabla_{\mathbf{x}} c$ is the gradient of c with respect to $\mathbf{x}$, under the condition that the Jacobi matrix $\mathbf{C}$, defined by

$$\begin{aligned} \mathbf{C} = \mathbf{C}(\mathbf{x}, \mathbf{m}(\mathbf{x})) &= D_{\mathbf{x}\mathbf{y}} c = \begin{pmatrix} \frac{\partial^2 c}{\partial x_1 \partial y_1} & \frac{\partial^2 c}{\partial x_1 \partial y_2} \\ \frac{\partial^2 c}{\partial x_2 \partial y_1} & \frac{\partial^2 c}{\partial x_2 \partial y_2} \end{pmatrix} \\ &= \frac{4}{\mathbf{N}(\mathbf{x}, \mathbf{y})^2} (-\mathbf{x} + |\mathbf{x}|^2 \mathbf{y}) (-\mathbf{y} + |\mathbf{y}|^2 \mathbf{x})^T + \frac{2}{\mathbf{N}(\mathbf{x}, \mathbf{y})} (\mathbf{I} - 2\mathbf{x} \mathbf{y}^T) \end{aligned} \quad (13)$$

is invertible.

In fact, we can verify by substituting the expression for c given in (9) into (12) and solving for $\mathbf{y} = \mathbf{m}(\mathbf{x})$, that this implicit mapping is identical to the mapping derived in equation (A.6) via the law of reflection.

By evaluation we note that the matrix $\mathbf{C}$ can be written as $\mathbf{C} = \beta \mathbf{I} + \mathbf{B}$, with β a constant and $\mathbf{B}$ is a skew-symmetric matrix, $\det(\mathbf{C}) > 0$ (which holds if $\hat{\mathbf{s}} \cdot \hat{\mathbf{t}} \neq 1$ and the reflector changes the direction of the light rays, as we have assumed above), and

$$\text{tr}(\mathbf{C}) = \frac{4(-1 + (\mathbf{m}(\mathbf{x}) + \mathbf{J} \mathbf{m}(\mathbf{x})) \cdot \mathbf{x})(-1 + (\mathbf{m}(\mathbf{x}) - \mathbf{J} \mathbf{m}(\mathbf{x})) \cdot \mathbf{x})}{\mathbf{N}(\mathbf{x}, \mathbf{m}(\mathbf{x}))^2} \quad (14)$$

where we introduced the symplectic matrix

$$\mathbf{J} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad (15)$$

which acts upon a vector by rotating it over $\pi/2$ in the counterclockwise direction.

A sufficient condition for a maximum/minimum solution requires

$$D^2 u_1(\mathbf{x}) - D_{\mathbf{x}\mathbf{x}} c(\mathbf{x}, \mathbf{m}(\mathbf{x})) = \mathbf{P} \quad (16)$$

to be symmetric positive/negative semi-definite (SPD/SND), respectively. Hence, for a c-convex pair we require $\text{tr}(\mathbf{P}) \geq 0$ and $\det(\mathbf{P}) \geq 0$. On the other hand, for a c-concave pair we need $\text{tr}(\mathbf{P}) \leq 0$ and $\det(\mathbf{P}) \geq 0$. Note that $\mathbf{P}$ is a symmetric matrix.

Differentiating equation (12) again with respect to $\mathbf{x}$ gives

$$\mathbf{D}_{\mathbf{x}\mathbf{x}}\mathbf{c}(\mathbf{x}, \mathbf{m}(\mathbf{x})) + \mathbf{C} \mathbf{D}\mathbf{m}(\mathbf{x}) = \mathbf{D}^2 u_1(\mathbf{x}), \quad (17)$$

where $\mathbf{D}\mathbf{m}(\mathbf{x})$ is the 2×2 Jacobi matrix of $\mathbf{m}$ with respect to $\mathbf{x}$, with $\mathbf{C}$ defined in (13). Combining equation (16) and (17) gives

$$\mathbf{C}(\mathbf{x}, \mathbf{m}(\mathbf{x})) \mathbf{D}\mathbf{m}(\mathbf{x}) = \mathbf{P}(\mathbf{x}). \quad (18)$$

We will elaborate further on the conditions for $\mathbf{P}(\mathbf{x})$ in Section 3.

2.2. Energy conservation

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

$$\int_{\mathcal{A}} f(\phi, \theta) \, d\mathbf{S}(\phi, \theta) = \int_{\hat{\mathbf{t}}(\mathcal{A})} g(\psi, \chi) \, d\mathbf{S}(\psi, \chi), \quad (19)$$

for an arbitrary set $\mathcal{A} \subset S^2$ and image set $\hat{\mathbf{t}}(\mathcal{A}) \subset S^2$. If we substitute $\hat{\mathbf{s}} = \hat{\mathbf{s}}(\mathbf{x})$ and $\hat{\mathbf{t}} = \hat{\mathbf{t}}(\mathbf{y})$ from equation (5) we can write equation (19) as

$$\int_{\mathbf{x}(\mathcal{A})} \tilde{f}(\mathbf{x}) \left| \frac{\partial \hat{\mathbf{s}}}{\partial x_1} \times \frac{\partial \hat{\mathbf{s}}}{\partial x_2} \right| d\mathbf{x} = \int_{\mathbf{y}(\hat{\mathbf{t}}(\mathcal{A}))} \tilde{g}(\mathbf{y}) \left| \frac{\partial \hat{\mathbf{t}}}{\partial y_1} \times \frac{\partial \hat{\mathbf{t}}}{\partial y_2} \right| d\mathbf{y}. \quad (20)$$

We can derive that

$$\left| \frac{\partial \hat{\mathbf{s}}}{\partial x_1} \times \frac{\partial \hat{\mathbf{s}}}{\partial x_2} \right| = \frac{4}{(1 + |\mathbf{x}|^2)^2}, \quad \left| \frac{\partial \hat{\mathbf{t}}}{\partial y_1} \times \frac{\partial \hat{\mathbf{t}}}{\partial y_2} \right| = \frac{4}{(1 + |\mathbf{y}|^2)^2}. \quad (21)$$

Substituting (21) and the mapping $\mathbf{y} = \mathbf{m}(\mathbf{x})$ into the energy conservation relation (20) gives

$$\int_{\mathbf{x}(\mathcal{A})} \tilde{f}(\mathbf{x}) \frac{4}{(1 + |\mathbf{x}|^2)^2} d\mathbf{x} = \int_{\mathbf{x}(\mathcal{A})} \tilde{g}(\mathbf{m}(\mathbf{x})) \frac{4}{(1 + |\mathbf{m}(\mathbf{x})|^2)^2} |\det(\mathbf{D}\mathbf{m}(\mathbf{x}))| d\mathbf{x}. \quad (22)$$

Using (18) we can rewrite equation (22) to the generalized Monge–Ampère equation

$$\det(\mathbf{D}\mathbf{m}(\mathbf{x})) = \frac{\tilde{f}(\mathbf{x}) (1 + |\mathbf{m}(\mathbf{x})|^2)^2}{\tilde{g}(\mathbf{m}(\mathbf{x})) (1 + |\mathbf{x}|^2)^2} = \frac{\det(\mathbf{P}(\mathbf{x}))}{\det(\mathbf{C}(\mathbf{x}, \mathbf{m}(\mathbf{x})))}, \quad (23)$$

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

Our goal is to find a mapping $\mathbf{y} = \mathbf{m}(\mathbf{x}) : \mathcal{X} \rightarrow \mathcal{Y}$ for a particular source domain $\mathcal{X}$ and target domain $\mathcal{Y}$. We require global energy conservation; i.e., equation (20) should hold with $\mathbf{x}(\mathcal{A}) = \mathcal{X}$ and $\mathbf{y}(\hat{\mathbf{t}}(\mathcal{A})) = \mathcal{Y}$ under any mapping $\mathbf{y} = \mathbf{m}(\mathbf{x}) : \mathcal{X} \rightarrow \mathcal{Y}$. Thus, the light from the entire source is mapped to the target and the reflector functions as a perfect mirror. For this reason, we can not choose any combination of $\mathcal{X}$, $\mathcal{Y}$, $\tilde{f}(\mathbf{x})$ and $\tilde{g}(\mathbf{y})$. We eliminate this dependency by normalizing the source intensity $\tilde{f}(\mathbf{x})$ by the total intensity over $\mathcal{X}$

$$I(\mathcal{X}) = \int_{\mathcal{X}} \tilde{f}(\mathbf{x}) \frac{4}{(1 + |\mathbf{x}|^2)^2} d\mathbf{x}, \quad (24a)$$

and normalizing the target intensity $\tilde{g}(\mathbf{y})$ by the total intensity over $\mathcal{Y}$

$$I(\mathcal{Y}) = \int_{\mathcal{Y}} \tilde{g}(\mathbf{y}) \frac{4}{(1 + |\mathbf{y}|^2)^2} d\mathbf{y}. \quad (24b)$$

Note that we do not need to perform this normalization step if we have chosen $\tilde{f}(\mathbf{x})$ and $\tilde{g}(\mathbf{y})$ such that $I(\mathcal{X}) = I(\mathcal{Y})$. Using (24) we rewrite the generalized Monge–Ampère equation in (23) to

$$\det(\mathbf{D}\mathbf{m}(\mathbf{x})) = \frac{\tilde{f}(\mathbf{x})/I(\mathcal{X})}{\tilde{g}(\mathbf{m}(\mathbf{x}))/I(\mathcal{Y})} \frac{(1 + |\mathbf{m}(\mathbf{x})|^2)^2}{(1 + |\mathbf{x}|^2)^2} = F(\mathbf{x}, \mathbf{m}(\mathbf{x})), \quad (25a)$$

where we introduce $F(\mathbf{x}, \mathbf{m}(\mathbf{x}))$ to equal the total right hand side. We define the corresponding transport boundary condition to (25a) as

$$\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}, \quad (25b)$$

stating that all light from the boundary of the source $\mathcal{X}$ is mapped to the boundary of the target $\mathcal{Y}$ [7,10], which is a consequence of the edge-ray principle [29] and explained in detail in Appendix B.

Summarizing, we aim to compute $\mathbf{m}$ from (25) and subsequently u_1 from (12). In Villani [4, p. 282], the Monge–Ampère equation (25a) and equation (12) are together referred to as the basic PDE of optimal transport. For results on the existence, regularity and uniqueness of solutions, see [4].

3. Numerical method

We first compute the mapping $\mathbf{m}$ from (25) by using a generalized least-squares method [7,10,25]. The mapping $\mathbf{m}$ can be calculated efficiently by an iterative procedure that involves finding the numerical solution of a constrained minimization problem, imposing the transport boundary condition using skew projections on line segments, and computing the numerical solution of a linear elliptic boundary value problem. Upon convergence the location of the optical surface u is calculated from the mapping using equation (12) also in a least-squares sense.

To compute a c-convex solution, we can write the Monge–Ampère equation (25a) as the matrix equation (18), with $\mathbf{P}(\mathbf{x})$ a SPD matrix satisfying $\det(\mathbf{P}(\mathbf{x})) = F(\mathbf{x}, \mathbf{m}(\mathbf{x})) \det(\mathbf{C}(\mathbf{x}, \mathbf{m}(\mathbf{x})))$. We write $\mathbf{m} = \mathbf{m}(\mathbf{x})$ and enforce the matrix equation (18) by minimizing the functional

$$J_I[\mathbf{m}, \mathbf{P}] = \frac{1}{2} \iint_{\mathcal{X}} \|\mathbf{C} \mathbf{D} \mathbf{m} - \mathbf{P}\|^2 d\mathbf{x}, \quad (26)$$

under the constraint $\det(\mathbf{P}) = F \det(\mathbf{C})$. The norm used is the Frobenius norm. To impose the transport boundary condition (25b) we minimize the functional

$$J_B[\mathbf{m}, \mathbf{b}] = \frac{1}{2} \oint_{\partial\mathcal{X}} |\mathbf{m} - \mathbf{b}|^2 ds \quad (27)$$

over $\mathbf{b}$, where $|\cdot|$ denotes the L_2 -norm and $\mathbf{b}$ is a function from the source boundary to the target boundary, i.e., $\mathbf{b} : \partial\mathcal{X} \rightarrow \partial\mathcal{Y}$. By minimizing this functional we aim to impose $\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}$, which holds if $J_B[\mathbf{m}, \mathbf{b}] = 0$. This is equivalent, under some restrictions, to $\mathbf{m}(\mathcal{X}) = \mathcal{Y}$, stating that all the light from the source arrives at the target; see Appendix B. We combine the functionals J_I and J_B by a weighted average as

$$J[\mathbf{m}, \mathbf{P}, \mathbf{b}] = \alpha J_I[\mathbf{m}, \mathbf{P}] + (1 - \alpha) J_B[\mathbf{m}, \mathbf{b}], \quad (28)$$

with $0 < \alpha < 1$.

Starting from an initial guess $\mathbf{m}^0$ and cost function matrix $\mathbf{C}(\cdot, \mathbf{m}^0)$ we perform the iteration:

$$\mathbf{b}^{n+1} = \operatorname{argmin}_{\mathbf{b} \in \mathcal{B}} J_B[\mathbf{m}^n, \mathbf{b}], \quad (29a)$$

$$\mathbf{P}^{n+1} = \operatorname{argmin}_{\mathbf{P} \in \mathcal{P}(\mathbf{m}^n)} J_I[\mathbf{m}^n, \mathbf{P}], \quad (29b)$$

$$\mathbf{m}^{n+1} = \operatorname{argmin}_{\mathbf{m} \in \mathcal{M}} J[\mathbf{m}, \mathbf{P}^{n+1}, \mathbf{b}^{n+1}], \quad (29c)$$

where the minimization steps are performed over the spaces

$$\mathcal{B} = \{\mathbf{b} \in C^1(\partial\mathcal{X})^2 \mid \mathbf{b}(\mathbf{x}) \in \partial\mathcal{Y}\}, \quad (30a)$$

$$\mathcal{P}(\mathbf{m}) = \{\mathbf{P} \in C^1(\mathcal{X})^{2 \times 2} \mid \mathbf{P} \text{ SPD}, \det(\mathbf{P}) = F(\cdot, \mathbf{m}) \det(\mathbf{C}(\cdot, \mathbf{m}))\}, \quad (30b)$$

$$\mathcal{M} = C^2(\mathcal{X})^2, \quad (30c)$$

where the smoothness of the spaces is the required smoothness for our numerical algorithm; see Section 3.3. After each iteration we update the matrix $\mathbf{C}(\cdot, \mathbf{m}^n)$. As initial guess $\mathbf{m}^0$ we map the smallest bounding box enclosing $\mathcal{X}$ to the smallest bounding box enclosing $\mathcal{Y}$. Without loss of generality we assume the bounding box of the source $\mathcal{X}$ has rectangular shape $[a_{\min}, a_{\max}] \times [b_{\min}, b_{\max}]$ and the bounding box of the target $\mathcal{Y}$ has rectangular shape $[c_{\min}, c_{\max}] \times [d_{\min}, d_{\max}]$. In order to find a c-convex u_1 , we specify the initial guess $\mathbf{m}^0 = (m_1^0, m_2^0)^T$ as

$$m_1^0 = \frac{x_1 - a_{\min}}{a_{\max} - a_{\min}} c_{\max} + \frac{a_{\max} - x_1}{a_{\max} - a_{\min}} c_{\min}, \quad (31a)$$

$$m_2^0 = \frac{x_2 - b_{\min}}{b_{\max} - b_{\min}} d_{\max} + \frac{b_{\max} - x_2}{b_{\max} - b_{\min}} d_{\min}. \quad (31b)$$

The corresponding Jacobi matrix $\mathbf{D}\mathbf{m}^0$ is symmetric positive definite.

If we would like to find a c-concave u_1 instead, we perform the minimization over SND matrices $\mathbf{P}$ and choose a slightly different initial guess $\mathbf{m}^0 = (m_1^0, m_2^0)^T$ given by

$$m_1^0 = \frac{x_1 - a_{\min}}{a_{\max} - a_{\min}} c_{\min} + \frac{a_{\max} - x_1}{a_{\max} - a_{\min}} c_{\max}, \quad (32a)$$

$$m_2^0 = \frac{x_2 - b_{\min}}{b_{\max} - b_{\min}} d_{\min} + \frac{b_{\max} - x_2}{b_{\max} - b_{\min}} d_{\max}. \quad (32b)$$

The corresponding Jacobi matrix $\mathbf{D}\mathbf{m}^0$ is symmetric negative definite.

Using the initial mapping in (31) or (32), we can show that $\det(\mathbf{D}\mathbf{m}^0) > 0$, and that $\text{tr}(\mathbf{P}^0) = \text{tr}(\mathbf{C}(\cdot, \mathbf{m}^0)) \text{tr}(\mathbf{D}\mathbf{m}^0)$, since the diagonal elements of $\mathbf{C}$ are equal and $\mathbf{D}\mathbf{m}^0$ is a diagonal matrix. Substituting $\mathbf{m}^0$ into (14) we find that $\text{tr}(\mathbf{C}(\cdot, \mathbf{m}^0)) \geq 0$. Hence, using the initial mapping in (31) or (32), $\mathbf{P}^0 = \mathbf{C}(\cdot, \mathbf{m}^0) \mathbf{D}\mathbf{m}^0$ is positive or negative semi-definite, respectively.

We remark that we need to choose a source domain and target domain such that $\mathbf{N}(\cdot, \mathbf{m}^0) > 0$ in (9b) and $\mathbf{C}(\cdot, \mathbf{m}^0)$ is invertible.

We discretize the source domain $\mathcal{X}$ using a standard rectangular $N_1 \times N_2$ grid for some $N_1, N_2 \in \mathbb{N}$ and introduce $\mathbf{x}_{ij} = (x_{1,i}, x_{2,j})$ with

$$x_{1,i} = a_{\min} + (i - 1) h_1, \quad h_1 = \frac{a_{\max} - a_{\min}}{N_1 - 1}, \quad i = 1, \dots, N_1, \quad (33a)$$

$$x_{2,j} = b_{\min} + (j - 1) h_2, \quad h_2 = \frac{b_{\max} - b_{\min}}{N_2 - 1}, \quad j = 1, \dots, N_2. \quad (33b)$$

After setting the initial guess $\mathbf{m}^0$ we perform the minimization steps in (29) and subsequently update $\mathbf{C}$ in every iteration. The minimization steps (29a), (29b), and (29c) are explained in detail in Section 3.1, 3.2, 3.3, respectively. Finally, we compute the location of the reflector surface u as described in Section 3.4.

3.1. Minimization procedure for $\mathbf{b}$

In this section, we introduce a novel method to impose the transport boundary equation. It is a modification to the method described in [7,10].

We assume $\mathbf{m} = \mathbf{m}^n$ is given and we need to minimize $J_B[\mathbf{m}, \mathbf{b}]$ over $\mathbf{b} \in \mathbf{B}$. The minimization can be performed point-wise because the integrand does not depend on derivatives of $\mathbf{b}$. We drop the indices n and $n + 1$ for ease of notation. We denote $\mathbf{m}_{ij} = \mathbf{m}(\mathbf{x}_{ij})$, $\mathbf{b}_{ij} = \mathbf{b}(\mathbf{x}_{ij})$ and perform the minimization

$$\min_{\mathbf{b}_{ij} \in \mathbf{B}} \frac{1}{2} |\mathbf{m}_{ij} - \mathbf{b}_{ij}|^2. \quad (34)$$

We discretize the boundary of $\mathcal{Y}$ using points $\mathbf{z}_k \in \partial\mathcal{Y}$, ($k = 1, 2, \dots, N_b$) with increasing index clockwise along the boundary, and we define $\mathbf{z}_{N_b+1} = \mathbf{z}_1$. We connect adjacent points by line segments $(\mathbf{z}_k, \mathbf{z}_{k+1})$ and determine the “closest” line segment to each $\mathbf{m}_{ij}$.

First, we define the outward normals $\mathbf{n}_k$ associated with each boundary point $\mathbf{z}_k$ as

$$\mathbf{n}_k = \frac{1}{2} \left(\frac{1}{|\mathbf{z}_{k+1} - \mathbf{z}_k|} \mathbf{J}(\mathbf{z}_{k+1} - \mathbf{z}_k) + \frac{1}{|\mathbf{z}_k - \mathbf{z}_{k-1}|} \mathbf{J}(\mathbf{z}_k - \mathbf{z}_{k-1}) \right), \quad (35)$$

where $\mathbf{J}$ is defined as in (15), thus $\mathbf{n}_k$ is the vector pointing in the average direction of the normals to the two adjacent line segments $\mathbf{z}_{k+1} - \mathbf{z}_k$ and $\mathbf{z}_k - \mathbf{z}_{k-1}$. It bisects the angle between the adjacent segments.

We define $l_k : \mathbf{y} = \mathbf{z}_k + \lambda \mathbf{n}_k$, ($k = 1, 2, \dots, N_b$), as the points $\mathbf{y} \in \mathbb{R}^2$ which are on the bisector through $\mathbf{z}_k$, and let ℓ be the line through $\mathbf{m}_{ij}$, parallel to the segment $(\mathbf{z}_k, \mathbf{z}_{k+1})$. We let $\mathbf{p}_k$ and $\mathbf{p}_{k+1}$ be the intersection points of ℓ with l_k and l_{k+1} , respectively, as shown in Fig. 4.

The intersection points $\mathbf{p}_k$ and $\mathbf{p}_{k+1}$ of ℓ with l_k and l_{k+1} , respectively, can be written as

$$\mathbf{p}_k = \mathbf{z}_k + k_1 \mathbf{n}_k, \quad (36a)$$

$$\mathbf{p}_{k+1} = \mathbf{z}_{k+1} + k_2 \mathbf{n}_{k+1}, \quad (36b)$$

with k_1 and k_2 constants. We can solve for k_1 and k_2 by setting $\mathbf{m}_{ij} - \mathbf{p}_k // \mathbf{z}_{k+1} - \mathbf{z}_k$ and $\mathbf{m}_{ij} - \mathbf{p}_{k+1} // \mathbf{z}_{k+1} - \mathbf{z}_k$, giving

$$k_1 = \frac{\det(\mathbf{m}_{ij} - \mathbf{z}_k, \mathbf{z}_{k+1} - \mathbf{z}_k)}{\det(\mathbf{n}_k, \mathbf{z}_{k+1} - \mathbf{z}_k)}, \quad (36c)$$

$$k_2 = \frac{\det(\mathbf{m}_{ij} - \mathbf{z}_{k+1}, \mathbf{z}_{k+1} - \mathbf{z}_k)}{\det(\mathbf{n}_{k+1}, \mathbf{z}_{k+1} - \mathbf{z}_k)}. \quad (36d)$$

Fig. 4. Calculation of the distance of $\mathbf{m}_{ij}$ to a line segment $(\mathbf{z}_k, \mathbf{z}_{k+1})$ with the point $\mathbf{m}_{i,j}^k$ as the projection of $\mathbf{m}_{ij}$ on the line through $\mathbf{z}_k$ and $\mathbf{z}_{k+1}$.

The bisector lines l_k and l_{k+1} may cross as illustrated in Fig. 4b. We determine whether this occurs by evaluating if both of the following two conditions hold for each line segment

$$d_k(\mathbf{p}_{k+1}) d_k(\mathbf{z}_{k+1}) < 0, \quad (37a)$$

$$d_{k+1}(\mathbf{p}_k) d_{k+1}(\mathbf{z}_k) < 0, \quad (37b)$$

where

$$d_k(\mathbf{y}) = \det(\mathbf{p}_k - \mathbf{z}_k, \mathbf{y} - \mathbf{z}_k), \quad (37c)$$

$$d_{k+1}(\mathbf{y}) = \det(\mathbf{p}_{k+1} - \mathbf{z}_{k+1}, \mathbf{y} - \mathbf{z}_{k+1}). \quad (37d)$$

Note that the line l_k through $\mathbf{z}_k$ and $\mathbf{p}_k$ are the points $\mathbf{y}$ for which $d_k(\mathbf{y}) = 0$. Likewise, the line l_{k+1} through $\mathbf{z}_{k+1}$ and $\mathbf{p}_{k+1}$ are the points $\mathbf{y}$ for which $d_{k+1}(\mathbf{y}) = 0$. The first condition (37a) checks whether $\mathbf{p}_{k+1}$ and $\mathbf{z}_{k+1}$ are located on opposite sides of the line segment $\mathbf{p}_k - \mathbf{z}_k$. Likewise, the second condition (37b) checks whether $\mathbf{p}_k$ and $\mathbf{z}_k$ are located on opposite sides of $\mathbf{p}_{k+1} - \mathbf{z}_{k+1}$. Together they determine whether $\mathbf{p}_k - \mathbf{z}_k$ crosses $\mathbf{p}_{k+1} - \mathbf{z}_{k+1}$.

The projection $\mathbf{m}_{i,j}^k$ of $\mathbf{m}_{ij}$ on the line segment $(\mathbf{z}_k, \mathbf{z}_{k+1})$ is given by

$$\mathbf{m}_{i,j}^k = \mathbf{z}_k + t_k (\mathbf{z}_{k+1} - \mathbf{z}_k), \quad (38a)$$

$$t_k = \frac{|\mathbf{m}_{ij} - \mathbf{p}^*|}{|\mathbf{p}_{k+1} - \mathbf{p}_k|}, \quad (38b)$$

where, if at least one of the inequalities (37a) and (37b) does not hold, i.e., the bisector lines do not cross as in Fig. 4a, we define $\mathbf{p}^* = \mathbf{p}_k$, and if (37a) and (37b) are both true, i.e., the bisector lines do cross as in Fig. 4b, we use $\mathbf{p}^* = \mathbf{p}_{k+1}$.

For a given line segment $(\mathbf{z}_k, \mathbf{z}_{k+1})$, we check if $\mathbf{m}_{ij}$ is located in between $\mathbf{p}_k$ and $\mathbf{p}_{k+1}$ by evaluating

$$0 \leq (\mathbf{m}_{ij} - \mathbf{p}_k) \cdot (\mathbf{p}_{k+1} - \mathbf{p}_k) \leq |\mathbf{p}_{k+1} - \mathbf{p}_k|^2. \quad (39)$$

If this does not hold, we set the parameter t_k in (38b) to an arbitrarily large number to exclude $\mathbf{m}_{ij}$ in the minimization procedure below.

The skew projection (38) ensures that the ratio between the distances $|\mathbf{m}_{ij} - \mathbf{p}^*|$ and $|\mathbf{p}_{k+1} - \mathbf{p}_k|$ is the same as the ratio between the distances $|\mathbf{m}_{i,j}^k - \mathbf{z}_k|$ and $|\mathbf{z}_{k+1} - \mathbf{z}_k|$.

Finally, we calculate the points $\mathbf{b}_{ij}$ as

$$\mathbf{b}_{ij} = \operatorname{argmin}_{\mathbf{m}_{i,j}^k} \frac{1}{2} |\mathbf{m}_{ij} - \mathbf{m}_{i,j}^k|^2. \quad (40)$$

This procedure is repeated for all $\mathbf{x}_{ij} \in \partial\mathcal{X}$.

3.2. Minimization procedure for $\mathbf{P}$

We assume $\mathbf{m} = \mathbf{m}^n$ is fixed. The minimization of $J_I[\mathbf{m}, \mathbf{P}]$ can be performed point-wise because the integrand does not depend on derivatives of $\mathbf{P}$. We minimize $\|\mathbf{Q} - \mathbf{P}\|$ for each grid point $\mathbf{x}_{ij} \in \mathcal{X}$, where $\mathbf{Q} = \mathbf{C}\mathbf{D} = (q_{ij})$, and $\mathbf{D} = (d_{ij})$ is the central difference approximation of $\mathbf{Dm}$. This gives rise to the minimization problem

$$\text{minimize} \quad H_S(p_{11}, p_{22}, p_{12}) = \frac{1}{2} \|\mathbf{Q}_S - \mathbf{P}\|^2, \quad (41a)$$

$$\text{subject to} \quad \det(\mathbf{P}) = p_{11} p_{22} - p_{12}^2 = F(\cdot, \mathbf{m}) \det(\mathbf{C}(\cdot, \mathbf{m})), \quad (41b)$$

where $\mathbf{Q}_S = \frac{1}{2}(\mathbf{Q} + \mathbf{Q}^T)$. We have replaced $\mathbf{Q}$ by its symmetric part $\mathbf{Q}_S$, since this gives the same minimizer (p_{11}, p_{22}, p_{12}) . For a c-convex solution we impose the additional constraint

$$\text{tr}(\mathbf{P}) = p_{11} + p_{22} \geq 0, \quad (41c)$$

while for a c-concave solution we require

$$\text{tr}(\mathbf{P}) = p_{11} + p_{22} \leq 0. \quad (41d)$$

Possible solutions are stationary points of the Lagrangian Λ defined as

$$\Lambda(p_{11}, p_{22}, p_{12}; \mu) = \frac{1}{2} \|\mathbf{Q}_S - \mathbf{P}\|^2 + \mu (\det(\mathbf{P}) - F(\cdot, \mathbf{m}) \det(\mathbf{C}(\cdot, \mathbf{m}))). \quad (42)$$

Setting all partial derivatives of Λ to 0 results in the algebraic system

$$p_{11} + \lambda p_{22} = q_{11}, \quad (43a)$$

$$\lambda p_{11} + p_{22} = q_{22}, \quad (43b)$$

$$(1 - \lambda) p_{12} = \frac{1}{2} (q_{12} + q_{21}), \quad (43c)$$

$$p_{11} p_{22} - p_{12}^2 = F(\cdot, \mathbf{m}) \det(\mathbf{C}(\cdot, \mathbf{m})), \quad (43d)$$

where $\lambda = \mu / \det(\mathbf{C}(\cdot, \mathbf{m}))$. We can always select the possible minimizers that satisfy the nonlinear constraints (41b), and the inequality constraint (41c) or (41d) [25]. Details of the remaining analytical procedure to select the minimizers are presented in [7,10,25].

3.3. Minimization procedure for $\mathbf{m}$

We minimize the combined functional $J[\mathbf{m}, \mathbf{P}, \mathbf{b}]$ over all $\mathbf{m} \in \mathcal{M}$. This step cannot be performed point-wise and we compute the first variation $\delta J[\mathbf{m}, \mathbf{P}, \mathbf{b}](\eta)$ with respect to $\mathbf{m}$ for $\eta \in \mathcal{M}$, i.e.,

$$\begin{aligned} \delta J[\mathbf{m}, \mathbf{P}, \mathbf{b}](\eta) &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \left(J[\mathbf{m} + \epsilon \eta, \mathbf{P}, \mathbf{b}] - J[\mathbf{m}, \mathbf{P}, \mathbf{b}] \right) \\ &= \lim_{\epsilon \rightarrow 0} \left[\frac{\alpha}{2} \int_{\mathcal{X}} 2 (\mathbf{C} \mathbf{D} \mathbf{m} - \mathbf{P}) : \mathbf{C} \mathbf{D} \eta + \epsilon \|\mathbf{C} \mathbf{D} \eta\|^2 \, d\mathbf{x} \right. \\ &\quad \left. + \frac{1 - \alpha}{2} \oint_{\partial \mathcal{X}} 2 (\mathbf{m} - \mathbf{b}) \cdot \eta + \epsilon |\eta|^2 \, ds \right] \\ &= \alpha \int_{\mathcal{X}} (\mathbf{C} \mathbf{D} \mathbf{m} - \mathbf{P}) : \mathbf{C} \mathbf{D} \eta \, d\mathbf{x} + (1 - \alpha) \oint_{\partial \mathcal{X}} (\mathbf{m} - \mathbf{b}) \cdot \eta \, ds. \end{aligned} \quad (44)$$

The minimizer is given by $\delta J[\mathbf{m}, \mathbf{P}, \mathbf{b}](\eta) = 0$ for all $\eta \in \mathcal{M}$. Using Gauss' theorem and the fundamental lemma of calculus of variations [30], we obtain the coupled elliptic boundary value problem

$$\nabla \cdot (\mathbf{C}^T \mathbf{C} \mathbf{D} \mathbf{m}) = \nabla \cdot (\mathbf{C}^T \mathbf{P}), \quad \mathbf{x} \in \mathcal{X}, \quad (45a)$$

$$(1 - \alpha) \mathbf{m} + \alpha (\mathbf{C}^T \mathbf{C} \nabla \mathbf{m}) \cdot \hat{\mathbf{n}} = (1 - \alpha) \mathbf{b} + \alpha \mathbf{C} \cdot \mathbf{P} \hat{\mathbf{n}}, \quad \mathbf{x} \in \partial \mathcal{X}. \quad (45b)$$

We solve this system for $\mathbf{m}$ using the finite volume method [25].

The required smoothness for the minimization spaces in (29) can be derived from (45). In equation (45a), we need $\mathbf{m}$ to be at least twice differentiable and $\mathbf{P}$ to be at least once differentiable. From equation (45b), we require $\mathbf{b}$ to be at least once differentiable since first order partial derivatives of $\mathbf{m}$ appear in the Robin boundary condition and $\mathbf{m}$ is at least twice differentiable. Rigorous regularity results of solutions to the basic PDE of optimal transport in (12) and (25) are beyond the scope of this paper and we refer to [4, Ch. 12] for a detailed overview.

3.4. Computation of u

Upon convergence of the iterative procedure for the mapping $\mathbf{m}$, we calculate the location of the optical surface from equation (12). We compute the generalized least-squares solution by minimizing the functional

$$I[u_1] = \frac{1}{2} \int_{\mathcal{X}} |\nabla u_1 - \nabla_{\mathbf{x}} c(\cdot, \mathbf{m})|^2 d\mathbf{x}. \quad (46)$$

We cannot perform this step point-wise and analogous to the minimization procedure for $\mathbf{m}$, we compute the first variation of $\delta I[u_1](v)$ for $v \in C^2(\mathcal{X})$ as

$$\begin{aligned} \delta I[u_1](v) &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \left(I[u_1 + \epsilon v] - I[u_1] \right) \\ &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \int_{\mathcal{X}} 2 (\nabla u_1 - \nabla_{\mathbf{x}} c(\cdot, \mathbf{m})) \cdot \nabla v + \epsilon |v|^2 d\mathbf{x} \\ &= \int_{\mathcal{X}} (\nabla u_1 - \nabla_{\mathbf{x}} c(\cdot, \mathbf{m})) \cdot \nabla v d\mathbf{x}. \end{aligned} \quad (47)$$

The minimizer is given by $\delta I[u_1](v) = 0$ for all $v \in C^2(\mathcal{X})$. Once more, using Gauss' theorem and the fundamental lemma of calculus of variations [30], we obtain the boundary value problem

$$\Delta u_1 = \nabla \cdot \nabla_{\mathbf{x}} c(\cdot, \mathbf{m}), \quad \mathbf{x} \in \mathcal{X}, \quad (48a)$$

$$\nabla u_1 \cdot \hat{\mathbf{n}} = \nabla_{\mathbf{x}} c(\cdot, \mathbf{m}) \cdot \hat{\mathbf{n}}, \quad \mathbf{x} \in \partial \mathcal{X}. \quad (48b)$$

This is a Neumann problem which has a unique solution up to an additive constant, and consequently a corresponding finite difference matrix with incomplete rank. We calculate a unique least-squares solution by using the QR -decomposition of the finite difference matrix. The compatibility condition is satisfied for this Neumann problem up to discretization errors. An alternative method to compute a unique solution is to fix the value of u_1 at one point. However, we find that this concentrates the residual in the compatibility condition at one point, instead of smoothing it out over the whole domain.

Finally, we compute the location of the optical surface u as a function of $\mathbf{x}_{ij}$ using $u(\mathbf{x}) = e^{u_1(\mathbf{x})} (1 + |\mathbf{x}|^2)$, since $u_1 = \log(u/(1 + |\mathbf{x}|^2))$. We transform our stereographic source coordinates defined in (3) to Cartesian coordinates denoting $(x, y, z) = u(\hat{\mathbf{s}}) \hat{\mathbf{s}}$ and plot the reflector surface using

$$x = \frac{2u(\mathbf{x})x_1}{1+x_1^2+x_2^2}, \quad y = \frac{2u(\mathbf{x})x_2}{1+x_1^2+x_2^2}, \quad z = \frac{u(\mathbf{x})(1-x_1^2-x_2^2)}{1+x_1^2+x_2^2}, \quad (49)$$

cf. (5), for all $\mathbf{x}_{ij} = (x_{1,i}, x_{2,j}) \in \mathcal{X}$.

4. Numerical results

We test our generalized least-squares method on a number of example problems. We compute the correct surface for a number of mappings corresponding to flat or spherical reflector surfaces. We also compute the freeform surface that transforms light into an image on a screen in the far field.

4.1. Comparison of boundary methods

The idea of the skew projection arose through preliminary experiments. With sharp-cornered target domains, we noticed that orthogonal projections do not uniformly distribute the points $\mathbf{b}$ along the boundary. Through orthogonal projections the points $\mathbf{b}$ are essentially 'not allowed to turn a sharp corner' but accumulate at the end of a segment. With skew projections this effect is mitigated resulting in a better and faster convergence in the corner points of the target domain.

We compare the minimization procedure for $\mathbf{b}$ described in Section 3.1 to the perpendicular projection previously presented in [10]. In our test problem, the source domain is given by the square $\mathcal{X} = [-0.5, 0.5]^2$ and the target domain by the parallelogram centred in the origin spanned by a vector $\mathbf{w}$ of length 1 parallel to the y_1 -axis and a vector $\mathbf{u}$ of length 1 which forms an angle of $\pi/6$ with $\mathbf{w}$, i.e., $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^2 \mid -1/2 + y_2\sqrt{3} \leq y_1 \leq 1/2 + y_2\sqrt{3}, -1/4 \leq y_2 \leq 1/4\}$. We have taken a uniform source distribution $\tilde{f}(\mathbf{x}) = 1$ and uniform target distribution $\tilde{g}(\mathbf{y}) = 1$, and in order to satisfy global energy conservation we calculate $I(\mathcal{X}) = 16/\sqrt{5} \cot^{-1}(\sqrt{5})$ and $I(\mathcal{Y}) \approx 1.5533$ using (24). ($I(\mathcal{Y})$ is evaluated by using MATLAB's inbuilt double numerical integration method). We solve the boundary value problem (25) for a 100×100 grid and 4 corner boundary points. With 4 boundary points each side of the parallelogram is one segment $(\mathbf{z}_k, \mathbf{z}_{k+1})$. We use the initial guess $\mathbf{m}^0$ given in (32) for a c -concave u_1 , in order to ensure $\mathbf{C}(\cdot, \mathbf{m}^0)$ is invertible. The convergence of J_I and J_B is shown in Fig. 5c and the final mappings are shown in Fig. 5a and Fig. 5b. Using a skew projection results in a better and faster convergence in the corner points of the target domain. Fig. 5d shows the influence of varying N_b on the final values of J_I and J_B , after 2000 iterations. The orthogonal and skew projections become equivalent when the number of boundary points reaches approximately 200, due to the fact that the skew projection becomes more and more orthogonal as the lengths of the line segments decrease.

Fig. 5. "Square-to-parallelogram" problem: the mapping on a 100×100 grid with $\alpha = 0.2$ and $N_b = 4$, $\tilde{f}(\mathbf{x}) = 1$ and $\tilde{g}(\mathbf{y}) = 1$.

4.2. Exact solution: tilted flat surface

One way to test our algorithm is by pre-computing the target domains corresponding to a surface of which we can derive the correct mapping. For example, we consider a tilted flat reflector surface $ax + by + cz = d$, with given constants a, b, c, d , and $c \neq 0, cd > 0$. By substituting $x = u s_1, y = u s_2, z = u s_3$ and $u = e^{u_1} (1 + |\mathbf{x}|^2)$, and changing to stereographic coordinates using (5), we can derive an expression for the c-convex/c-concave solution u_1

$$u_1(\mathbf{x}) = \log(d) - \log\left(2\mathbf{a} \cdot \mathbf{x} + c(1 - |\mathbf{x}|^2)\right), \quad \mathbf{a} = \begin{pmatrix} a \\ b \end{pmatrix}. \quad (50)$$

Note that we require our source domain $\mathcal{X}$ to lie within the interior of the ellipse $c^2 |\mathbf{x} - \frac{\mathbf{a}}{c}|^2 = |\mathbf{a}|^2 + c^2$, since we require that $2\mathbf{a} \cdot \mathbf{x} + c(1 - |\mathbf{x}|^2) > 0$. Subsequently calculating $\frac{\partial u_1}{\partial x_1}$ and $\frac{\partial u_1}{\partial x_2}$, and substituting into equation (A.6) we obtain the mapping as

$$\mathbf{y} = \mathbf{m}(\mathbf{x}) = \frac{\mathbf{B}(\mathbf{x}) \mathbf{x} + \mathbf{a} c (-1 + |\mathbf{x}|^2)}{c^2 + 2\mathbf{a} \cdot \mathbf{x} + |\mathbf{a}|^2 |\mathbf{x}|^2}, \quad (51a)$$

where

$$\mathbf{B}(\mathbf{x}) = \begin{pmatrix} -a^2 + b^2 + c^2 & -2ab \\ 2ab & a^2 - b^2 + c^2 \end{pmatrix}. \quad (51b)$$

Fig. 6. “Tilted-flat-surface” problem: the mapping and reflector surface after 1000 iterations on a 100×100 grid with $\alpha = 0.2$ and $N_b = 1000$, $\tilde{f}(\mathbf{x}) = 1$ and $\tilde{g}(\mathbf{y}) = 1$. The absolute difference between the computed reflector surface u_1 and exact solution u_1^* is displayed in Fig. 6d.

We consider a square source domain $\mathcal{X} = [-0.2, 0.2]^2$ and a 100×100 grid with $N_b = 1000$. We choose $a = 2$, $b = 1$, $c = 3$, and $d = 1$. Using equation (51) we can compute the target domain $\mathcal{Y}$, the corresponding target boundary, and the correct mapping. Next, we solve the boundary value problem (25) for a uniform source and target distribution, i.e., $\tilde{f}(\mathbf{x}) = 1$ on $\mathcal{X}$ and $\tilde{g}(\mathbf{y}) = 1$ on $\mathcal{Y}$. In order to satisfy global energy conservation we calculate $I(\mathcal{X})$ and $I(\mathcal{Y})$. In fact, we know that $I(\mathcal{X}) = I(\mathcal{Y})$ is satisfied already by our choice of target domain $\mathcal{Y}$ corresponding to the mapping of a flat surface and choice of equal uniform intensities. We use the initial guess $\mathbf{m}^0$ given in (31) for a c-convex u_1 . Fig. 6 shows the results after 1000 iterations. The converged mapping is displayed in Fig. 6a and the reflector surface upon convergence in Fig. 6b. The error convergence is given in Fig. 6c. The maximum absolute difference between the reflector surface and the exact solution restricted to the grid after 1000 iterations is shown in Fig. 6d as a function of grid size $N = N_1 = N_2$. The slope of a logarithmic least-squares fit indicates second-order convergence to the exact surface.

4.3. Square-to-circle problem

We consider a square source domain $\mathcal{X} = [-0.5, 0.5]^2$ and circular target domain $\mathcal{Y} = \{(y_1, y_2) \in \mathbb{R}^2 \mid |y|^2 \leq 0.5\}$. We solve the boundary value problem (25) for a uniform source and target distribution, i.e., $\tilde{f}(\mathbf{x}) = 1$ on $\mathcal{X}$ and $\tilde{g}(\mathbf{y}) = 1$ on $\mathcal{Y}$. In order to satisfy global energy conservation we calculate $I(\mathcal{X}) = 16/\sqrt{5} \cot^{-1}(\sqrt{5})$ and $I(\mathcal{Y}) = \frac{4}{5}\pi$ using equation (24). We use the initial guess $\mathbf{m}^0$ given in (32) for a c-concave u_1 . Fig. 7a shows the converged mapping on a 50×50 grid.

Fig. 7. “Square-to-circle” problem: convergence history for several grid sizes. We calculate a c-concave solution u_1 and parameter values are $\alpha = 0.2$, $N_b = 1000$.

To investigate convergence of the numerical algorithm, we introduce the norms

$$\|A\|_{2 \times 2} = \left(\iint_{\mathcal{X}} \|A\|^2 d\mathbf{x} \right)^{1/2}, \quad \|a\|_2 = \left(\oint_{\partial \mathcal{X}} |a|^2 ds \right)^{1/2}, \quad (52)$$

for $A \in [C^1(\mathcal{X})]^{2 \times 2}$ and $a \in [C^1(\partial \mathcal{X})]^2$, as described in [10]. Let $J_I^n = J_I[\mathbf{m}^n, \mathbf{P}^n]$ and $J_B^n = J_B[\mathbf{m}^n, \mathbf{b}^n]$. We can derive that

$$\left| \sqrt{J_I^{n+1}} - \sqrt{J_I^n} \right| \leq \frac{1}{\sqrt{2}} \left(\|C^{n+1} Dm^{n+1} - C^n Dm^n\|_{2 \times 2} + \|\mathbf{P}^{n+1} - \mathbf{P}^n\|_{2 \times 2} \right) := c_I^n, \quad (53a)$$

$$\left| \sqrt{J_B^{n+1}} - \sqrt{J_B^n} \right| \leq \frac{1}{\sqrt{2}} \left(\|\mathbf{m}^{n+1} - \mathbf{m}^n\|_2 + \|\mathbf{b}^{n+1} - \mathbf{b}^n\|_2 \right) := c_B^n. \quad (53b)$$

Fig. 7 shows J_I , J_B for several $N \times N$ grids. It also shows the changes in CDm , $m|_{\partial \mathcal{X}}$ (m on the boundary), $\mathbf{P}$ and $\mathbf{b}$, with the updates in $\mathbf{P}$ and $\mathbf{b}$ plotted from the second iteration onwards (since they have no initial values). We used $N_b = 1000$ and $\alpha = 0.2$. The functionals J_I and J_B reach a plateau at a certain iteration number while the individual error terms continue to decrease up to machine precision. For the numerical simulations in the current section we introduce the stopping criterion

$$c_I^n \leq 0.1 \sqrt{J_I^n}, \quad \text{and} \quad c_B^n \leq 0.1 \sqrt{J_B^n}. \quad (54)$$

Table 1

Number of iterations, computing time (in seconds) and residuals in the least-squares algorithm for the “Square-to-circle” problem.

Grids	100 × 100	200 × 200	300 × 300	400 × 400	500 × 500	Fits
Iterations	208	351	466	563	650	$\propto N^{0.7}$
Time	22	115	333	772	1370	$\propto N^{2.6}$
J_I	4.0×10^{-7}	3.9×10^{-8}	1.2×10^{-8}	5.5×10^{-9}	3.3×10^{-9}	$\propto N^{-3.0}$
J_B	1.9×10^{-7}	4.6×10^{-8}	2.4×10^{-8}	1.6×10^{-8}	1.2×10^{-8}	$\propto N^{-1.7}$

Fig. 8. “Square-to-circle” problem: values of J_I and J_B as function of the iteration number for different values of N_b on a 200×200 grid, $\alpha = 0.2$.

Fig. 7 and Table 1 show the results using this stopping criterion, along with performed logarithmic least-squares fits. The number of iterations required increases sublinearly with N . We see that J_I has approximately third-order convergence and J_B second-order convergence, when using the stopping criterion. In general, we will not use this stopping criterion in the remainder of this paper, only when indicated.

Fig. 8 shows the influence of N_b for a 200×200 grid. Choosing a larger N_b decreases J_I and J_B , but this effect becomes smaller for large enough N_b .

Fig. 9a shows the calculation times of the minimization procedures for $\mathbf{P}$, $\mathbf{b}$, $\mathbf{m}$, and the computation of u_1 as a function of $N = N_1 = N_2$. As expected, we conclude that the calculation time for the minimization procedure for $\mathbf{P}$ is quadratic in N , and thus linear in the number of grid points. The minimization procedure for $\mathbf{b}$ is linear in N . The calculation of $\mathbf{m}$ and u_1 should be at least linear in the number of grid points and thus quadratic in N . As shown in Fig. 9b, the minimization procedure for $\mathbf{b}$ increases approximately linearly with increasing N_b . The slopes of performed logarithmic least-squares fits are also displayed. The total calculation time for one iteration is approximately proportional to N^2 , and with the number of iterations growing sublinearly in N , see Table 1, the total calculation time scales roughly with N^3 .

The choice of the value of α is examined in Fig. 10. We plot the functional $J = (1 - \alpha) J_I + \alpha J_B$, introduced in (28). We use the stopping criterion in (54) and see that the criterion and a lower value of J is reached sooner for small but not too small values of α , i.e., $\alpha = 0.1$ and $\alpha = 0.2$. This result is independent of grid size. For large values of α , i.e., $\alpha = 0.8$, and small grid size 50×50 we see that the value of J increases before reaching the stopping criterion, which happens during the minimization procedure for $\mathbf{P}$. When the mapping $\mathbf{m}$ has changed, the set $\mathcal{P}(\mathbf{m})$ has changed over which we minimize in (29b).

4.4. Picture on a projection screen in the far field

As a final example we challenge our algorithm to compute a freeform reflector surface that converts the light from a point source into a far-field target intensity distribution corresponding to a picture.

For an incoming cone-shaped beam the source domain $\mathcal{X}$ is circular. We perform a change of coordinates for the stereographic variables $\mathbf{x} \in \mathcal{X}$ on the source domain to polar stereographic coordinates $\boldsymbol{\omega} = (\rho, \zeta)$ with the transformation where $\rho \geq 0$ is the radial coordinate and $0 \leq \zeta < 2\pi$ the azimuth (angle with respect to positive x_1 -axis). We maintain Cartesian stereographic coordinates for the target domain. The derivation of the corresponding generalized Monge–Ampère equation in polar stereographic coordinates and modifications to the least-squares algorithm are presented in [31].

We consider the source domain $\mathcal{X} = \{(\rho, \zeta) \in \mathbb{R}^2 \mid 0 \leq \rho \leq 0.2, 0 \leq \zeta < 2\pi\}$ using polar stereographic coordinates (ρ, ζ) for the source domain and a uniform light distribution $\tilde{f}(\rho, \zeta) = 1$ and consequently $I(\mathcal{X}) = \frac{2}{13}\pi$. The opening angle of the narrow cone-shaped bundle is approximately 1.1 rad = 63° . The reflected rays are projected on a screen P in the far field, parallel to the xz -plane. The required illumination $L(\xi, \eta)$ [lm/m^2], with (ξ, η) the Cartesian coordinates on the projection

Fig. 9. “Square-to-circle” problem: average calculation time per iteration (50 iterations total) as function of N and N_b . We calculate a c-concave solution u_1 and use parameter value $\alpha = 0.2$. The dashed lines are least-squares fits. The laptop used for the calculations has an Intel Core i7-7700HQ CPU 2.80 GHz with 32.0 GB of RAM.

screen, is derived from the grey scale values of a picture of a koala, see Fig. 11. The target distribution $\tilde{g}(\mathbf{y})$ is a deformation of the illuminance $L(\xi, \eta)$; the conversion from $L(\xi, \eta)$ to $\tilde{g}(\mathbf{y})$ is explained in detail in [7,32].

The grey scale values of the picture prescribe the illuminance. However, the conversion from the coloured image to grey scale values creates black regions in the target distribution for which $\tilde{g}(\mathbf{y}) = 0$. To avoid division by 0 in the right hand side of equation (25a), we increase values of $\tilde{g}(\mathbf{y})$ which are below a threshold of 5% of its maximum value to this threshold. We determine $I(\mathcal{V})$ by dividing the region on the projection screen into quadrants with the pixels of the picture as corners and approximating the integral of $L(\xi, \eta)$ using the 2D composite trapezoidal rule.

We use the least-squares algorithm to compute the optical map $\mathbf{m}$ and the reflector surface. We discretize the source domain using a 1000×1000 grid and use $\alpha = 0.5$ to give the domain and boundary equal weights in the minimization procedure. We use the initial guess $\mathbf{m}^0$ given in equation (43) in [31] to compute a c-convex u_1 . The optical map, plotted using a coarsened version of the source grid, the reflector surface, and convergence results are shown in Fig. 12. The error J_I in the interior converges to a larger value than the error J_B in the boundary, as the finite grid is unable to capture the finest details of the koala.

Subsequently, we validated the resulting reflector image using ray tracing. We traced approximately 3000×3000 uniformly distributed rays from source to far field. The resulting target illuminance $L(\xi, \eta)$ is plotted in Fig. 12. The ray trace image closely resembles the original picture, showing details such as the hairs on the koala’s coat.

5. Concluding remarks

In this paper, we presented a method to compute the shape of a single freeform reflector which transforms the light distribution of a point source to a desired far-field distribution. First, we showed that the optical mapping can be derived using either the geometrical-optics law of reflection or the logarithmic cost function in the corresponding optimal transport problem. We subsequently combined the optical mapping with energy conservation to derive the generalized Monge–Ampère equation. We developed a generalized least-squares method for the numerical solution of this nonlinear second order elliptic partial differential equation. The method first computes the optical map in an iterative procedure, and proceeds to calculate the location of the optical surface from the converged mapping. We tested our algorithm for a couple of examples. For an exact solution, such as a flat reflector surface, we checked the accuracy of the algorithm. We also constructed a reflector converting the light of a point source into a picture on a screen in the far field. To the best of our knowledge, there are no known algorithms for the point source which are capable of calculating reflector surfaces that produce this level of detail in the far field.

A natural extension is to apply this algorithm to the generalized Monge–Ampère equation corresponding to a lens surface. In fact, the formulation of the Monge–Ampère equation using the cost function and corresponding solution method using c-convexity theory allows for an extension of the method to any optical system with a continuously differentiable cost function and a relation of the form $u_1(\mathbf{x}) + u_2(\mathbf{y}) = c(\mathbf{x}, \mathbf{y})$, where u_1 and u_2 are some geometrical parameters describing the optical element(s). We aim to apply the method to real optical design problems, using a different variety of source and target distributions, e.g., target intensity profiles for street lighting purposes.

Fig. 10. “Square-to-circle” problem; values of $J = (1 - \alpha) J_I + \alpha J_B$ as function of the iteration number for several grid sizes and different values of α with $N_b = 1000$. We use the stopping criterion in (54).

Fig. 11. “Square-to-koala” problem: the original image (left) and target distribution $\tilde{g}(\mathbf{y})$ and target boundary (right).

Fig. 12. “Square-to-koala” problem: the optical mapping (upper left), reflector surface (upper right), J_I and J_B (lower left), and ray traced image (lower right).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

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

Appendix A. The geometrical-optics derivation of the optimal map

We derive the mapping $\mathbf{y} = \mathbf{m}(\mathbf{x})$ in stereographic coordinates using the expression for the reflected ray $\hat{\mathbf{t}}$ in (2).

First, let $u_1(\mathbf{x}) = \log(u(\mathbf{x})/(1 + |\mathbf{x}|^2))$. The division within the logarithm is motivated in Section 2.1.2; by defining (8) we arrive at a relation of the form $u_1(\mathbf{x}) + u_2(\mathbf{y}) = c(\mathbf{x}, \mathbf{y})$ with $c(\mathbf{x}, \mathbf{y})$ as compactly formulated as possible). We rewrite the gradient of u restricted to the surface $r = \text{constant}$ in (1b) to

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

eliminating the term u in the denominator, cf. (1b). We assume $u_1 = u_1(\mathbf{x})$ and apply the chain rule to the partial derivatives of u_1 , i.e.,

$$\frac{\partial u_1}{\partial \phi} = \frac{\partial u_1}{\partial x_1} \frac{\partial x_1}{\partial \phi} + \frac{\partial u_1}{\partial x_2} \frac{\partial x_2}{\partial \phi}, \quad \frac{\partial u_1}{\partial \theta} = \frac{\partial u_1}{\partial x_1} \frac{\partial x_1}{\partial \theta} + \frac{\partial u_1}{\partial x_2} \frac{\partial x_2}{\partial \theta}. \quad (\text{A.2})$$

Using the definition of the stereographic coordinates of the source in equation (3) we can rewrite the partial derivatives of $\mathbf{x}$ occurring in (A.1) and (A.2) in $\hat{\mathbf{s}}$ -coordinates to

$$\frac{\partial x_1}{\partial \phi} = \frac{\cos(\theta)}{1 + \cos(\phi)} = \frac{s_1}{(1 + s_3) \sqrt{s_1^2 + s_2^2}}, \quad \frac{\partial x_1}{\partial \theta} = -\frac{\sin(\phi) \sin(\theta)}{1 + \cos(\phi)} = -\frac{s_2}{1 + s_3}, \quad (\text{A.3a})$$

$$\frac{\partial x_2}{\partial \phi} = \frac{\sin(\theta)}{1 + \cos(\phi)} = \frac{s_2}{(1 + s_3) \sqrt{s_1^2 + s_2^2}}, \quad \frac{\partial x_2}{\partial \theta} = \frac{\sin(\phi) \cos(\theta)}{1 + \cos(\phi)} = \frac{s_1}{1 + s_3}. \quad (\text{A.3b})$$

Second, we can express the basis vectors $\hat{\mathbf{e}}_\phi$ and $\hat{\mathbf{e}}_\theta$ in $\mathbf{x}$ -coordinates ($\hat{\mathbf{e}}_r = \hat{\mathbf{s}}$, see (5)) as follows:

$$\begin{aligned} \hat{\mathbf{e}}_\phi &= \begin{pmatrix} \cos(\phi) \cos(\theta) \\ \cos(\phi) \sin(\theta) \\ -\sin(\phi) \end{pmatrix} = \frac{1}{\sqrt{s_1^2 + s_2^2}} \begin{pmatrix} s_1 s_3 \\ s_2 s_3 \\ -s_1^2 - s_2^2 \end{pmatrix} \\ &= \frac{1}{|\mathbf{x}| (1 + |\mathbf{x}|^2)} \begin{pmatrix} x_1 (1 - |\mathbf{x}|^2) \\ x_2 (1 - |\mathbf{x}|^2) \\ -2 |\mathbf{x}|^2 \end{pmatrix}, \\ \hat{\mathbf{e}}_\theta &= \begin{pmatrix} -\sin(\theta) \\ \cos(\theta) \\ 0 \end{pmatrix} = \frac{1}{\sqrt{s_1^2 + s_2^2}} \begin{pmatrix} -s_2 \\ s_1 \\ 0 \end{pmatrix} = \frac{1}{|\mathbf{x}|} \begin{pmatrix} -x_2 \\ x_1 \\ 0 \end{pmatrix}. \end{aligned} \quad (\text{A.3c})$$

Third, substituting $\sin(\phi) = \sqrt{s_1^2 + s_2^2} = 2 |\mathbf{x}| / (1 + |\mathbf{x}|^2)$, (A.2) and (A.3) into (A.1) and using the inverse projection in (5), we get an expression for $\nabla_r u$ in terms of $\mathbf{x}$ and the partial derivatives of u_1 with respect to $\mathbf{x}$, as

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

where

$$\nabla u_1 = \begin{pmatrix} \frac{\partial u_1}{\partial x_1} \\ \frac{\partial u_1}{\partial x_2} \end{pmatrix}. \quad (\text{A.4b})$$

Fourth, substituting this expression into the reflected ray $\hat{\mathbf{t}}$ in (2) and applying the inverse stereographic projection in (5) to $\hat{\mathbf{e}}_r = \hat{\mathbf{s}}$, we find the outgoing ray $\hat{\mathbf{t}}(\mathbf{x})$ expressed in $\mathbf{x}$ -coordinates as

$$\hat{\mathbf{t}}(\mathbf{x}) = \frac{1}{\mathbf{D}(\mathbf{x})} \begin{pmatrix} 4 \frac{\partial u_1}{\partial x_1} + 2 |\nabla u_1|^2 x_1 \\ 4 \frac{\partial u_1}{\partial x_2} + 2 |\nabla u_1|^2 x_2 \\ 2 |\nabla u_1|^2 - \mathbf{D}(\mathbf{x}) \end{pmatrix}, \quad (\text{A.5a})$$

where

$$\mathbf{D}(\mathbf{x}) = 4 + 4 \mathbf{x} \cdot \nabla u_1 + |\nabla u_1|^2 (1 + |\mathbf{x}|^2). \quad (\text{A.5b})$$

Finally, transforming $\hat{\mathbf{t}}(\mathbf{x})$ to $\mathbf{y}$ using the definition of the stereographic coordinates of the target in (4), we arrive at $\mathbf{y} = \mathbf{m}(\mathbf{x})$

$$\mathbf{y} = \frac{2 \nabla u_1 + \mathbf{x} |\nabla u_1|^2}{4 + 4 \mathbf{x} \cdot \nabla u_1 + (|\mathbf{x}| |\nabla u_1|)^2}. \quad (\text{A.6})$$

Appendix B. The transport boundary condition and edge-ray principle

In this section, we show that the transport boundary condition in (25b) follows from the implicit boundary condition $\mathbf{m}(\mathcal{X}) = \mathcal{Y}$, stating that all the light from the source $\mathcal{X}$ must be transferred to the target domain $\mathcal{Y}$. The equivalence of the boundary conditions follows from the edge-ray principle [29] and convexity of the optical surface.

As in section 2.1, we define our source domain $\mathcal{X}$ as the closed support of $\tilde{f}(\mathbf{x}) = f(\phi(\mathbf{x}), \theta(\mathbf{x}))$, and our target domain $\mathcal{Y}$ as the image under the mapping $\mathbf{m}$, i.e., $\mathcal{Y} = \mathbf{m}(\mathcal{X})$. We refer to $\mathbf{m} : \mathcal{X} \rightarrow \mathcal{Y}$ as the optical map $\mathbf{y} = \mathbf{m}(\mathbf{x})$ from the source set of stereographic coordinates $\mathcal{X}$ to the target set of stereographic coordinates $\mathcal{Y}$. We use the implicit boundary condition $\mathbf{m}(\mathcal{X}) = \mathcal{Y}$.

The equivalence of the two boundary conditions $\mathbf{m}(\mathcal{X}) = \mathcal{Y}$ and $\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}$ follows from a basic principle of topology, which is known as the edge-ray principle when applied to optics [29]. The closure of a set $\mathcal{A}$ is denoted by a bar and defined as $\overline{\mathcal{A}} = \text{int}(\mathcal{A}) \cup \partial\mathcal{A}$, where $\text{int}(\mathcal{A})$ denotes the union of all open subsets of $\mathcal{A}$. Note that $\mathcal{X}$ is also defined as the closure of the subset of all points where $\tilde{f}(\mathbf{x})$ is nonzero, but we omit the bar notation.

Theorem 1. Let $\mathcal{X}$ and $\mathcal{Y}$ be topological spaces. Let $\mathbf{m} : \mathcal{X} \rightarrow \mathcal{Y}$ be a homeomorphism, i.e., $\mathbf{m}$ is a bijection that is continuous and open, so $\mathbf{m}^{-1}$ is also continuous. Then, for any subset $\mathcal{A} \subseteq \mathcal{X}$, we have $\mathbf{m}(\partial\mathcal{A}) = \partial\mathbf{m}(\mathcal{A})$. In particular, $\mathbf{m}(\partial\mathcal{X}) = \partial\mathbf{m}(\mathcal{X}) = \partial\mathcal{Y}$.

Proof. Let $\mathcal{A} \subseteq \mathcal{X}$ and let $\mathbf{y} \in \mathbf{m}(\partial\mathcal{A})$. Then $\mathbf{x} = \mathbf{m}^{-1}(\mathbf{y}) \in \partial\mathcal{A}$. Hence, $\mathbf{x}$ is a boundary point of $\mathcal{A}$ and so $\mathbf{x} \in \overline{\mathcal{A}} \setminus \text{int}(\mathcal{A})$. This implies that $\mathbf{y} \in \mathbf{m}(\overline{\mathcal{A}} \setminus \text{int}(\mathcal{A}))$, because $\mathbf{m}$ is an injective mapping. Furthermore, we have that $\mathbf{m}$ is a homeomorphism, implying that $\mathbf{m}(\overline{\mathcal{A}}) = \overline{\mathbf{m}(\mathcal{A})}$ and $\mathbf{m}(\text{int}(\mathcal{A})) = \text{int}(\mathbf{m}(\mathcal{A}))$. Therefore, $\mathbf{y} \in \overline{\mathbf{m}(\mathcal{A})} \setminus \text{int}(\mathbf{m}(\mathcal{A})) = \partial\mathbf{m}(\mathcal{A})$, and we have $\mathbf{m}(\partial\mathcal{A}) \subseteq \partial\mathbf{m}(\mathcal{A})$. Reversing our line of thought above we can also show $\partial\mathbf{m}(\mathcal{A}) \subseteq \mathbf{m}(\partial\mathcal{A})$. We conclude that $\mathbf{m}(\partial\mathcal{A}) = \partial\mathbf{m}(\mathcal{A})$. $\square$

For the equivalence of the two boundary conditions it is essential that the optical mapping $\mathbf{m}$ is a homeomorphism. In the remainder of this section, we write the mapping $\mathbf{m}$ as $\mathbf{m} = (\gamma_+ \circ \hat{\mathbf{t}} \circ \gamma_-^{-1})(\mathbf{x})$, where we denote $\gamma_{\pm}$ as the stereographic maps (- for south pole in equation (3), + for north pole in equation (4)) and $\hat{\mathbf{t}}$ the vectorial law of reflection in equation (2). We prove that $\mathbf{m}$ is a homeomorphism by showing in Lemma 1 that $\gamma_{\pm}$ is a bijection and in Lemma 2 that $\hat{\mathbf{t}}$ is a bijection. Subsequently, we can conclude that $\mathbf{m} = (\gamma_+ \circ \hat{\mathbf{t}} \circ \gamma_-^{-1})(\mathbf{x})$ is a bijection using the fact that a composition of bijective functions is bijective.

The stereographic projections in (3) and (4) can be rewritten as

$$\gamma_{\pm}(\mathbf{w}) = \frac{1}{1 - \mathbf{w} \cdot \mathbf{p}_{\pm}} \mathbf{w}^*, \quad (\text{B.1})$$

where the vector $\hat{\mathbf{p}}_{\pm} = (0, 0, \pm 1)^T$ denotes the pole, and we consider $\mathbf{w} \in S^2 \setminus \hat{\mathbf{p}}_{\pm} = \{(w_1, w_2, w_3)^T \in \mathbb{R}^3 \mid |\mathbf{w}|^2 = 1\} \setminus \hat{\mathbf{p}}_{\pm}$ and $\mathbf{w}^* \in \mathbb{R}^2 = \{(w_1, w_2, 0)^T \in \mathbb{R}^3 \mid (w_1, w_2)^T \in \mathbb{R}^2\}$.

Lemma 1. $\gamma_{\pm}$ is a bijection from the unit sphere $S^2 \setminus \hat{\mathbf{p}}_{\pm}$ to $\mathbb{R}^2$, i.e., from S^2 without the north/south pole to $\mathbb{R}^2$.

Proof. A function is bijective if and only if it has an inverse. It is sufficient to show that $\gamma_{\pm}(\mathbf{w})$ has an inverse. The inverse is given in equation (5) and can be written as

$$\gamma_{\pm}^{-1}(\mathbf{v}) = \hat{\mathbf{p}}_{\pm} + \frac{2(\mathbf{v} - \hat{\mathbf{p}}_{\pm})}{|\mathbf{v} - \hat{\mathbf{p}}_{\pm}|^2}, \quad (\text{B.2})$$

where $\mathbf{v} \in \mathbb{R}^2 = \{(v_1, v_2, 0)^T \in \mathbb{R}^3 \mid (v_1, v_2)^T \in \mathbb{R}^2\}$ and the vector $\hat{\mathbf{p}}_{\pm} = (0, 0, \pm 1)^T$ is a unit vector perpendicular to the plane of $\mathbf{v}$. Note that $|\mathbf{v} - \hat{\mathbf{p}}_{\pm}|^2$ is nowhere 0 in $\mathbb{R}^3$ since the third component of $\mathbf{v} - \hat{\mathbf{p}}_{\pm}$ is never 0, and that we retrieve a vector $\mathbf{w} \neq \hat{\mathbf{p}}_{\pm}$ for $\mathbf{v}$ finite. In fact, $\lim_{|\mathbf{v}| \rightarrow \infty} \gamma_{\pm}^{-1} = \hat{\mathbf{p}}_{\pm}$. The length of $\mathbf{w}$ indeed becomes

$$\begin{aligned} |\gamma_{\pm}^{-1}(\mathbf{v})| &= \sqrt{\left(\frac{2v_1^2}{1+v_1^2+v_2^2}\right)^2 + \left(\frac{2v_2^2}{1+v_1^2+v_2^2}\right)^2 + \left(\frac{\pm(-1+v_1^2+v_2^2)}{1+v_1^2+v_2^2}\right)^2} \\ &= \sqrt{\frac{1+2v_1^2+v_1^4+2v_2^2+2v_1^2v_2^2+v_2^4}{(1+v_1^2+v_2^2)^2}} \\ &= \sqrt{\frac{(1+v_1^2+v_2^2)^2}{(1+v_1^2+v_2^2)^2}} = 1. \quad \square \end{aligned} \quad (\text{B.3})$$

Lemma 2. The vectorial law of reflection $\hat{\mathbf{t}} = \hat{\mathbf{s}} - 2(\hat{\mathbf{s}} \cdot \hat{\mathbf{n}})\hat{\mathbf{n}}$ with $\hat{\mathbf{s}} = \hat{\mathbf{e}}_r$ is a bijection for strictly convex optical surfaces. In our case, if the reflector surface $z(x, y) = \sqrt{u(\phi, \theta)^2 - (x^2 + y^2)}$ is convex with $(x, y, z)^T$ denoting the Cartesian coordinate vector, then the vectorial law of reflection is a bijection.

Fig. B.13. As $\hat{s}$ moves counterclockwise from $\hat{s}_1$ to $\hat{s}_2$, the normal $\hat{n}$ to the surface must also move counterclockwise from $\hat{n}_1$ to $\hat{n}_2$ if $\hat{t}_1 = \hat{t}_2$.

Proof. The law of reflection is surjective by definition since we require $\mathcal{Y}$ to be the image of $\mathcal{X}$ under the mapping $\mathbf{m}$. Using bijectivity of the stereographic projections and $\mathcal{Y} = \mathbf{m}(\mathcal{X}) = (\gamma_+ \circ \hat{\mathbf{t}} \circ \gamma_-^{-1})(\mathcal{X})$, we require the inverse stereographic projection of $\mathcal{Y}$, written as $\gamma_+^{-1}(\mathcal{Y})$, to be the image of the inverse stereographic projection of $\mathcal{X}$, written as $\gamma_-^{-1}(\mathcal{X})$, under the law of reflection $\hat{\mathbf{t}}$.

To show injectivity, we assume $\hat{s}_1 \neq \hat{s}_2$ and aim to prove $\hat{\mathbf{t}}(\hat{s}_1) \neq \hat{\mathbf{t}}(\hat{s}_2)$. We write

$$\hat{\mathbf{t}}_1 = \hat{s}_1 - 2(\hat{s}_1 \cdot \hat{\mathbf{n}}_1)\hat{\mathbf{n}}_1, \quad (\text{B.4a})$$

$$\hat{\mathbf{t}}_2 = \hat{s}_2 - 2(\hat{s}_2 \cdot \hat{\mathbf{n}}_2)\hat{\mathbf{n}}_2, \quad (\text{B.4b})$$

with $\hat{\mathbf{t}}_1 = \hat{\mathbf{t}}(\hat{s}_1)$, $\hat{\mathbf{t}}_2 = \hat{\mathbf{t}}(\hat{s}_2)$, $\hat{\mathbf{n}}_1 = \hat{\mathbf{n}}(\hat{s}_1)$ and $\hat{\mathbf{n}}_2 = \hat{\mathbf{n}}(\hat{s}_2)$.

The vectorial law of reflection states that the incident ray $\hat{s}$, the reflected ray $\hat{t}$ and the normal $\hat{n}$ to the surface all lie in the same plane. Hence, if $\text{span}(\hat{s}_1, \hat{\mathbf{n}}_1) \neq \text{span}(\hat{s}_2, \hat{\mathbf{n}}_2)$ then it immediately follows that $\hat{\mathbf{t}}_1 \neq \hat{\mathbf{t}}_2$. Therefore, it remains to show that if $\text{span}(\hat{s}_1, \hat{\mathbf{n}}_1) = \text{span}(\hat{s}_2, \hat{\mathbf{n}}_2)$, i.e., $\hat{s}_1$ and $\hat{s}_2$ lie in the same plane, it follows that $\hat{\mathbf{t}}_1 \neq \hat{\mathbf{t}}_2$. In the remainder of this proof, we will give a geometric argument to show that this holds for a convex optical surface.

Fig. B.13 shows the plane spanned by $\hat{s}_1$ and $\hat{s}_2$. We draw the vectors $\hat{s}_1$ and $\hat{s}_2$ as vectors on the unit sphere. The vectorial law of reflection states that the vector $\hat{\mathbf{t}} - \hat{s} = -2(\hat{s} \cdot \hat{\mathbf{n}})\hat{\mathbf{n}}$ is a multiple of the unit normal vector $\hat{\mathbf{n}}$. If we assume that $\hat{\mathbf{t}}_1 = \hat{\mathbf{t}}_2$ we see that as $\hat{s}$ moves counterclockwise from $\hat{s}_1$ to $\hat{s}_2$, the normal $\hat{\mathbf{n}}$ to the surface must also move counterclockwise from $\hat{\mathbf{n}}_1$ to $\hat{\mathbf{n}}_2$. Here $\hat{\mathbf{n}}_1$ and $\hat{\mathbf{n}}_2$ are taken to be directed towards the point source $\mathcal{O}$.

In Fig. 14b we see a parabolic reflector with focal point at $\mathcal{O}$ and, hence, multiple parallel outgoing rays with $\hat{\mathbf{t}}_1 = \hat{\mathbf{t}}_2$ while $\hat{s}_1 \neq \hat{s}_2$. The optical surface is strictly concave and we can verify that as $\hat{s}$ moves counterclockwise from $\hat{s}_1$ to $\hat{s}_2$, the unit normal $\hat{\mathbf{n}}$ moves in the same direction.

If the optical surface is convex, as for instance shown in Fig. 14a, this situation can not occur. As $\hat{s}$ moves counterclockwise from $\hat{s}_1$ to $\hat{s}_2$, the downward unit normal $\hat{\mathbf{n}}$ moves in the opposite direction (or keeps pointing in the same direction in case of a flat surface). This contradicts with the assumption that $\hat{\mathbf{t}}_1 = \hat{\mathbf{t}}_2$ and hence, $\hat{\mathbf{t}}_1 \neq \hat{\mathbf{t}}_2$. $\square$

In summary, we have shown that the stereographic projections $\gamma_{\pm}$ are bijective functions in Lemma 1. The law of reflection is bijective for convex optical surfaces as proven in Lemma 2. Combining Lemma 1 and 2, $\mathbf{m} = (\gamma_+ \circ \hat{\mathbf{t}} \circ \gamma_-^{-1})(\mathbf{x})$ is a homeomorphism under the assumption of a convex optical surface by the composition of bijective functions. Subsequently, the equivalence of the boundary conditions $\mathbf{m}(\mathcal{X}) = \mathcal{Y}$ and $\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}$ follows from Theorem 1.

The equivalence proof presented in this section holds for a convex optical surface. During the iterative procedure of the numerical algorithm we require that the matrix $D\mathbf{m}$ is nonsingular, and we have $\forall \mathbf{x} \in \mathcal{X}$

$$\det(D\mathbf{m}(\mathbf{x})) = \frac{\tilde{f}(\mathbf{x})(1 + |\mathbf{m}(\mathbf{x})|^2)^2}{\tilde{g}(\mathbf{m}(\mathbf{x}))(1 + \mathbf{x}^2)^2} > 0, \quad (\text{B.5})$$

cf. equation (23), since we take $\mathcal{X}$ as the closed support of $\tilde{f}(\mathbf{x})$ and $\mathcal{Y} = \mathbf{m}(\mathcal{X})$, i.e., we require that all energy from an arbitrary subset of the source domain is mapped to a subset of the target domain and never collapses to a single point. It is also continuous since we take $f: S^2 \rightarrow [0, \infty)$ and $g: S^2 \rightarrow [0, \infty)$ to be continuous intensity functions and $\mathbf{m} \in C^2(\mathcal{X})^2$. Starting from an injective initial mapping in equation (31) or (32) we locally enforce a one-to-one correspondence by requiring $\det(D\mathbf{m}) > 0$ with $F(\mathbf{x}, \mathbf{m}(\mathbf{x})) > 0$ in equation (25a). In the numerical examples we have considered so far, this local enforcement leads to a globally injective $\mathbf{m}$. Moreover, the implicit function theorem requires that $\det(\nabla_{\mathbf{y}} \nabla_{\mathbf{x}} c(\mathbf{x}, \mathbf{y})) \neq 0$ for the mapping $\mathbf{y}$ defined as $\mathbf{m}(\mathbf{x}) = (\nabla_{\mathbf{x}} c(\mathbf{x}, \cdot))^{-1} \circ \nabla u_1(\mathbf{x})$ to exist locally, cf. equation (12) in the article, where $(\nabla_{\mathbf{x}} c)^{-1}$ is the local inverse of $\mathbf{x} \mapsto \nabla_{\mathbf{x}} c(\mathbf{x}, \mathbf{y})$. In other words, for every point $\mathbf{y}$ there exists a small neighbourhood for which $\nabla_{\mathbf{y}} \nabla_{\mathbf{x}} c$ is invertible and $\mathbf{x} \mapsto \nabla_{\mathbf{x}} c(\mathbf{x}, \mathbf{y})$ has a local inverse $\mathbf{y} = (\nabla_{\mathbf{x}} c(\mathbf{x}, \cdot))^{-1} \circ \nabla u_1(\mathbf{x})$. We know that

$$\det(\nabla_{\mathbf{y}} \nabla_{\mathbf{x}} c(\mathbf{x}, \mathbf{y})) = \det(\mathbf{C}) = \frac{4}{N(\mathbf{x}, \mathbf{y})^2} \neq 0, \quad (\text{B.6})$$

Fig. 14. (a) A convex parabolic reflector surface gives an injective $\hat{\mathbf{t}}$ but (b) a strictly concave parabolic reflector surface may not.

with $\mathbf{N}(\mathbf{x}, \mathbf{y})$ from (9b). This inequality holds everywhere since we can write $\mathbf{N}(\mathbf{x}, \mathbf{y})$ as

$$\mathbf{N}(\mathbf{x}(\hat{\mathbf{s}}), \mathbf{y}(\hat{\mathbf{t}})) = \frac{1}{2} (1 + |\mathbf{x}(\hat{\mathbf{s}})|^2) (1 + |\mathbf{y}(\hat{\mathbf{t}})|^2) (1 - \hat{\mathbf{s}} \cdot \hat{\mathbf{t}}), \quad (\text{B.7})$$

cf. (7), (8) and (9), and we assume that $\hat{\mathbf{s}} \cdot \hat{\mathbf{t}} \neq 1$, which means that the reflector changes the direction of the light rays.

Upon convergence of the numerical procedure in section 3.4, we do not enforce convexity/concavity of the resulting optical surface. The cost function in equation (7) and (9) is convex and we can prove that taking a c-convex (maximum) pair results in a convex u_1 [28, p.61]. This gives a convex $u(\mathbf{x}) = e^{u_1(\mathbf{x})} (1 + |\mathbf{x}|^2)$, but we can not conclude a priori that z in equation (49) is convex. Therefore, we can not conclude a priori that the optimal mapping $\mathbf{m}$ for a particular source domain $\mathcal{X}$ and target domain $\mathcal{Y}$ should be a homeomorphism. However, in order to locally enforce energy conservation and ensure the regularity of $\mathbf{m}$, we assume that the optical mapping $\mathbf{m}$ is globally injective and always use the transport boundary condition $\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}$.

References

- [1] A.S. Glassner, *An Introduction to Ray Tracing*, Elsevier, 1989.
- [2] C. Filosa, J.H.M. ten Thije Boonkkamp, W.L. IJzerman, Ray tracing method in phase space for two-dimensional optical systems, *Appl. Opt.* 55 (2016) 3599–3606.
- [3] C. Filosa, J.H.M. ten Thije Boonkkamp, W.L. IJzerman, Phase space ray tracing for a two-dimensional parabolic reflector, *Math. Stat.* 5 (2017) 135–142.
- [4] C. Villani, *Optimal Transport: Old and New*, vol. 338, Springer Science & Business Media, 2009.
- [5] C. Villani, *Topics in Optimal Transportation*, vol. 58, American Mathematical Soc., 2003.
- [6] 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.
- [7] C.R. Prins, *Inverse Methods for Illumination Optics*, Ph.D. thesis, Eindhoven University of Technology, Eindhoven, 2014.
- [8] N.K. Yadav, L.B. Romijn, J.H.M. ten Thije Boonkkamp, W.L. IJzerman, A least-squares method for the design of two-reflector optical systems, *J. Phys. Photonics* 1 (2019) 034001.
- [9] B.D. Froese, A numerical method for the elliptic Monge–Ampère equation with transport boundary conditions, *SIAM J. Sci. Comput.* 34 (2012) A1432–A1459.
- [10] 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 (2015) B937–B961.
- [11] V. Oliker, E. Newman, The energy conservation equation in the reflector mapping problem, *Appl. Math. Lett.* 6 (1993) 91–95.
- [12] X.-J. Wang, On the design of a reflector antenna, *Inverse Probl.* 12 (1996) 351.
- [13] X.-J. Wang, On the design of a reflector antenna II, *Calc. Var. Partial Differ. Equ.* 20 (2004) 329–341.
- [14] W. Gangbo, V.I. Oliker, Existence of optimal maps in the reflector-type problems, *ESAIM Control Optim. Calc. Var.* 13 (2007) 93–106.
- [15] V. Oliker, Mathematical aspects of design of beam shaping surfaces in geometrical optics, in: *Trends in Nonlinear Analysis*, Springer, 2003, pp. 193–224.
- [16] S.A. Kochengin, V.I. Oliker, Computational algorithms for constructing reflectors, *Comput. Vis. Sci.* 6 (2003) 15–21.
- [17] H. Ries, J. Muschawek, Tailored freeform optical surfaces, *J. Opt. Soc. Am. A* 19 (2002) 590–595.
- [18] T. Glimm, V.I. Oliker, Optical design of single reflector systems and the Monge–Kantorovich mass transfer problem, *J. Math. Sci.* 117 (2003) 4096–4108.
- [19] F.R. Fournier, W.J. Cassarly, J.P. Rolland, Fast freeform reflector generation using source-target maps, *Opt. Express* 18 (2010) 5295–5304.
- [20] C. Canavesi, W.J. Cassarly, J.P. Rolland, Target flux estimation by calculating intersections between neighboring conic reflector patches, *Opt. Lett.* 38 (2013) 5012–5015.
- [21] R. Wu, L. Xu, P. Liu, Y. Zhang, Z. Zheng, H. Li, X. Liu, Freeform illumination design: a nonlinear boundary problem for the elliptic Monge–Ampère equation, *Opt. Lett.* 38 (2013) 229–231.
- [22] K. Brix, Y. Hafizogullari, A. Platen, Designing illumination lenses and mirrors by the numerical solution of Monge–Ampère equations, *J. Opt. Soc. Am. A* 32 (2015) 2227–2236.
- [23] 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.
- [24] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, G. Peyré, Iterative Bregman projections for regularized transportation problems, *SIAM J. Sci. Comput.* 37 (2015) A1111–A1138.
- [25] N.K. Yadav, J.H.M. ten Thije Boonkkamp, W.L. IJzerman, A Monge–Ampère problem with non-quadratic cost function to compute freeform lens surfaces, *J. Sci. Comput.* 80 (2019) 475–499.

- [26] V.I. Olier, Freeform Optical Systems With Prescribed Irradiance Properties in Near-Field, International Optical Design Conference 2006, vol. 6342, International Society for Optics and Photonics, 2006, p. 634211.
- [27] R.K. Luneburg, Mathematical Theory of Optics, University of California Press, 1964.
- [28] N.K. Yadav, Monge–Ampère Problems with Non-Quadratic Cost Function: Application to Freeform Optics, Ph.D. thesis, Eindhoven University of Technology, Eindhoven, 2018.
- [29] H. Ries, A. Rabl, Edge-ray principle of nonimaging optics, *J. Opt. Soc. Am. A* 11 (1994) 2627–2632.
- [30] R. Courant, D. Hilbert, Wiley classic edition, *Methods Math. Phys.* 1 (1989).
- [31] L.B. Romijn, J.H.M. ten Thije Boonkkamp, W.L. IJzerman, Freeform lens design for a point source and far-field target, *J. Opt. Soc. Am. A* 36 (2019) 1926–1939.
- [32] M. Born, E. Wolf, *Principles of Optics*, 7 ed., Cambridge University Press, 1999.