Journal of Scientific Computing (2019) 80:475–499
https://doi.org/10.1007/s10915-019-00948-9

# A Monge–Ampère Problem with Non-quadratic Cost Function to Compute Freeform Lens Surfaces

**N. K. Yadav[1] · J. H. M. ten Thije Boonkkamp[1] · W. L. IJzerman[1,2]**

Received: 28 July 2018 / Revised: 22 January 2019 / Accepted: 19 March 2019 /
Published online: 27 March 2019


**Abstract**
In this article, we present a least-squares method to compute freeform surfaces of a lens with parallel incoming and outgoing light rays, which is a transport problem corresponding to a *non-quadratic* cost function. The lens can transfer a given emittance of the source into a desired illuminance at the target. The freeform lens design problem can be formulated as a Monge–Ampère type differential equation with transport boundary condition, expressing conservation of energy combined with the law of refraction. Our least-squares algorithm is capable to handle a non-quadratic cost function, and provides two solutions corresponding to either convex or concave lens surfaces.

**Keywords** Monge–Ampère equation · Transport boundary conditions · Non-quadratic cost function · Least-squares method · Freeform lens surfaces · Optical design · Inverse problem

**Mathematics Subject Classification** 35A15 · 35J15 · 35J20 · 35J96 · 4900 · 65N08 · 65N21 · 78A05

## 1 Introduction

The optical design problem involving freeform surfaces is a challenging problem, even for a single mirror/lens surface which transfers a given intensity/emittance distribution of the source into a desired intensity/illuminance distribution at the target [1–3]. More specifically, the freeform design problem is an *inverse problem*: "Find an optical system containing freeform refractive/reflective surfaces that provides the desired target light distribution for

✉ N. K. Yadav
n.k.yadav@tue.nl

J. H. M. ten Thije Boonkkamp
j.h.m.tenthijeboonkkamp@tue.nl

W. L. IJzerman
wilbert.ijzerman@signify.com

[1] Department of Mathematics and Computer Science, CASA, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands

[2] Signify Research, High Tech Campus 7, 5656 AE Eindhoven, The Netherlands

a given source distribution". Inverse optical design has a wide range of applications from LED based optical products for street lighting and car headlights to applications in medical science, image processing and lithography [1,4].

To convert a given emittance profile with parallel light rays into a desired illuminance profile with parallel light rays, one requires at least two freeform lens/mirror surfaces [2,5]. This freeform problem can be formulated as a second order partial differential equation of Monge–Ampère (MA) type, with transport boundary conditions, applying the laws of geometrical optics and energy conservation [2,3,6,7]. For the two-reflector problem [2,8], one can obtain the following mathematical formulation using properties of geometrical optics, i.e.,

$$u_1(\mathrm{x}) + u_2(\mathrm{y}) = c(\mathrm{x}, \mathrm{y}) := a_1 + a_2|\mathrm{x} - \mathrm{y}|^2, \tag{1}$$

where $a_1, a_2$ are constants and $u_1(\mathrm{x})$, $u_2(\mathrm{y})$ represent the location of the optical surfaces, and $|.|$ denotes the 2-norm for vectors. The right hand side function $c(\mathrm{x}, \mathrm{y})$ in the above expression is a quadratic (cost) function. Assuming convex/concave reflective surfaces the ray-trace map can be uniquely expressed as the gradient of $u_1$, i.e.,

$$\boldsymbol{m}(\mathrm{x}) = \nabla u_1(\mathrm{x}). \tag{2}$$

Furthermore, using conservation of energy, one can derive a second order partial differential equation (PDE) of MA-type. We refer to this equation as the standard MA-equation, representing optical systems characterized by a quadratic cost function.

In this article, we show that a similar mathematical expression can be obtained for the freeform surfaces of a lens with parallel ingoing and outgoing light rays applying the laws of geometrical optics:

$$u_1(\mathrm{x}) + u_2(\mathrm{y}) = c(\mathrm{x}, \mathrm{y}) := b_1 - \sqrt{b_2 + b_3|\mathrm{x} - \mathrm{y}|^2}, \tag{3}$$

where $b_1, b_2, b_3$ are constants, and $u_1(\mathrm{x})$, $u_2(\mathrm{y})$ represent the first and second refractive surface of the lens, respectively. For the freeform lens the cost function $c(\mathrm{x}, \mathrm{y})$ is no longer a quadratic function, and the ray-trace map can not be expressed as the gradient of some function, we provide more details in Sect. 2. Energy conservation results in a complicated MA-type equation. In this article, we will present a numerical algorithm to compute the freeform surfaces of a lens characterized by a non-quadratic cost function. A rigorous analysis of the existence and uniqueness of weak solutions of similar lens design problems is presented in Oliker's work [9,10].

There are several numerical methods which can be employed to compute freeform surfaces of optical systems characterized by a quadratic cost function. However, to the best of our knowledge, this paper is the first to describe a numerical method for the MA-equation with non-quadratic cost function. Froese et al. [11–13] solve the standard MA-equation within the framework of optimal mass transport (OMT). Applying the theory of viscosity solutions, they refine the solution using an iterative Fourier-transform algorithm with overcompensation. In recent publications [5,14,15], the authors obtain freeform optical surfaces by solving the standard MA-equation using Newton iteration. These numerical methods require an initial guess which is obtained through the OMT problem. Brix et al. [3,16] solve the standard inverse design problem using a collocation method with a tensor-product B-spline basis. Glimm and Oliker [8,17] show that the illuminance control problem can be solved using an optimization approach instead of solving a MA-type differential equation. Further, a similar approach to design freeform surfaces of a lens is developed by Rubinstein and Wolansky [18].

A least-squares (LS) method [2,7] has been presented to solve the standard MA-equation to compute single reflector/lens or double reflector freeform surfaces optical systems. The method provides the optical mapping which transfers the given emittance of the source into the desired illuminance at the target, and the freeform surfaces are obtained via this mapping.

However, the coupled freeform lens surfaces design problem corresponds to a non-quadratic cost function. The goal of this paper is to present a numerical method which is applicable to design an optical system corresponding to a non-quadratic cost function. Here, we present a fast and effective extended least-squares (ELS) method to construct the freeform surfaces of the lens. The ELS-method is a two-stage procedure like the LS-method: first we determine an optimal mapping by minimizing three functionals iteratively, next, we compute the freeform surfaces from the converged mapping. In the first stage, there are two nonlinear minimization steps, which can be performed point-wise, like in the LS-method. In the third step two elliptic partial differential equations have to be solved. For the LS-method, these are decoupled Poisson equations. However, in the ELS-method these are coupled elliptic equations.

Our least-squares method is quite generally applicable since it can handle arbitrary twice differentiable cost functions $c(\mathrm{x}, \mathrm{y})$, also in other fields of science and engineering such as optimal transport theory, shape optimization, compression modeling, relativistic theory, incompressible fluid flow, economics, astrophysics, atmospheric sciences etc. For the interested reader, we refer to the following: Evans' survey notes [19], articles of Bouchittè-Buttazzo [20,21], Gangbo's lecture notes [22], and paper of Benamou–Brenier [23]. However, we restrict ourselves to the computation of freeform optical systems.

This paper is structured as follows. In Sect. 2 we explain the geometrical structure of the optical system and formulate the mathematical model. The detailed procedure of the proposed least-squares method is shown in Sect. 3. We apply the numerical method to four test problems in Sect. 4 and verify the solutions using a ray tracing algorithm [2]. Finally, a brief discussion and concluding remarks are given in Sect. 5.

## 2 Formulation of the Problem

The geometrical structure of a lens optical system is shown schematically in Fig. 1. Let $(x_1, x_2, z) \in \mathbb{R}^3$ denote the Cartesian coordinates with $z$ the horizontal coordinate and $\mathrm{x} = (x_1, x_2) \in \mathbb{R}^2$ the coordinates in the plane $z = 0$, denoted by $\alpha_1$, and let $\mathcal{S}$ be a bounded source domain in the plane $\alpha_1$. The source $\mathcal{S}$ emits parallel light rays which propagate in the positive $z$-direction. The emittance, i.e., luminous flux per unit area (for an introduction to photometry quantities see e.g. [24, p. 7–9]), of the source is given by $f(\mathrm{x})$ [lm/m$^2$], $\mathrm{x} \in \mathcal{S}$, where $f$ is a non-negative integrable function on the domain $\mathcal{S}$. The target at a distance $\ell > 0$ from the plane $\alpha_1$ is denoted by $\mathcal{T}$.

The incoming light rays are refracted at the first lens surface $\mathcal{L}_1$, propagate through the lens and are refracted again at the second lens surface $\mathcal{L}_2$, to create a parallel bundle of light rays in the positive $z$-direction. The index of refraction of the lens $n > 1$ and the surrounding medium is air with refractive index of unity. The lens surfaces are defined as $z \equiv u_1(\mathrm{x})$, $\mathrm{x} \in \mathcal{S}$ and $w \equiv \ell - z = u_2(\mathrm{y})$, $\mathrm{y} \in \mathcal{T}$, respectively, where $\mathrm{y} = (y_1, y_2) \in \mathbb{R}^2$ are the Cartesian coordinates of the target plane $\alpha_2$.

The goal is to design a lens system such that after two refractions the refracted rays must form a parallel beam, propagating in the positive $z$-direction, and provide a prescribed illuminance $g(\mathrm{y})$ [lm/m$^2$] at the plane $\alpha_2 : z = \ell$, where $g > 0$ is a positive integrable

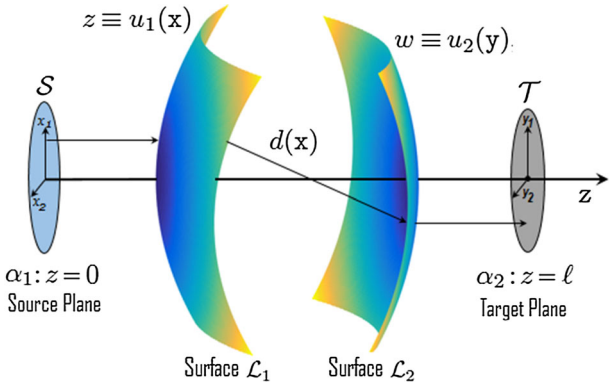



**Fig. 1** Sketch of a freeform lens optical system

function on the domain $\mathcal{T}$. It is assumed that both $\mathcal{L}_1$ and $\mathcal{L}_2$ are perfect lens surfaces and no energy is lost in the refraction.

### 2.1 Geometrical Formulation of the Freeform Lens

In this section, we first give an expression for the ray-trace map, and secondly we derive a mathematical formulation for the location of the freeform surfaces using the laws of geometrical optics.

The mapping $\boldsymbol{m}$ can be derived by tracing a typical ray through the optical system. Let us consider a ray emitted from a position $\mathrm{x} \in \mathcal{S}$ on the source and propagating in the positive $z$-direction, let $\hat{\boldsymbol{s}}$ be the unit direction of the incident ray. The ray strikes the first lens surface $\mathcal{L}_1$, refracts off in direction $\hat{\boldsymbol{t}}$, strikes the second lens surface $\mathcal{L}_2$, and reflects off, again in the direction $\hat{\boldsymbol{s}}$. The unit surface normal of the first lens surface $\mathcal{L}_1$, directed towards the light source, is given by

$$\hat{\boldsymbol{n}}_1 = \frac{(\nabla u_1, -1)}{\sqrt{|\nabla u_1|^2 + 1}}. \tag{4}$$

Throughout this article, we use the convention that a hat denotes a unit vector. According to Snell's law [24,25], the direction $\hat{\boldsymbol{t}} = \hat{\boldsymbol{t}}(\mathrm{x})$ of the refracted ray can be expressed as

$$\hat{\boldsymbol{t}} = \eta \hat{\boldsymbol{s}} + F(|\nabla u_1|; \eta)\hat{\boldsymbol{n}}_1, \tag{5}$$

where $\eta = 1/n < 1$ with $n$ the refractive index of the lens and

$$F(z; \eta) = \frac{1}{\sqrt{z^2 + 1}}\left[\eta - \sqrt{1 + (1 - \eta^2)z^2}\right]. \tag{6}$$

If we write $\hat{\boldsymbol{t}} = (t_1, t_2, t_3)^T$ then the first two components of the vector $\hat{\boldsymbol{t}}$, can be written as a function of the third component of the vector $\hat{\boldsymbol{t}}$ as

$$\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = (\eta - t_3)\nabla u_1. \tag{7}$$

The image on the target of the point $\mathrm{x} \in \mathcal{S}$ is the point $\mathrm{y} \in \mathcal{T}$ under the ray trace mapping $\boldsymbol{m}$, i.e., $\mathrm{y} = \boldsymbol{m}(\mathrm{x})$, $\mathrm{x} \in \mathcal{S}$. This mapping can be obtained by the projection of $\hat{\boldsymbol{t}}$ on the plane $\alpha_1$, i.e.,

$$\boldsymbol{m}(\mathrm{x}) = \mathrm{x} + \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} d(\mathrm{x}), \tag{8}$$

where $d(\mathrm{x})$ is the distance between surfaces $\mathcal{L}_1$ and $\mathcal{L}_2$ along the ray refracted in the direction $\hat{\boldsymbol{t}}(\mathrm{x})$. The distance $d(\mathrm{x})$ between the lens surfaces can be obtained using properties of geometrical optics: The total optical path length $L(\mathrm{x})$ corresponding to the ray associated with a point $\mathrm{x} \in \mathcal{S}$, is given by

$$L(\mathrm{x}) = u_1(\mathrm{x}) + nd(\mathrm{x}) + u_2(\mathrm{y}). \tag{9}$$

The theorem of Malus and Dupin (the principle of equal optical path lengths) states that the total optical path length between any two orthogonal wavefronts is the same for all rays [26, p. 130]. As we deal with two parallel beams of light rays, the wavefront coincides with planes $\alpha_1$ and $\alpha_2$. Therefore, the total optical path length will be independent of the position vector $\mathrm{x}$, i.e., $L(\mathrm{x}) = L$. The horizontal distance $\ell$ between the source and the target plane is given by

$$\ell = u_1(\mathrm{x}) + (\hat{\boldsymbol{s}} \cdot \hat{\boldsymbol{t}})d(\mathrm{x}) + u_2(\mathrm{y}). \tag{10}$$

Subtracting Eqs. (9) and (10), and using Eq. (5), we obtain the following expression

$$d(\mathrm{x}) = \frac{\beta}{n - t_3}, \tag{11}$$

where where $\beta = L - \ell$ is the "reduced" optical path length. Substituting (7) and (11) in (8), we have

$$\boldsymbol{m}(\mathrm{x}) = \mathrm{x} + \beta \frac{\eta - t_3}{n - t_3}. \tag{12}$$

Now, substituting $t_3$ in the above equation from the law of refraction (5), the mapping $\boldsymbol{m}$ is given by the relation

$$\boldsymbol{m}(\mathrm{x}) = \mathrm{x} - \frac{\beta \nabla u_1(\mathrm{x})}{\sqrt{n^2 + (n^2 - 1)|\nabla u_1|^2}}. \tag{13}$$

Next, we derive a mathematical expressions for the location of the lens surfaces. An alternative expression for the distance $d$ reads

$$d^2 = \big(\ell - u_1(\mathrm{x}) - u_2(\mathrm{y})\big)^2 + |\mathrm{x} - \mathrm{y}|^2. \tag{14}$$

Thus, from Eqs. (9) and (14), we obtain

$$n^2\big(\ell - u_1(\mathrm{x}) - u_2(\mathrm{y})\big)^2 + n^2|\mathrm{x} - \mathrm{y}|^2 = \big(L - u_1(\mathrm{x}) - u_2(\mathrm{y})\big)^2,$$

which can be rewritten as

$$\left[u_1(\mathrm{x}) + u_2(\mathrm{y}) + \frac{L - n^2\ell}{n^2 - 1}\right]^2 + \frac{n^2}{n^2 - 1}|\mathrm{x} - \mathrm{y}|^2 = \left(\frac{n\beta}{n^2 - 1}\right)^2,$$

and after elementary algebraic derivations, we obtain

$$u_1(\mathrm{x}) + u_2(\mathrm{y}) = \frac{n^2\ell - L}{n^2 - 1} \pm \frac{n}{n^2 - 1}\sqrt{\beta^2 - (n^2 - 1)|\mathrm{x} - \mathrm{y}|^2}. \tag{15}$$

This is a mathematical expression for the location of the lens surfaces but the sign in front of the square root is unknown yet. To determine this we proceed as follows. Using Eqs. (9)

(with $L(\mathrm{x}) = L$) and (14), we can show that $\beta^2 - (n^2-1)|\mathrm{x}-\mathrm{y}|^2 = (n\beta - d(n^2-1))^2 \geq 0$. Substituting, this expression in Eq. (15), we obtain

$$u_1(\mathrm{x}) + u_2(\mathrm{y}) = \frac{n^2\ell - L}{n^2-1} \pm \frac{n}{n^2-1}\big|n\beta - d(n^2-1)\big|. \tag{16}$$

First, we check the sign of the expression $n\beta - d(n^2-1)$. Substituting $d$ from Eq. (11), the expression becomes

$$n\beta - d(n^2-1) = \beta\frac{1-nt_3}{n-t_3}.$$

Since $\beta > 0$ and $n - t_3 > 0$, it remains to check the sign of $1 - nt_3$. Using the vectorial form of the law of refraction (5) and expression (6), we can write

$$1 - nt_3 = \frac{1 - \sqrt{n^2 + (n^2-1)|\nabla u_1|^2}}{|\nabla u_1|^2 + 1} < 0, \tag{17}$$

as $n > 1$. Thus we have to choose the negative sign in front of the absolute value in Eq. (16). Hence, we obtain

$$u_1(\mathrm{x}) + u_2(\mathrm{x}) = \frac{n^2\ell - L}{n^2-1} \pm \frac{nd(n^2-1) - n^2\beta}{n^2-1}.$$

Substituting $d$ from relation (9), the above expression becomes

$$u_1(\mathrm{x}) + u_2(\mathrm{y}) = \frac{n^2\ell - L}{n^2-1} \pm \Big(L - (u_1(\mathrm{x}) + u_2(\mathrm{y})) - \frac{n^2}{n^2-1}\beta\Big). \tag{18}$$

In the above equation, the right hand side equals the left hand side for the minus sign, therefore we have to choose minus sign in (15). Thus the mathematical expression for the lens surfaces becomes

$$\begin{aligned} u_1(\mathrm{x}) + u_2(\mathrm{y}) &= c(\mathrm{x},\mathrm{y}),\\ c(\mathrm{x},\mathrm{y}) &= \ell - \frac{\beta}{n^2-1} - \frac{n}{n^2-1}\sqrt{\beta^2 - (n^2-1)|\mathrm{x}-\mathrm{y}|^2}. \end{aligned} \tag{19}$$

These kind of freeform optical design problems are closely related to the mass transport problem [10,27]. The right hand side function $c(\mathrm{x},\mathrm{y})$ is known as the cost function in OMT theory.

To conclude, we have derived a mathematical formulation representing the freeform lens optical system which is given in (19). Also, we obtained the expression (13) for the ray-trace mapping $\boldsymbol{m}$. Next, we formulate a second order partial differential equation for the freeform lens.

### 2.2 Energy Conservation for the Freeform Lens

Recall, that $f \geq 0$ and $g > 0$ are integrable functions and no energy is lost in the light transfer process. Thus energy conservation is given by

$$\iint_{\mathcal{S}} f(\mathrm{x})\mathrm{d}\mathrm{x} = \iint_{\mathcal{T}} g(\mathrm{y})\mathrm{d}\mathrm{y}. \tag{20}$$

The key tool for the design of such an optical system is to find a mapping $\mathrm{y} = \boldsymbol{m}(\mathrm{x}) : \mathcal{S} \to \mathcal{T}$ that satisfies the energy conservation constraint (20) for each measurable set $\mathcal{A} \subset \mathcal{S}$, i.e.,

$$\iint_{\mathcal{A}} f(\mathrm{x})\mathrm{d}\mathrm{x} = \iint_{\boldsymbol{m}(\mathcal{A})} g(\mathrm{y})\mathrm{d}\mathrm{y}, \tag{21}$$

and after a change of variables the constraint becomes

$$f(\mathrm{x}) = g(\boldsymbol{m}(\mathrm{x}))|\det(\mathrm{D}\boldsymbol{m}(\mathrm{x}))|, \quad \forall \mathrm{x} \in \mathcal{S}, \tag{22}$$

where $\mathrm{D}\boldsymbol{m}$ is the Jacobian of the mapping $\boldsymbol{m}$, which measures the expansion/contraction of a tube of rays due to the two refractions. The accompanying boundary condition is derived from the condition that all the light from the source domain $\mathcal{S}$ must be transferred into the target domain $\mathcal{T}$, i.e.,

$$\boldsymbol{m}(\partial\mathcal{S}) = \partial\mathcal{T}, \tag{23}$$

stating that the boundary of the source $\mathcal{S}$ is mapped to the boundary of the target $\mathcal{T}$. This is a consequence of the edge ray principle [28].

Next, we derive a MA-type equation for the freeform lens using the energy conservation constraint (22) and the mathematical formulation (19) for the location of the lens surfaces. We assume that both lens surfaces $u_1$ and $u_2$ are either c-convex or c-concave functions. According to the following definition, the lens surfaces $u_1$ and $u_2$ are c-convex if

$$u_1(\mathrm{x}) = \max_{\mathrm{y}\in\mathcal{T}}\{c(\mathrm{x},\mathrm{y}) - u_2(\mathrm{y})\} \quad \forall \quad \mathrm{x} \in \mathcal{S}, \tag{24a}$$

$$u_2(\mathrm{y}) = \max_{\mathrm{x}\in\mathcal{S}}\{c(\mathrm{x},\mathrm{y}) - u_1(\mathrm{x})\} \quad \forall \quad \mathrm{y} \in \mathcal{T}, \tag{24b}$$

alternatively, these are c-concave if

$$u_1(\mathrm{x}) = \min_{\mathrm{y}\in\mathcal{T}}\{c(\mathrm{x},\mathrm{y}) - u_2(\mathrm{y})\} \quad \forall \quad \mathrm{x} \in \mathcal{S}, \tag{25a}$$

$$u_2(\mathrm{y}) = \min_{\mathrm{x}\in\mathcal{S}}\{c(\mathrm{x},\mathrm{y}) - u_1(\mathrm{x})\} \quad \forall \quad \mathrm{y} \in \mathcal{T}. \tag{25b}$$

For a continuously differentiable function $c \in C^1(\mathcal{S} \times \mathcal{T})$, the c-convex/concave functions $u_1$ and $u_2$ are Lipschitz continuous [27,29], and the mapping $\mathrm{y} = \boldsymbol{m}(\mathrm{x})$ is implicitly given by the relation

$$\nabla_{\mathrm{x}} u_1(\mathrm{x}) = \nabla_{\mathrm{x}} c(\mathrm{x}, \boldsymbol{m}(\mathrm{x})), \tag{26}$$

which is a necessary condition for (24b) and (25b), and holds under the condition that the Jacobi matrix $\boldsymbol{C} = \mathrm{D}_{\mathrm{xy}} c$ defined by

$$\boldsymbol{C} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} = \begin{pmatrix} \frac{\partial^2 c}{\partial x_1 \partial y_1} & \frac{\partial^2 c}{\partial x_1 \partial y_2} \\ \frac{\partial^2 c}{\partial x_2 \partial y_1} & \frac{\partial^2 c}{\partial x_2 \partial y_2} \end{pmatrix}, \tag{27}$$

is invertible. For our optical problem the mapping $\boldsymbol{m}$ given by relation (13) satisfies relation (26) indeed.

The matrix $\boldsymbol{C}$ is symmetric negative semi-definite which is a consequence of the fact that the function $c$ depends on $|\mathrm{x} - \mathrm{y}|$. This can be verified as follows: let us rewrite the cost function (19) as

$$c(\mathrm{x},\mathrm{y}) = \ell - \frac{\beta}{n^2 - 1} + \tilde{c}(\mathrm{x},\mathrm{y}), \tag{28a}$$

$$\tilde{c}(\mathrm{x},\mathrm{y}) = -\frac{n}{n^2 - 1}\sqrt{\beta^2 - (n^2 - 1)|\mathrm{x} - \mathrm{y}|^2}. \tag{28b}$$

By differentiating (28) with respect to $\mathrm{x}$ and $\mathrm{y}$, we obtain

$$\nabla_{\mathrm{x}} c(\mathrm{x},\mathrm{y}) = -\frac{n^2}{n^2-1}\frac{1}{\tilde{c}}(\mathrm{x}-\mathrm{y}), \tag{29a}$$

$$\nabla_{\mathrm{y}} c(\mathrm{x},\mathrm{y}) = \frac{n^2}{n^2-1}\frac{1}{\tilde{c}}(\mathrm{x}-\mathrm{y}), \tag{29b}$$

which gives

$$\nabla_{\mathrm{x}} c(\mathrm{x},\mathrm{y}) + \nabla_{\mathrm{y}} c(\mathrm{x},\mathrm{y}) = 0. \tag{29c}$$

Differentiating one more time with respect to $\mathrm{x}$, we conclude that

$$\boldsymbol{C} = \mathrm{D}_{\mathrm{xy}} c = -\mathrm{D}_{\mathrm{xx}} c. \tag{30}$$

Evaluating all derivatives, we obtain the following expression

$$\boldsymbol{C} = \frac{n^2}{n^2-1}\frac{\boldsymbol{I}}{\tilde{c}} + \left(\frac{n^2}{n^2-1}\right)^2 \frac{1}{\tilde{c}^3}\begin{pmatrix} (x_1-y_1)^2 & (x_1-y_1)(x_2-y_2) \\ (x_1-y_1)(x_2-y_2) & (x_2-y_2)^2 \end{pmatrix}. \tag{31}$$

We can rewrite the above expression as follows:

$$\boldsymbol{C} = \frac{\gamma^2}{\tilde{c}^3}\left(\frac{\tilde{c}^2}{\gamma}\boldsymbol{I} + (\mathrm{x}-\mathrm{y})(\mathrm{x}-\mathrm{y})^T\right), \tag{32}$$

where $\gamma = n^2/(n^2-1) > 0$. Since $\tilde{c} < 0$, we conclude that $\det(\boldsymbol{C}) > 0$ and $\mathrm{tr}(\boldsymbol{C}) \leq 0$ hence the matrix $\boldsymbol{C}$ is symmetric negative semi-definite.

Since the function $c(\mathrm{x},\mathrm{y})$ defined in (19) is continuously differentiable, from relation (26), we deduce

$$\boldsymbol{C}\mathrm{D}\boldsymbol{m}(\mathrm{x}) = \mathrm{D}^2 u_1(\mathrm{x}) - \mathrm{D}_{\mathrm{xx}} c \equiv \boldsymbol{P}, \tag{33}$$

where $\mathrm{D}^2 u_1$ is the Hessian of $u_1$. The matrix $\boldsymbol{P} = \mathrm{D}^2 u_1(\mathrm{x}) - \mathrm{D}_{\mathrm{xx}} c$ is negative semi-definite for a c-concave pair $(u_1, u_2)$ and positive semi-definite for a c-convex pair $(u_1, u_2)$. In the following, we discuss the convex case, thus we require the matrix $\boldsymbol{P}$ to be positive semi-definite. Substituting $\mathrm{D}\boldsymbol{m}$ from (33) into the energy conservation condition (22), we obtain

$$\frac{\det(\boldsymbol{P}(\mathrm{x}))}{\det(\boldsymbol{C}(\mathrm{x},\boldsymbol{m}(\mathrm{x})))} = \frac{f(\mathrm{x})}{g(\boldsymbol{m}(\mathrm{x}))}, \quad \forall \ \ \mathrm{x} \in \mathcal{S}. \tag{34}$$

We know that the $2 \times 2$ matrix $\boldsymbol{P}$ is positive semi-definite if and only if

$$\mathrm{tr}(\boldsymbol{P}) \geq 0 \quad \text{and} \quad \det(\boldsymbol{P}) \geq 0. \tag{35}$$

Because $\det(\boldsymbol{C}) > 0$ and the right hand side functions $f \geq 0$, $g > 0$ in Eq. (34), it is obvious that $\det(\boldsymbol{P}) \geq 0$. So, the only requirement left is $\mathrm{tr}(\boldsymbol{P}) \geq 0$ for convex optical surfaces.

In the following section, we give a detailed description of the ELS-algorithm to solve the MA-equation (34) with the boundary condition (23) and constraints (35). The method presented here is based on [7]. Compared to [7] we deal with a non-quadratic cost function that results in the presence of the matrix $\boldsymbol{C}$ in (34).

## 3 Numerical Algorithm

Prins et al. [7] introduced a least-squares method to compute single freeform surfaces governed a quadratic cost function. Further, we applied the method to design a two-reflector

optical system [2], which is also a quadratic cost problem. Our version of the least-squares method was inspired by publications by Caboussat et al. [30,31], who developed a least-squares method for the Monge–Ampère–Dirichlet problem. An extension of their method to the three-dimensional equation is presented in [32].

In this section, we extend the least-squares method to compute the freeform surfaces of a lens characterized by a non-quadratic cost function. The ELS-method is a two-stage procedure. In the first stage we calculate the optimal mapping by minimizing three functionals iteratively, and in the second stage we compute the freeform surfaces from the mapping in the least squares sense.

### 3.1 First Stage: Calculation of the Mapping

First, we calculate the mapping $\boldsymbol{m}$ using the least-squares method for the lens optical system as follows: we enforce the equality $\boldsymbol{C}\mathrm{D}\boldsymbol{m} = \boldsymbol{P}$ by minimizing the following functional

$$J_{\mathrm{I}}(\boldsymbol{m}, \boldsymbol{P}) = \frac{1}{2} \iint_{\mathcal{S}} ||\boldsymbol{C}\mathrm{D}\boldsymbol{m} - \boldsymbol{P}||^2 \mathrm{d}\mathrm{x}. \tag{36}$$

The norm used in this functional is the Frobenius norm, which is defined as follows. Let $\boldsymbol{A} : \boldsymbol{B}$ denote the Frobenius inner product of the matrices $\boldsymbol{A} = (a_{ij})$ and $\boldsymbol{B} = (b_{ij})$, defined by

$$\boldsymbol{A} : \boldsymbol{B} = \sum_{i,j} a_{ij} b_{ij}, \tag{37}$$

the Frobenius norm is then defined as $||\boldsymbol{A}|| = \sqrt{\boldsymbol{A} : \boldsymbol{A}}$. Next, we address the boundary by minimizing the functional

$$J_{\mathrm{B}}(\boldsymbol{m}, \boldsymbol{b}) = \frac{1}{2} \oint_{\partial\mathcal{S}} |\boldsymbol{m} - \boldsymbol{b}|^2 \mathrm{d}s. \tag{38}$$

We combine the functionals $J_{\mathrm{I}}$ for the interior and $J_{\mathrm{B}}$ for the boundary domain by a weighted average:

$$J(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b}) = \alpha J_{\mathrm{I}}(\boldsymbol{m}, \boldsymbol{P}) + (1 - \alpha) J_{\mathrm{B}}(\boldsymbol{m}, \boldsymbol{b}). \tag{39}$$

The parameter $\alpha$ $(0 < \alpha < 1)$ controls the weight of the first functional compared to the second functional. The variables $\boldsymbol{b}$, $\boldsymbol{P}$, and $\boldsymbol{m}$ are elements of the following spaces

$$\mathcal{B} = \{\boldsymbol{b} \in [C(\partial\mathcal{S})]^2 \mid \boldsymbol{b}(\mathrm{x}) \in \partial\mathcal{T}\}, \tag{40a}$$

$$\mathcal{P}(\boldsymbol{m}) = \left\{ \boldsymbol{P} \in [C^1(\mathcal{S})]^{2\times 2} \;\middle|\; \frac{\det(\boldsymbol{P})}{\det(\boldsymbol{C}(\cdot, \boldsymbol{m}))} = \frac{f}{g(\boldsymbol{m})} \right\}, \tag{40b}$$

$$\mathcal{M} = [C^2(\mathcal{S})]^2, \tag{40c}$$

respectively. The minimizer gives us the mapping $\boldsymbol{m}$ which is implicitly related to the surface function $u_1$. We calculate this minimizer by repeatedly minimizing over the three spaces separately. We start with an initial guess $\boldsymbol{m}^0$, which will be specified shortly, and we calculate the matrix $\boldsymbol{C}(\mathrm{x}, \boldsymbol{m}^0)$ at the initial guess $\boldsymbol{m}^0$. Subsequently, we perform the iteration

$$\boldsymbol{b}^{n+1} = \operatorname*{argmin}_{\boldsymbol{b}\in\mathcal{B}} J_{\mathrm{B}}(\boldsymbol{m}^n, \boldsymbol{b}), \tag{41a}$$

$$\boldsymbol{P}^{n+1} = \operatorname*{argmin}_{\boldsymbol{P}\in\mathcal{P}(\boldsymbol{m}^n)} J_{\mathrm{I}}(\boldsymbol{m}^n, \boldsymbol{P}), \tag{41b}$$

$$\boldsymbol{m}^{n+1} = \operatorname*{argmin}_{\boldsymbol{m}\in\mathcal{M}} J(\boldsymbol{m}, \boldsymbol{P}^{n+1}, \boldsymbol{b}^{n+1}). \tag{41c}$$

Next, we compute the matrix $\boldsymbol{C}(\mathrm{x}, \boldsymbol{m}^n)$ before going to the next iteration.

We initialize our minimization procedure by constructing an initial guess $\boldsymbol{m}^0$ which maps a bounding box of the source area $\mathcal{S}$ to a bounding box of the target area $\mathcal{T}$. Without loss of generality we assume the smallest bounding box of the source and the target are rectangular and denote these by $[a_{\min}, a_{\max}] \times [b_{\min}, b_{\max}]$ and $[c_{\min}, c_{\max}] \times [d_{\min}, d_{\max}]$, respectively. Then the initial guess reads:

$$m_1^0 = \frac{x_1 - a_{\min}}{a_{\max} - a_{\min}} c_{\min} + \frac{a_{\max} - x_1}{a_{\max} - a_{\min}} c_{\max}, \tag{42a}$$

$$m_2^0 = \frac{x_2 - b_{\min}}{b_{\max} - b_{\min}} d_{\min} + \frac{b_{\max} - x_2}{b_{\max} - b_{\min}} d_{\max}. \tag{42b}$$

Note that the corresponding Jacobi matrix $\mathrm{D}\boldsymbol{m}^0$ of the initial condition is symmetric (in fact diagonal) negative definite. The matrix $\boldsymbol{C}$ is also negative definite, moreover from relation (32) we conclude that $c_{11}, c_{22} < 0$ which implies that the matrix $\boldsymbol{P} = \boldsymbol{C}(\mathrm{x}, \boldsymbol{m}^0)\mathrm{D}\boldsymbol{m}^0$ is positive definite. Thus this initialization satisfies our requirement $\mathrm{tr}(\boldsymbol{P}) \geq 0$.

Obviously, the minimization steps in (41) as well as the computation of the optical surfaces is done numerically. To that purpose we discretize the source $\mathcal{S}$ with a standard rectangular $N_1 \times N_2$ grid for some $N_1, N_2 \in \mathbb{N}$, so the grid points $\mathrm{x}_{ij} = (x_{1,i}, x_{2,j})$ are defined as

$$x_{1,i} = a_{\min} + (i-1)h_1, \quad h_1 = \frac{a_{\max} - a_{\min}}{N_1 - 1}, \quad i = 1, \ldots, N_1, \tag{43a}$$

$$x_{2,j} = b_{\min} + (j-1)h_2, \quad h_2 = \frac{b_{\max} - b_{\min}}{N_2 - 1}, \quad j = 1, \ldots, N_2. \tag{43b}$$

We start the iteration process (41) using initial guess $\boldsymbol{m}^0$. Here each iteration consists of four steps: we perform the three minimization steps (41a)–(41c), and fourthly we update the matrix $C$ at every iteration. In this article, we give a detailed description of the minimization steps (41b) and (41c). The minimization step first (41a) is simple and direct, and performed point-wise because no derivative of $\boldsymbol{b}$ with respect to x appears in the functional, more details can be found in [7].

Finally, from the converged mapping $\boldsymbol{m}$, we compute the first lens surface $u_1$ via relation (26) in a least-squares sense, and the second lens surface $u_2$ from relation (19), see Sect. 3.2.

**Minimizing procedure for $\boldsymbol{P}$**

We assume $\boldsymbol{m}$ fixed and minimize $J_1(\boldsymbol{m}, \boldsymbol{P})$ over the matrices under the condition (34). Since the integrand of $J_1(\boldsymbol{m}, \boldsymbol{P})$ does not contain derivatives of $\boldsymbol{P}$, the minimization procedure can be done pointwise. So, we need to minimize $||\boldsymbol{C}\boldsymbol{D} - \boldsymbol{P}||$ for each grid point $\mathrm{x}_{ij} \in \mathcal{S}$, where $\boldsymbol{D}$ is the central difference approximation of $\mathrm{D}\boldsymbol{m}$. Let's define

$$\boldsymbol{P} = \begin{pmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix}, \quad \boldsymbol{Q} = \boldsymbol{C}\boldsymbol{D} = \begin{pmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{pmatrix}, \tag{44a}$$

with

$$d_{11} = \delta_{x_1} m_1, \quad d_{12} = \delta_{x_2} m_1, \quad d_{21} = \delta_{x_1} m_2, \quad d_{22} = \delta_{x_2} m_2, \tag{44b}$$

where $\delta_{x_1}$ and $\delta_{x_2}$ are the central difference approximations of $\partial/\partial x_1$ and $\partial/\partial x_2$, respectively. Note that, the matrices $\boldsymbol{C}$, $\boldsymbol{D}$, $\boldsymbol{P}$ and $\boldsymbol{Q}$ all depend on $\mathrm{x}_{ij}$. For the sake of brevity we omit $\mathrm{x}_{ij}$. Moreover, we want to avoid crossing grid lines, i.e., intersection of images of grid lines in $\mathcal{S}$, and for that reason we require $d_{11}, d_{22} > 0$. This can be achieved by imposing

$$d_{11} = \max(\delta_{x_1} m_1, \varepsilon), \quad d_{22} = \max(\delta_{x_2} m_2, \varepsilon), \tag{45}$$

for a threshold value $\varepsilon > 0$. This implies that $m_1(\mathrm{x}_{i+1,j}) < m_1(\mathrm{x}_{i,j})$ and $m_2(\mathrm{x}_{i,j+1}) < m_2(\mathrm{x}_{i,j})$ for all $\mathrm{x}_{i,j} \in \mathcal{S}$, which assures that there is no crossing of grid lines. In our computations we choose $\varepsilon = 10^{-8}$.

Note that the matrix $\boldsymbol{P}$ is symmetric but the matrix $\boldsymbol{D}$ need not be symmetric and $d_{12}, d_{21} < 0$ is possible. Next we define the function

$$H(p_{11}, p_{22}, p_{12}) = \frac{1}{2}||\boldsymbol{Q} - \boldsymbol{P}||^2. \tag{46}$$

Also, we define the matrix $\boldsymbol{Q}_\mathrm{S}$ as the symmetric part of the matrix $\boldsymbol{Q}$, i.e.,

$$\boldsymbol{Q}_\mathrm{S} = \frac{1}{2}(\boldsymbol{Q} + \boldsymbol{Q}^T) = \begin{pmatrix} q_{11} & q_\mathrm{S} \\ q_\mathrm{S} & q_{22} \end{pmatrix}, \tag{47}$$

with $q_\mathrm{S} = \frac{1}{2}(q_{12} + q_{21})$. The function $H_S$ corresponding to the symmetric matrix $\boldsymbol{Q}_S$ is defined as

$$\begin{aligned} H_\mathrm{S}(p_{11}, p_{22}, p_{12}) &= \frac{1}{2}||\boldsymbol{Q}_\mathrm{S} - \boldsymbol{P}||^2 \\ &= H(p_{11}, p_{22}, p_{12}) - \frac{1}{4}(q_{12} - q_{21})^2. \end{aligned} \tag{48}$$

Since $(q_{12} - q_{21})^2$ is independent of $p_{11}$, $p_{22}$ and $p_{12}$, and because we are only interested in the minimizer $(p_{11}, p_{22}, p_{12})$ and not in its value $H(p_{11}, p_{22}, p_{12})$, we minimize $H_\mathrm{S}$ instead of $H$. For each grid point $\mathrm{x}_{ij} = (x_{1,i}, x_{2,j}) \in \mathcal{S}$ we have the following quadratic minimization problem

$$\text{minimize} \quad H_\mathrm{S}(p_{11}, p_{22}, p_{12}), \tag{49a}$$

$$\text{subject to} \quad \det(\boldsymbol{P}) = \frac{f}{g}\det(\boldsymbol{C}), \tag{49b}$$

$$\mathrm{tr}(\boldsymbol{P}) \geq 0. \tag{49c}$$

This problem can be solved analytically, and we will show that for given $q_{11}, q_{22}, q_\mathrm{S}$ and $f/g$ there exist at least one and at most four real solutions, see "Appendix A". From these we have to select the ones that give rise to a negative semi-definite matrix $\boldsymbol{P}$, and we will also show that this is always possible. Finally, we compare the values of $H_\mathrm{S}(p_{11}, p_{22}, p_{12})$ to find the global minimum.

The possible minimizers of (49) are obtained introducing the Lagrangian function $\Lambda$, defined as

$$\Lambda(p_{11}, p_{22}, p_{12}; \mu) = \frac{1}{2}\left\| \boldsymbol{Q}_\mathrm{S} - \boldsymbol{P} \right\|^2 + \mu\left(\det(\boldsymbol{P}) - \frac{f}{g}\det(\boldsymbol{C})\right), \tag{50}$$

where $\mu$ is the Lagrange multiplier. By setting all partial derivatives of $\Lambda$ to 0 we find the critical points of $\Lambda$ and this gives the following algebraic system

$$p_{11} + \lambda p_{22} = q_{11}, \tag{51a}$$

$$\lambda p_{11} + p_{22} = q_{22}, \tag{51b}$$

$$(1 - \lambda)p_{12} = q_\mathrm{S}, \tag{51c}$$

$$p_{11}p_{22} - p_{12}^2 = \frac{f}{g}\det(\boldsymbol{C}), \tag{51d}$$

where $\lambda = \mu/\det(\boldsymbol{C})$. The system (51a)–(51c) is linear in $p_{11}$, $p_{22}$ and $p_{12}$, and is regular if $\lambda \neq \pm 1$ (we discuss the singular cases in the "Appendix A"). In the case when $\lambda \neq \pm 1$,

we calculate the critical points by inverting the system, i.e., we express $p_{11}$, $p_{22}$ and $p_{12}$ in terms of $\lambda$ as

$$p_{11} = \frac{\lambda q_{22} - q_{11}}{\lambda^2 - 1}, \quad p_{22} = \frac{\lambda q_{11} - q_{22}}{\lambda^2 - 1}, \quad p_{12} = \frac{q_S}{1 - \lambda}. \tag{52}$$

Substituting these expressions in Eq. (51d) gives the following quartic equation

$$F(\lambda) = a_4\lambda^4 + a_2\lambda^2 + a_1\lambda + a_0 = 0, \tag{53a}$$

with coefficients given by

$$a_4 = \frac{f}{g}\det(\boldsymbol{C}) \geq 0, \tag{53b}$$

$$a_2 = -2\frac{f}{g}\det(\boldsymbol{C}) - \det(\boldsymbol{Q}_S) = -2a_4 - \det(\boldsymbol{Q}_S), \tag{53c}$$

$$a_1 = \|\boldsymbol{Q}_S\|^2 \geq 0, \tag{53d}$$

$$a_0 = \frac{f}{g}\det(\boldsymbol{C}) - \det(\boldsymbol{Q}_S) = a_4 - \det(\boldsymbol{Q}_S). \tag{53e}$$

Furthermore, from Eqs. (51a)–(51b) the condition (49c) becomes

$$\mathrm{tr}(\boldsymbol{P}) = \frac{\mathrm{tr}(\boldsymbol{Q}_S)}{1 + \lambda} \geq 0, \tag{54}$$

and consequently, we need to select Lagrange multipliers that satisfy the above condition. It can be shown that the quartic Eq. (53) has at least two real roots, and one of them is less than $-1$ and other one is greater than $-1$ (see "Appendix A"). The convexity condition (54) can be satisfied by choosing the appropriate values of $\lambda$ and $\mathrm{tr}(\boldsymbol{Q}_S)$, and the minimizers of $H_S$ are given by (52).

**Minimizing procedure for $\boldsymbol{m}$**

In this section, we describe the minimization step (41c). The minimizing procedure for $\boldsymbol{m}$ differs from the procedure given in [7] because we have an extra matrix $\boldsymbol{C}$ in the function $J_I$ which results in two coupled elliptic equations for the components of the mapping $\boldsymbol{m}$ instead of decoupled Poisson equations. We assume $\boldsymbol{P}$ and $\boldsymbol{b}$ are fixed, and minimize $J(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b})$ over the functions $\boldsymbol{m} \in \mathcal{M}$ using calculus of variations, i.e., $\boldsymbol{P}$ and $\boldsymbol{b}$ are given in all grid points $\mathrm{x}_{ij} \in \mathcal{S}$. We want to compute $\boldsymbol{m}$ on the grid covering $\mathcal{S}$. Here, we drop the indices $n$ and $n+1$, for ease of notation. In the calculations that follow, we use the identity for the Frobenius norm of matrices, i.e.,

$$\|\boldsymbol{A} + \boldsymbol{B}\|^2 = \|\boldsymbol{A}\|^2 + 2\boldsymbol{A} : \boldsymbol{B} + \|\boldsymbol{B}\|^2. \tag{55}$$

The first variation of the functional $J$ with respect to $\boldsymbol{m}$ in the direction $\boldsymbol{\eta} \in [C^2(\mathcal{S})]^2$ is given by

$$\begin{aligned}
\delta J(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b})[\boldsymbol{\eta}] &= \lim_{\epsilon \to 0} \frac{1}{\epsilon}\big[J(\boldsymbol{m} + \epsilon\boldsymbol{\eta}, \boldsymbol{P}, \boldsymbol{b}) - J(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b})\big] \\
&= \lim_{\epsilon \to 0}\left[\frac{\alpha}{2}\iint_{\mathcal{S}} 2(\boldsymbol{C}\mathrm{D}\boldsymbol{m} - \boldsymbol{P}) : \boldsymbol{C}\mathrm{D}\boldsymbol{\eta} + \epsilon\|\boldsymbol{C}\mathrm{D}\boldsymbol{\eta}\|^2 \mathrm{d}\mathrm{x}\right. \\
&\quad \left. + \frac{1-\alpha}{2}\oint_{\partial\mathcal{S}} 2(\boldsymbol{m} - \boldsymbol{b}) \cdot \boldsymbol{\eta} + \epsilon\|\boldsymbol{\eta}\|^2 \mathrm{d}s\right] \\
&= \alpha\iint_{\mathcal{S}} (\boldsymbol{C}\mathrm{D}\boldsymbol{m} - \boldsymbol{P}) : \boldsymbol{C}\mathrm{D}\boldsymbol{\eta}\mathrm{d}\mathrm{x} + (1-\alpha)\oint_{\partial\mathcal{S}} (\boldsymbol{m} - \boldsymbol{b}) \cdot \boldsymbol{\eta}\mathrm{d}s.
\end{aligned} \tag{56}$$

The minimizer is obtained by setting the variation equal to 0, i.e.,

$$\delta J(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b})[\boldsymbol{\eta}] = 0, \quad \forall \boldsymbol{\eta} \in [C^2(\mathcal{S})]^2. \tag{57}$$

Let us define the column vectors $\boldsymbol{p}_1$, $\boldsymbol{p}_2$, $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ as

$$\boldsymbol{P} = [\boldsymbol{p}_1\ \boldsymbol{p}_2],\ \boldsymbol{C} = [\boldsymbol{c}_1\ \boldsymbol{c}_2],\ \boldsymbol{p}_i = \begin{pmatrix} p_{1i} \\ p_{2i} \end{pmatrix},\ \boldsymbol{c}_i = \begin{pmatrix} c_{1i} \\ c_{2i} \end{pmatrix}, \quad i = 1, 2. \tag{58}$$

We can split the first integrand of the final expression in (56) as follows

$$\begin{aligned} (\boldsymbol{C}\mathrm{D}\boldsymbol{m} - \boldsymbol{P}) : \boldsymbol{C}\mathrm{D}\boldsymbol{\eta} &= \boldsymbol{C}^T(\boldsymbol{C}\mathrm{D}\boldsymbol{m} - \boldsymbol{P}) : \mathrm{D}\boldsymbol{\eta} \\ &= \sum_{k=1}^{2} \boldsymbol{C}^T\Big(\boldsymbol{C}\frac{\partial \boldsymbol{m}}{\partial x_k} - \boldsymbol{p}_k\Big) \cdot \frac{\partial \boldsymbol{\eta}}{\partial x_k} \\ &= \boldsymbol{v}_1 \cdot \frac{\partial \boldsymbol{\eta}}{\partial x_1} + \boldsymbol{v}_2 \cdot \frac{\partial \boldsymbol{\eta}}{\partial x_2}, \end{aligned} \tag{59}$$

where the vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are column vectors of the matrix $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2]$, given by

$$\boldsymbol{v}_1 = \begin{pmatrix} v_{11} \\ v_{21} \end{pmatrix} = \boldsymbol{C}^T\Big(\boldsymbol{C}\frac{\partial \boldsymbol{m}}{\partial x_1} - \boldsymbol{p}_1\Big), \quad \boldsymbol{v}_2 = \begin{pmatrix} v_{12} \\ v_{22} \end{pmatrix} = \boldsymbol{C}^T\Big(\boldsymbol{C}\frac{\partial \boldsymbol{m}}{\partial x_2} - \boldsymbol{p}_2\Big),$$

and by defining $\boldsymbol{W} = \boldsymbol{V}^T = [\boldsymbol{w}_1, \boldsymbol{w}_2]$, we can rewrite the first integral of the final expression in (56) as

$$\iint_{\mathcal{S}} (\boldsymbol{C}\mathrm{D}\boldsymbol{m} - \boldsymbol{P}) : \boldsymbol{C}\mathrm{D}\boldsymbol{\eta}\mathrm{d}\mathrm{x} = \sum_{k=1}^{2} \iint_{\mathcal{S}} \boldsymbol{w}_k \cdot \nabla\eta_k \mathrm{d}\mathrm{x}. \tag{60}$$

Let $\hat{\boldsymbol{n}}$ denote the unit outward normal at the boundary $\partial\mathcal{S}$. Using the vector-scalar product rule [33, p. 576] and the identity

$$\iint_{\mathcal{S}} \nabla v \cdot \boldsymbol{F} + v\nabla \cdot \boldsymbol{F}\mathrm{d}\mathrm{x} = \oint_{\partial\mathcal{S}} v\boldsymbol{F} \cdot \hat{\boldsymbol{n}}\mathrm{d}s, \tag{61}$$

derived from the Gauss's theorem, the integrals in the rhs of (56) become

$$\iint_{\mathcal{S}} \boldsymbol{w}_k \cdot \nabla\eta_k \mathrm{d}\mathrm{x} = \oint_{\partial\mathcal{S}} \boldsymbol{w}_k \cdot \hat{\boldsymbol{n}}\eta_k \mathrm{d}s - \iint_{\mathcal{S}} \nabla \cdot \boldsymbol{w}_k\eta_k \mathrm{d}\mathrm{x}. \tag{62}$$

Substituting the integral in the final expression of (56), the minimizer can be obtained from the following relation

$$\sum_{k=1}^{2} \Big( \oint_{\partial\mathcal{S}} \big(\alpha \boldsymbol{w}_k \cdot \hat{\boldsymbol{n}} + (1-\alpha)(m_k - b_k)\big)\eta_k \mathrm{d}s - \alpha \iint_{\mathcal{S}} \nabla \cdot \boldsymbol{w}_k\eta_k \mathrm{d}\mathrm{x}\Big) = 0 \quad \forall \boldsymbol{\eta} \in [C^2(\mathcal{S})]^2, \tag{63}$$

where $m_k$ and $b_k$ ($k = 1, 2$) are the components of the vectors $\boldsymbol{m}$ and $\boldsymbol{b}$, respectively. Choosing $\eta_2 = 0$ and applying the fundamental lemma of calculus of variations [34, p. 15] for $\eta_1$, we find that $m_1$ and $m_2$ satisfies, almost everywhere, the equation

$$\begin{aligned} &\frac{\partial}{\partial x_1}\left[|\boldsymbol{c}_1|^2 \frac{\partial m_1}{\partial x_1} + \boldsymbol{c}_1 \cdot \boldsymbol{c}_2 \frac{\partial m_2}{\partial x_1}\right] + \frac{\partial}{\partial x_2}\left[|\boldsymbol{c}_1|^2 \frac{\partial m_1}{\partial x_2} + \boldsymbol{c}_1 \cdot \boldsymbol{c}_2 \frac{\partial m_2}{\partial x_2}\right] \\ &\quad = \frac{\partial}{\partial x_1}(\boldsymbol{c}_1 \cdot \boldsymbol{p}_1) + \frac{\partial}{\partial x_2}(\boldsymbol{c}_1 \cdot \boldsymbol{p}_2) \quad \mathrm{x} \in \mathcal{S}, \end{aligned} \tag{64a}$$

$$(1-\alpha)m_1+\alpha(|\boldsymbol{c}_1|^2\nabla m_1\cdot\hat{\boldsymbol{n}}+\boldsymbol{c}_1\cdot\boldsymbol{c}_2\nabla m_2\cdot\hat{\boldsymbol{n}})=(1-\alpha)b_1+\alpha\boldsymbol{c}_1\cdot\boldsymbol{P}\hat{\boldsymbol{n}}\quad \mathrm{x}\in\partial\mathcal{S}. \tag{64b}$$

Similarly, choosing $\eta_1=0$ and applying the fundamental lemma of calculus of variations for $\eta_2$, we obtain

$$\begin{aligned}&\frac{\partial}{\partial x_1}\left[\boldsymbol{c}_1\cdot\boldsymbol{c}_2\frac{\partial m_1}{\partial x_1}+|\boldsymbol{c}_2|^2\frac{\partial m_2}{\partial x_1}\right]+\frac{\partial}{\partial x_2}\left[\boldsymbol{c}_1\cdot\boldsymbol{c}_2\frac{\partial m_1}{\partial x_2}+|\boldsymbol{c}_2|^2\frac{\partial m_2}{\partial x_2}\right]\\&\quad=\frac{\partial}{\partial x_1}(\boldsymbol{c}_2\cdot\boldsymbol{p}_1)+\frac{\partial}{\partial x_2}(\boldsymbol{c}_2\cdot\boldsymbol{p}_2)\quad \mathrm{x}\in\mathcal{S},\end{aligned} \tag{65a}$$

$$(1-\alpha)m_2+\alpha(\boldsymbol{c}_1\cdot\boldsymbol{c}_2\nabla m_1\cdot\hat{\boldsymbol{n}}+|\boldsymbol{c}_2|^2\nabla m_2\cdot\hat{\boldsymbol{n}})=(1-\alpha)b_2+\alpha\boldsymbol{c}_2\cdot\boldsymbol{P}\hat{\boldsymbol{n}}\quad \mathrm{x}\in\partial\mathcal{S}. \tag{65b}$$

We can rewrite these equations as follows in matrix-vector form

$$\nabla\cdot(\boldsymbol{C}^T\boldsymbol{C}\mathrm{D}\boldsymbol{m})=\nabla\cdot(\boldsymbol{C}^T\boldsymbol{P}),\qquad \mathrm{x}\in\mathcal{S}, \tag{66a}$$

$$(1-\alpha)\boldsymbol{m}+\alpha(\boldsymbol{C}^T\boldsymbol{C}\nabla\boldsymbol{m})\cdot\hat{\boldsymbol{n}}=(1-\alpha)\boldsymbol{b}+\alpha\boldsymbol{C}\cdot\boldsymbol{P}\hat{\boldsymbol{n}},\quad \mathrm{x}\in\partial\mathcal{S}. \tag{66b}$$

These are two coupled elliptic equations with Robin boundary conditions for the two components $m_1$ and $m_2$ of the mapping $\boldsymbol{m}$ [35, p. 160]. The above equations are in divergence form which motivates us to discretize the equations using the finite volume method [35, p. 84–88], for more details see "Appendix B".

### 3.2 Second Stage: Calculation of the Freeform Surfaces

We compute the lens surfaces assuming that a numerical approximation of $\boldsymbol{m}$ on the grid is available. We compute the first lens surface $u_1(\mathrm{x})$ from the converged mapping $\boldsymbol{m}$ using relation (26) in the least-squares senses, i.e.,

$$u_1(\mathrm{x})=\underset{\phi}{\operatorname{argmin}}\, I(\phi),\quad I(\phi)=\frac{1}{2}\iint_{\mathcal{S}}|\nabla\phi(\mathrm{x})-\nabla_{\mathrm{x}}c(\mathrm{x},\boldsymbol{m}(\mathrm{x}))|^2\mathrm{d}\mathrm{x},\quad \forall\,\phi\in\mathrm{C}^1(\mathcal{S}). \tag{67}$$

We calculate the minimizing function $u_1(\mathrm{x})$ using calculus of variations. The first variation of the functional $I$ in (67) in a direction $v$ is given by

$$\begin{aligned}\delta I(u_1)[v]&=\lim_{\epsilon\to0}\frac{1}{\epsilon}\big[I(u_1+\epsilon v)-I(u_1)\big]\\&=\lim_{\epsilon\to0}\frac{1}{2}\left[\iint_{\mathcal{S}}\epsilon|\nabla v|^2+2(\nabla u_1-\nabla_{\mathrm{x}}c)\cdot\nabla v\mathrm{d}\mathrm{x}\right]\\&=\iint_{\mathcal{S}}(\nabla u_1-\nabla_{\mathrm{x}}c)\cdot\nabla v\mathrm{d}\mathrm{x}.\end{aligned} \tag{68}$$

The minimizer is given by

$$\delta I(u_1)[v]=0,\quad \forall\, v\in C^1(\mathcal{S}). \tag{69}$$

Using the Gauss's identity (61), we conclude from (69) that

$$\oint_{\partial\mathcal{S}}v(\nabla u_1-\nabla_{\mathrm{x}}c)\cdot\hat{\boldsymbol{n}}\mathrm{d}s-\iint_{\mathcal{S}}v(\Delta u_1-\nabla\cdot\nabla_{\mathrm{x}}c)\mathrm{d}\mathrm{x}=0,\quad \forall\, v\in C^1(\mathcal{S}). \tag{70}$$

Applying the fundamental lemma of calculus of variations [34, p. 15], we find

$$\Delta u_1 = \nabla \cdot \nabla_{\mathrm{x}} c(\mathrm{x}, \boldsymbol{m}), \quad \mathrm{x} \in \mathcal{S}, \tag{71a}$$

$$\nabla u_1 \cdot \hat{\boldsymbol{n}} = \nabla_{\mathrm{x}} c \cdot \hat{\boldsymbol{n}}, \quad \mathrm{x} \in \partial\mathcal{S}. \tag{71b}$$

This is a Neumann problem, and only has a solution if the compatibility condition is satisfied, which reads

$$\iint_{\mathcal{S}} \nabla \cdot \nabla_{\mathrm{x}} c \mathrm{d}\mathrm{x} - \oint_{\partial\mathcal{S}} \nabla_{\mathrm{x}} c \cdot \hat{\boldsymbol{n}} \mathrm{d}s = 0. \tag{72}$$

By Gauss's theorem, this is satisfied automatically. The solution of the Poisson equation with Neumann boundary condition is unique up to an additive constant. To make the solution unique, we have added the constraint $u_1(x_1, x_2) = 1$ at the first discretized left most corner point. We solve this problem using standard finite differences, and the discretized system is solved in Matlab using LU decomposition. The second lens surface is calculated from the relation (19), by substituting the converged mapping $\boldsymbol{m}(\mathrm{x})$ and the first lens surface $u_1(\mathrm{x})$, thus we have

$$u_2(\boldsymbol{m}(\mathrm{x})) = c(\mathrm{x}, \boldsymbol{m}(\mathrm{x})) - u_1(\mathrm{x}) \quad \forall \mathrm{x} \in \mathcal{S}. \tag{73}$$

The numerical algorithm is summarized as follows. We start the minimization procedure using the initial guess $\boldsymbol{m}^0$ given by (42) for the discretized source domain $\mathcal{S}$. Subsequently, we iteratively perform the steps given by (41a), (41b) and (41c). The first and second steps are minimization procedure for $\boldsymbol{b}$ and $\boldsymbol{P}$, respectively, and both are performed pointwise. The third step is a minimization procedure for the mapping $\boldsymbol{m}$ and is performed by solving two coupled elliptic boundary value problems given by (64) and (65). Next, we update the matrix $\boldsymbol{C}$ given by (27). Finally, after convergence of the iteration (41), the first lens surface is computed from the mapping $\boldsymbol{m}$ by solving Poisson problem (71), and the second lens surface is computed from relation (73).

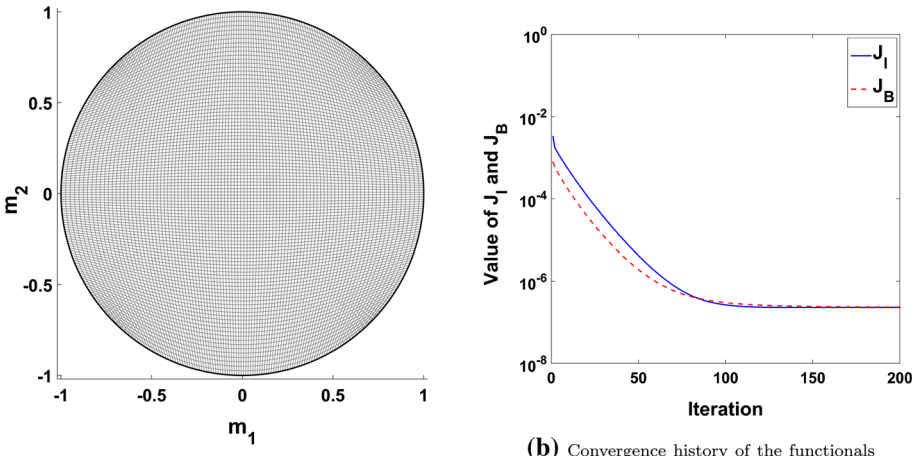
## 4 Numerical Results

We apply the algorithm to four test problems to compute c-convex lens surfaces: first, we map a square with uniform emittance into a circle with uniform illuminance, second, we map an ellipse with uniform emittance into another ellipse with uniform illuminance, third, we map a square with uniform emittance into a non-convex (flower) target distribution, and finally, we challenge our algorithm to map the same distribution into a light pattern given by a picture on the target screen. The numerical results are verified by our self-built ray tracer based on the Monte–Carlo method [2].

### 4.1 From a Square to a Circle

In the first test problem, we design an optical system of lens surfaces that transforms the uniform emittance of a square into a circle with uniform illuminance. The source domain is given by $\mathcal{S} = [-1, 1] \times [-1, 1]$ and the target domain by $\mathcal{T} = \{\mathrm{y} \in \mathbb{R}^2 \mid \ ||\mathrm{y}||_2 \leq 1\}$. The light source emits a parallel beam of light rays with uniform emittance, i.e.,

$$f(\mathrm{x}) = \frac{1}{4} \quad \forall\, \mathrm{x} \in \mathcal{S}. \tag{74}$$




**(a)** The mapping after 200 iterations.

**(b)** Convergence history of the functionals $J_{\mathrm{I}}$ and $J_{\mathrm{B}}$.

**Fig. 2** Square-to-circle problem: the mapping and convergence history

The target plane is at a distance $\ell = 40$ from the source plane, and we have fixed the refractive index $n = 1.5$ and the reduced optical path length $\beta = 3\pi$ for all numerical problems. The target $\mathcal{T}$ is illuminated by a parallel beam of light rays with uniform illuminance, i.e.,

$$g(\mathrm{y}) = \begin{cases} \frac{1}{\pi} & \text{if } \mathrm{y} \in \mathcal{T}, \\ 0 & \text{otherwise.} \end{cases} \tag{75}$$

Note that the energy conservation condition (20) is satisfied. We discretize the source domain $\mathcal{S}$ uniformly with $200 \times 200$ grid points. We have a different grid for the boundary, and found from various experiments that the number of boundary grid points $N_{\mathrm{b}}$ does not influence the convergence of the algorithm if it is chosen large enough. Since a large value of $N_{\mathrm{b}}$ does not significantly increase the calculation time, we have chosen $N_{\mathrm{b}} = 1000$. We also observed from various experiments that $\alpha = 0.65$ is a good choice for $\alpha$ to have residuals in $J_{\mathrm{I}}$ and $J_{\mathrm{B}}$ close together. Choosing $\alpha$ too large or too small slows down convergence. We stopped the algorithm after 200 iterations because $J_{\mathrm{I}}$ and $J_{\mathrm{B}}$ stall. The resulting mapping after 200 iterations is shown in Fig. 2a, and the convergence history of the algorithm is shown in Fig. 2b. The algorithm performed efficiently, the boundary and interior functionals for the circular target have converged well with residuals of approximately $2.35 \times 10^{-7}$.

### 4.2 From an Ellipse to Another Ellipse

In the second test case, we apply the algorithm to map a uniform emittance of an ellipse into another ellipse with uniform illuminance. The source domain is given by $\mathcal{S} = \{(x_1, x_2) \in \mathbb{R}^2 \mid \ 4x_1^2 + x_2^2 \leq 4\}$, see Fig. 3a, and the target domain by $\mathcal{T} = \{(y_1, y_2) \in \mathbb{R}^2 \mid \ y_1^2 + 4y_2^2 \leq 4\}$, i.e., rotate an ellipse distribution to another ellipse over $\pi/2$. We have $f(x_1, x_2) = g(y_1, y_2) = 1/4\pi$. We use a $200 \times 200$ grid with 1000 points on the boundary, the reduced optical path length $\beta = 3\pi$, and the weight parameter $\alpha = 0.65$. The mapping after 200 iterations is shown in Fig. 3b, and the source distribution in grid format shown in Fig. 3a. The algorithm exhibits almost the same convergence as shown in Fig. 2b.

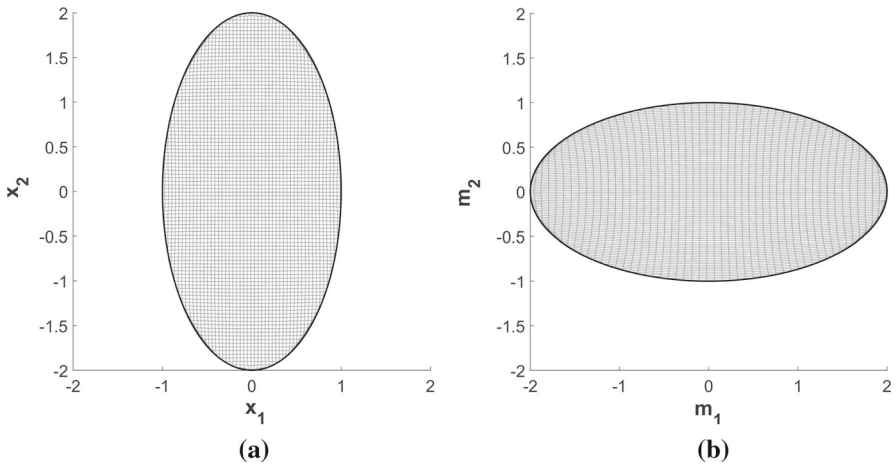



**Fig. 3** **a** An ellipse on the source $\mathcal{S}$ and **b** its image under the mapping $\boldsymbol{m}$ on the target $\mathcal{T}$

### 4.3 From a Square to a Non-convex (Flower) Target

In the third test case, we test the algorithm for non-convex flower shaped targets. We apply the algorithm to map a uniform emittance of a square into uniformly illuminated non-convex targets. The source domain is given by the square $[-1, 1] \times [-1, 1]$ with $f(x_1, x_2) = 1/4$, and the target domain is defined in polar coordinates as

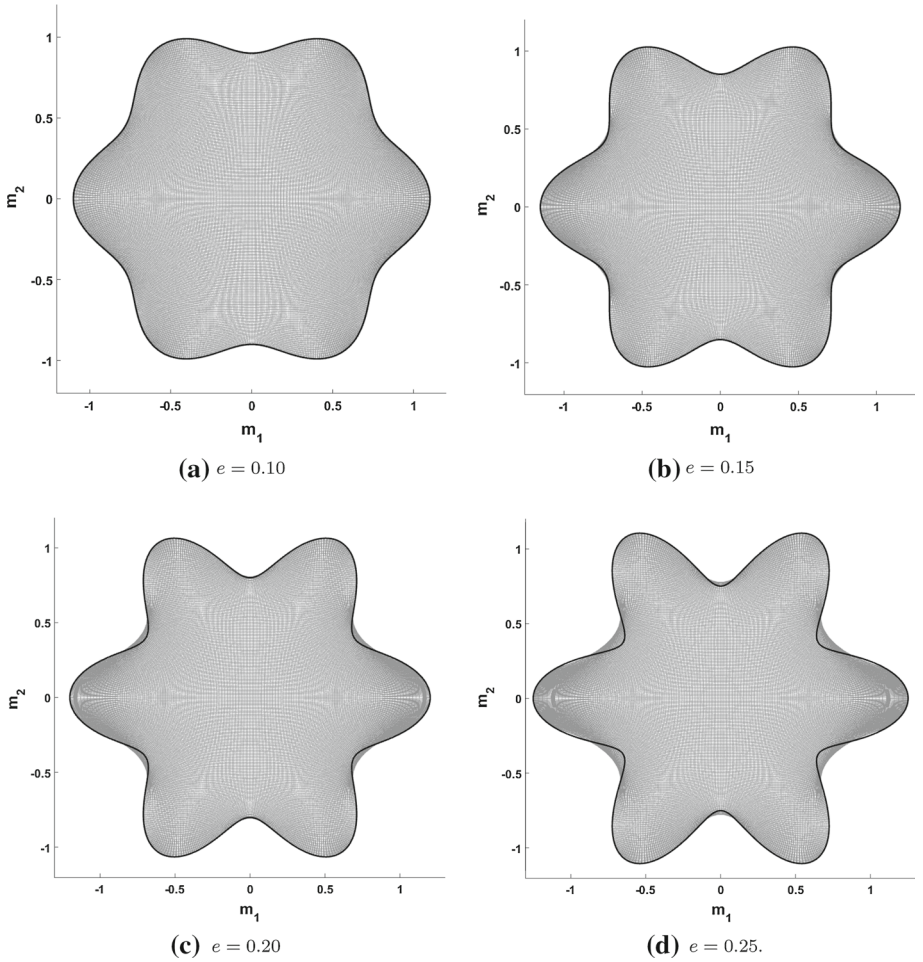
$$\rho(\theta) = 1 + e\cos(6\theta), \quad 0 \leq \theta < 2\pi, \tag{76}$$

where $\rho(\theta)$ is the distance to the origin and $\theta$ is the counterclockwise angle with respect to the $y_1$-axis in the target plane. We test the algorithm for the four values $e \in \{0.1, 0.15, 0.20, 0.25\}$ which represent the deviation of the target domain from a convex shape. We use $200 \times 200$ grid with 1000 points on the boundary, the reduced optical path length $\beta = 3\pi$, and the weight parameter $\alpha = 0.50$. The mappings after 250 iterations are shown in Fig. 4. The residual $J$ after 250 iterations is $2.73 \times 10^{-7}$, $7.05 \times 10^{-6}$, $3.93 \times 10^{-5}$ and $9.99 \times 10^{-5}$, respectively. The convergence problem arises for target domains which strongly deviate from a convex shape, but if the shape deviates mildly from convex, the algorithm performs satisfactorily, see Fig. 4a–d.

### 4.4 From a Square to a Picture

The fourth test problem is to design an optical system of lens surfaces which transforms a square uniform bundle of parallel light rays into a target distribution corresponding to a given picture. Here, we challenge our algorithm for a picture showing costumes of the Indian classical dance Bharatanatyam, see Fig. 5. The emittance of the light source is again the same as defined in (74) and the parameters of the optical system are also the same as defined in Sect. 4.1. The desired target illumination $g(y_1, y_2)$ is given by the grayscale test image shown in Fig. 5. Because the target distribution contains many details, e.g., the pattern of costumes and jewellery, it provides a challenging test for our algorithm.

Note that the picture is converted into grayscale and contains some black regions, which results in $g(y_1, y_2) = 0$ for some $(y_1, y_2) \in \mathcal{T}$. This would give division by 0 in the least-




**(a)** $e = 0.10$ **(b)** $e = 0.15$

**(c)** $e = 0.20$ **(d)** $e = 0.25$.

**Fig. 4** *Non-convex* flower shaped target problem: mapping after 200 iterations

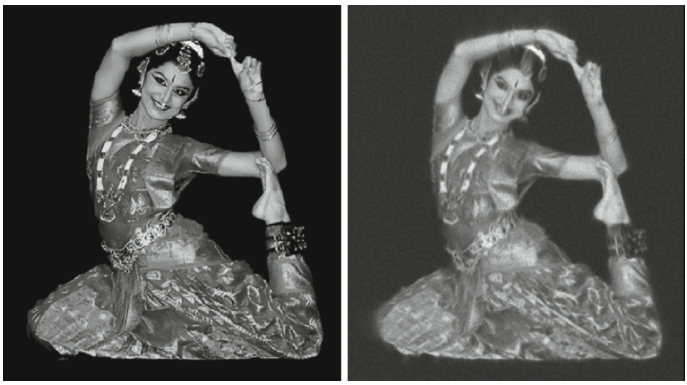
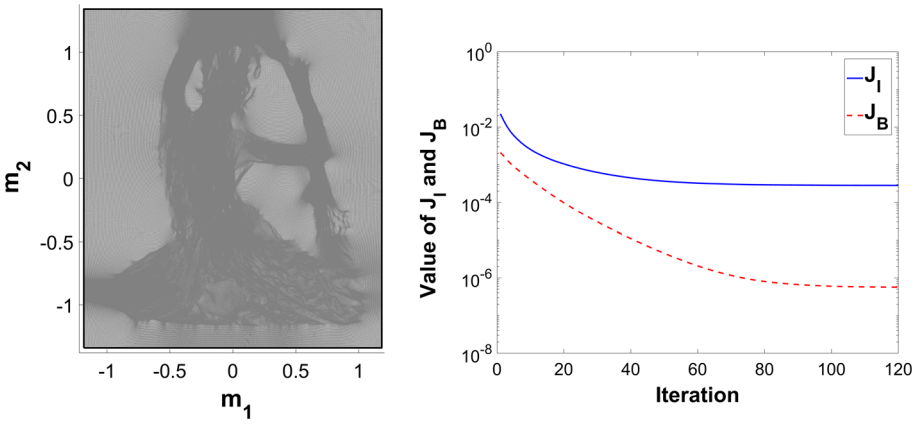


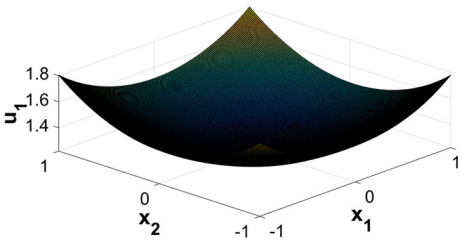
**Fig. 5** The picture showing costumes of Indian classical dance Bharatanatyam. The original is shown on the left and the ray trace result is shown on the right. Courtesy Wikipedia



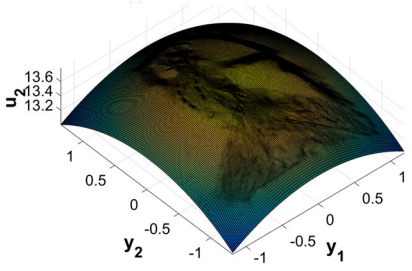

**(a)** The converged mapping. **(b)** Convergence history of the mapping.

**Fig. 6** The mapping and its convergence history for the image target

**Fig. 7** Freeform surfaces of the lens for the image target




**(a)** The first surface $\mathcal{L}_1$.




**(b)** The second surface $\mathcal{L}_2$.

squares algorithm. Therefore, we have increased the illuminance to 5% of the maximum value if it is less than this threshold value. We discretized the source $\mathcal{S}$ on a $500 \times 500$ grid, with 1000 boundary points. The convergence history of the algorithm is shown in Fig. 6b for $\alpha = 0.70$. We stopped the algorithm after 150 iterations, because $J_{\mathrm{I}}$ and $J_{\mathrm{B}}$ did no longer seem to decrease. The resulting mapping is shown in Fig. 6a, the image details can be recognized in the grid. The optical system is verified using the ray tracing algorithm [2]. We ran our ray tracing algorithm for 10 million uniformly distributed random points on the source to compute the actual illumination pattern produced on the target. The target illuminance for 10 million rays is plotted in the Fig. 5. The output images is very close to the corresponding original image, although the image is slightly blurred, but even complex

details can be identified. The functions $u_1(\mathrm{x})$ and $u_2(\mathrm{y})$ representing the freeform surfaces $\mathcal{L}_1$ and $\mathcal{L}_2$ of the lens, respectively, are shown in Fig. 7. The lens surfaces are convex on their respective domains and an alternative representation of the mapping can be seen as contour of grids on the second lens surface.

## 5 Discussion and Conclusion

We introduced a least-squares method to compute freeform surfaces of an optical system corresponding to a non-quadratic cost function. The method is an extended version of the least-squares method, earlier introduced in [7]. Furthermore, we presented a new generic (in term of cost function) minimization procedure of $\boldsymbol{P}$ for the functional $J_{\mathrm{I}}$. Moreover, we have shown that the minimization procedure of the mapping $\boldsymbol{m}$ for the total functional $J$ consists of coupled elliptic PDEs.

We presented the extended least-squares method to compute coupled freeform surfaces of a lens. Our method can compute freeform surfaces of any optical system corresponding to a twice continuously differential cost function, which demonstrates the wider applicability of the method. The ELS-method also shows good performance for a non-convex target domain: as long as the domain does not deviate too much from a convex shape.

The algorithm is very time and memory efficient, and provides both convex and concave optical surfaces which makes it very suitable to use for these type of problems. Furthermore, we have applied the method to a very challenging problem containing the details of the costumes of the Indian classical dance Bharatanatyam and obtained a high resolution, preserving details of the original picture.

In future work we would like to apply the algorithm to more complex cost functions, e.g., point light sources and far-field problems. Also, we would to explore the applicability of the Monge–Ampère solver in other fields of science and engineering.



## Appendices

## A Solution of the Quartic Equation

We obtain four possible solutions of the quartic equation (53) using Ferrari's method [36, p. 32]. The key idea is to rewrite the quartic equation as two quadratic equations, and by solving both we get solutions of the quartic equation. For detailed solution of the quadratic equations we refer to [7].

**Solution of (53) when $f > 0$.**
It can be shown that the problem (53) has at least two real roots. For the real symmetric matrix $\boldsymbol{Q}_{\mathrm{S}}$, we can deduce

$$\mathrm{tr}(\boldsymbol{Q}_{\mathrm{S}})^2 - 4\det(\boldsymbol{Q}_{\mathrm{S}}) = (q_{11} - q_{22})^2 + 4q_{\mathrm{S}}^2 \geq 0, \tag{77}$$

and using the above relation, we conclude that $F(-1) = -\operatorname{tr}(\boldsymbol{Q}_S)^2 < 0$ and $F(1) = \operatorname{tr}(\boldsymbol{Q}_S)^2 - 4\det(\boldsymbol{Q}_S) \geq 0$, and the coefficient of $\lambda^4$ in the quartic equation (53) is positive. Which imply that (53) has at least two real roots, more precisely one of them is less than $-1$ and other one is greater than $-1$. The solution of (53) are given by

$$\lambda_1 = -\sqrt{\frac{y}{2}} + \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} + \frac{a_1}{2a_4\sqrt{2y}}}, \tag{78a}$$

$$\lambda_2 = -\sqrt{\frac{y}{2}} - \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} + \frac{a_1}{2a_4\sqrt{2y}}}, \tag{78b}$$

$$\lambda_3 = \sqrt{\frac{y}{2}} + \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} - \frac{a_1}{2a_4\sqrt{2y}}}, \tag{78c}$$

$$\lambda_4 = \sqrt{\frac{y}{2}} - \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} - \frac{a_1}{2a_4\sqrt{2y}}}, \tag{78d}$$

where $y$ is the solution of a cubic equation in the Ferrari's method. The real roots satisfying the convexity condition (54) are substituted in (52) and (49), yielding the possible minimizers of $H_S(p_{11}, p_{22}, p_{12})$. Note that we have division by zero in (78) if $y = 0$. We find that this happens only when $a_1 = 0$, i.e., $q_{11} = q_{22} = q_S = 0$. This is a special case which corresponds to the possibility $\lambda = 1$, which we discuss later.

**Solution of (53) when $f = 0$.**
If the source density $f = 0$, the quartic equation (53) reduced to a quadratic equation because $a_4 = 0$. The solutions are obtained by solving the corresponding quadratic equation, and the roots are given by

$$\lambda = \frac{-a_1 \pm \sqrt{a_1^2 - 4a_2a_0}}{2a_2}. \tag{79}$$

We can verify that the discriminant of this quadratic equation is always positive by substituting the coefficients in (53) in the discriminant, it becomes

$$a_1^2 - 4a_2a_0 = (q_{11}^2 - q_{22})^2 + 4q_S^2(q_{11}^2 + q_{22}^2)^2 \geq 0. \tag{80}$$

Furthermore, also in this case $F(-1) = -\operatorname{tr}(\boldsymbol{Q}_S)^2 < 0$ and $F(1) = \operatorname{tr}(\boldsymbol{Q}_S)^2 - 4\det(\boldsymbol{Q}_S) \geq 0$, and consequently it shows that (53) has at least one solution $\lambda > -1$.

**Solution of (53) when $\lambda = 1$.**
If $\lambda = 1$, i.e., $q_{11} = q_{22}$ and $q_S = 0$, and in this case, we cannot invert the system (51a)–(51c) for $(p_{11}, p_{22}, p_{12})$. Therefore we determine the minimum of $H_S(p_{11}, p_{22}, p_{12})$ as follows. Using $q_S = 0$ and $q_{11} = q_{22}$, the minimization (49) simplifies to

$$\underset{(p_{11},p_{22},p_{12})\in\mathbb{R}^3}{\operatorname{argmin}} \frac{1}{2}\big((p_{11} - q_{11})^2 + 2p_{12}^2 + (p_{22} - q_{11})^2\big), \tag{81}$$

with the conditions (49b) and (49c). By solving this minimization problem we obtained the following four solutions

$$p_{11} = p_{22} = \frac{q_{11}}{2}, \quad p_{12} = \pm\sqrt{\frac{q_{11}^2}{4} - \frac{f}{g}\det(\boldsymbol{C})}, \tag{82a}$$

or

$$p_{11} = p_{22} = \pm\sqrt{\frac{f}{g}\det(\boldsymbol{C})}, \quad p_{12} = 0. \tag{82b}$$

For the detailed solution see [7]. Also there exists at least one solution which satisfies the convexity condition $\mathrm{tr}(\boldsymbol{P}) \geq 0$.

**Solution of (53) when $\boldsymbol{\lambda = -1}$.**

If $\lambda = -1$, i.e., $q_{11} = -q_{22}$. In this case again we can not invert the system (51a)–(51c) for $(p_{11}, p_{22}, p_{12})$. We determine the minimizers using a different method. From (51a) and (51b), we find

$$p_{22} = p_{11} - q_{11}, \quad p_{12} = q_S/2. \tag{83}$$

Substituting these in (51d), we conclude

$$p_{11}^2 - q_{11}p_{11} - \frac{q_S^2}{4} - \frac{f}{g}\det(\boldsymbol{C}) = 0. \tag{84}$$

Solving for $p_{11}$, we find two solutions:

$$p_{11} = \frac{q_{11}}{2} + \frac{\sqrt{q_{11}^2 + q_S^2 + 4\det(\boldsymbol{C})f/g}}{2}, \tag{85a}$$

$$p_{12} = \frac{q_S}{2}, \tag{85b}$$

$$p_{22} = -\frac{q_{11}}{2} + \frac{\sqrt{q_{11}^2 + q_S^2 + 4\det(\boldsymbol{C})f/g}}{2}, \tag{85c}$$

and

$$p_{11} = \frac{q_{11}}{2} - \frac{\sqrt{q_{11}^2 + q_S^2 + 4\det(\boldsymbol{C})f/g}}{2}, \tag{86a}$$

$$p_{12} = \frac{q_S}{2}, \tag{86b}$$

$$p_{22} = -\frac{q_{11}}{2} - \frac{\sqrt{q_{11}^2 + q_S^2 + 4\det(\boldsymbol{C})f/g}}{2}, \tag{86c}$$

which are always real. The second solution satisfies the convexity condition $\mathrm{tr}(\boldsymbol{P}) \geq 0$.

## B Finite Volume Discretisation of the Coupled Elliptic Equations (66)

We can write the differential equation (64a) as

$$\frac{\partial f_{11}}{\partial x_1} + \frac{\partial f_{12}}{\partial x_2} = \frac{\partial r_{11}}{\partial x_1} + \frac{\partial r_{12}}{\partial x_2}, \tag{87}$$

where

$$f_{11} = |\boldsymbol{c}_1|^2 \frac{\partial m_1}{\partial x_1} + \boldsymbol{c}_1 \cdot \boldsymbol{c}_2 \frac{\partial m_2}{\partial x_1}, \quad r_{11} = \boldsymbol{c}_1 \cdot \boldsymbol{p}_1,$$

$$f_{12} = |\boldsymbol{c}_1|^2 \frac{\partial m_1}{\partial x_2} + \boldsymbol{c}_1 \cdot \boldsymbol{c}_2 \frac{\partial m_2}{\partial x_2}, \quad r_{12} = \boldsymbol{c}_1 \cdot \boldsymbol{p}_2.$$

The above equation can be written in the divergence form as

$$\nabla \cdot \boldsymbol{f}_1 = \nabla \cdot \boldsymbol{r}_1, \tag{88}$$

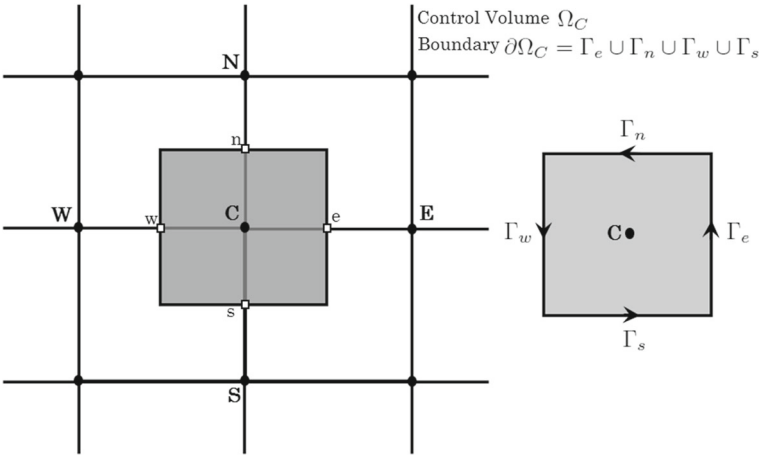



**Fig. 8** The control volume for a cell-centred finite volume method

where, $\boldsymbol{f}_1 = (f_{11}, f_{12})^T$ and $\boldsymbol{r}_1 = (r_{11}, r_{12})^T$. Integrating Eq. (88) over each $\mathcal{A} \subset \mathcal{S}$ and using Gauss's theorem [33, p. 925], we obtain

$$\oint_{\partial\mathcal{A}} \boldsymbol{f}_1 \cdot \hat{\boldsymbol{n}} \mathrm{d}s = \oint_{\partial\mathcal{A}} \boldsymbol{r}_1 \cdot \hat{\boldsymbol{n}} \mathrm{d}s, \tag{89}$$

where $\hat{\boldsymbol{n}}$ is the unit outward normal on the boundary $\partial\mathcal{A}$ of $\mathcal{A}$. Now, we create a set of non-overlapping control volumes for the computational grid of the domain $\mathcal{S}$ and apply the cell-centred finite volume method, i.e., grid points are located in the centre of the control volume.

Let us consider the control volume $\mathcal{A} \equiv \Omega_C = [x_{1,w}, x_{1,e}] \times [x_{2,s}, x_{2,n}]$ as shown in Fig. 8, where $x_{1,w}$ is the $x_1$-value at centre of the western cell face $\Gamma_w$, i.e., $x_{1,w} = x_{1,i-1/2}$ and approximated as $x_{1,w} = \big(x_1(W) + x_1(C)\big)/2$, etc, and $x_1(C) = x_{1,i}$, $x_1(W) = x_{1,i-1}$, etc. The finite volume method is used to transform equation (89) to a system of discrete equations for the centre point C of the control volume $\Omega_C$. First, Eq. (89) is applied for the control volume $\Omega_C$. This reduces the equation to one involving only first derivatives. These first order derivatives are replaced with central difference approximations, for more details see [35]. Finally, the integral equation (89) can be descretized as follows

$$\begin{aligned} &a_E m_1(E) + a_W m_1(W) + a_N m_1(N) + a_S m_1(S) + a_C m_1(C) \\ &\quad + b_E m_2(E) + b_W m_2(W) + b_N m_2(N) + b_S m_2(S) + b_C m_2(C) = r_1(C), \end{aligned} \tag{90}$$

where

$$\begin{aligned} a_E &= \frac{|\boldsymbol{c}_1|_e^2}{h_1^2}, \quad a_W = \frac{|\boldsymbol{c}_1|_w^2}{h_1^2}, \quad a_N = \frac{|\boldsymbol{c}_1|_n^2}{h_2^2}, \quad a_S = \frac{|\boldsymbol{c}_1|_s^2}{h_2^2}, \\ b_E &= \frac{(\boldsymbol{c}_1 \cdot \boldsymbol{c}_2)_e^2}{h_1^2}, \quad b_W = \frac{(\boldsymbol{c}_1 \cdot \boldsymbol{c}_2)_w^2}{h_1^2}, \quad b_N = \frac{(\boldsymbol{c}_1 \cdot \boldsymbol{c}_2)_n^2}{h_2^2}, \quad b_S = \frac{(\boldsymbol{c}_1 \cdot \boldsymbol{c}_2)_s^2}{h_2^2}, \\ a_C &= -(a_E + a_W + a_N + a_S), \quad b_C = -(b_E + b_W + b_N + b_S), \\ r_1(C) &= \frac{1}{h_1}\big[(\boldsymbol{c}_1 \cdot \boldsymbol{p}_1)_e - (\boldsymbol{c}_1 \cdot \boldsymbol{p}_1)_w\big] + \frac{1}{h_2}\big[(\boldsymbol{c}_1 \cdot \boldsymbol{p}_2)_n - (\boldsymbol{c}_1 \cdot \boldsymbol{p}_2)_s\big]. \end{aligned}$$

Similarly, the discrete form of Eq. (65a) is

$$\begin{aligned} &b_E m_1(E) + b_W m_1(W) + b_N m_1(N) + b_S m_1(S) + b_C m_1(C) \\ &\quad + d_E m_2(E) + d_W m_2(W) + d_N m_2(N) + d_S m_2(S) + d_C m_2(C) = r_2(C), \end{aligned} \tag{91}$$

where

$$\begin{aligned} d_E &= \frac{|\boldsymbol{c}_2|_e^2}{h_1^2},\ d_W = \frac{|\boldsymbol{c}_2|_w^2}{h_1^2},\ d_N = \frac{|\boldsymbol{c}_2|_n^2}{h_2^2},\ d_S = \frac{|\boldsymbol{c}_2|_s^2}{h_2^2}, \\ d_C &= -\left(d_E + d_W + d_N + d_S\right), \\ r_2(C) &= \frac{1}{h_1}\left[(\boldsymbol{c}_2 \cdot \boldsymbol{p}_1)_e - (\boldsymbol{c}_2 \cdot \boldsymbol{p}_1)_w\right] + \frac{1}{h_2}\left[(\boldsymbol{c}_2 \cdot \boldsymbol{p}_2)_n - (\boldsymbol{c}_2 \cdot \boldsymbol{p}_2)_s\right]. \end{aligned}$$

Calculation of the above coefficients requires values at the interfaces of the control volumes. We calculate the interface values using linear interpolation. We solve these linear systems (90)–(91) iteratively for $m_1$ and $m_2$ with boundary conditions (64b)–(65b), using MATLAB's inbuilt function *mldivide*, and therefore the coupled discrete elliptic equations can be solved very efficiently.

## References


1. Adrien, B., Axel, B., Rolf, W., Jochen, S., Peter, L.: High resolution irradiance tailoring using multiple freeform surfaces. Opt. Express **21**(9), 10563–10571 (2013)
2. Yadav, N.K., ten Thije Boonkkamp, J.H.M., IJzerman, W.L.: A least-squares method for the design of two-reflector optical system. J. Phys. Photonics (accepted)
3. Brix, K., Hafizogullari, Y., Platen, A.: Designing illumination lenses and mirrors by the numerical solution of Monge–Ampère equations. J. Opt. Soc. Am. A **32**(10), 803–837 (2015)
4. Fang, F.Z., Zhang, X.D., Weckenmann, A., Zhang, G.X., Evans, C.: Manufacturing and measurement of freeform optics. CIRP Ann. Manuf. Technol. **62**(2), 823–846 (2013)
5. Chang, S., Wu, R., Zheng, Z.: Design beam shapers with double freeform surfaces to form a desired wavefront with prescribed illumination pattern by solving a Monge–Ampère type equation. J. Opt. **18**, 125602 (2016)
6. Oliker, V.: Differential equations for design of a freeform single lens with prescribed irradiance properties. Opt. Eng. **53**(3), 031302 (2013)
7. Prins, C.R., Beltman, R., ten Thije Boonkkamp, J.H.M., IJzerman, W.L., Tukker, T.W., : A least-squares method for optimal transport using the Monge–Ampère equation. SIAM J. Sci. Comput. **37**(6), B937–B961 (2015)
8. Glimm, T., Oliker, V.: Optical design of two-reflector systems, the Monge–Kantorovich mass transfer problem and Fermat's principle. Indiana Univ. Math. J. **53**(5), 11070–11078 (2004)
9. Oliker, V.: On design of free-form refractive beam shapers, sensitivity to figure error, and convexity of lenses. J. Opt. Soc. Am. A **25**(12), 3067–3076 (2008)
10. Oliker, V.: Designing freeform lenses for intensity and phase control of coherent light with help from geometry and mass transport. Arch. Ration. Mech. Anal. **201**(3), 1013–1045 (2011)
11. Froese, B.D.: A numerical method for the elliptic Monge–Ampère equation with transport boundary conditions. SIAM J. Sci. Comput. **34**, A1432–A1459 (2012)
12. Benamou, J.D., Froese, B.D., Oberman, A.M.: A viscosity solution approach to the Monge–Ampère formulation of the optimal transportation problem. arXiv:1208.4873 (2013)
13. Benamou, J.D., Froese, B.D., Oberman, A.M.: Numerical solution of the optimal transportation problem using the Monge–Ampère equation. J. Comput. Phys. **260**, 107–126 (2014)
14. Bösel, C., Gross, H.: Single freeform surface design for prescribed input wavefront and target irradiance. J. Opt. Soc. Am. A **34**(9), 1490–1499 (2017)
15. Wu, R., Xu, L., Liu, P., Zhang, Y., Zheng, Z., Li, H., Liu, X.: Freeform illumination design: a nonlinear boundary problem for the elliptic Monge–Ampère equation. Opt. Lett. **38**(2), 229–231 (2013)
16. Brix, K., Hafizogullari, Y., Platen, A.: Solving the Monge–Ampère equation for the inverse reflector problem. Math. Models Methods Appl. Sci. **25**, 803–837 (2015)

17. Glimm, T., Oliker, V.: Optical design of single-reflector systems and the Monge–Kantorovich mass transfer problem. J. Math. Sci. **117**(3), 4096–4108 (2003)
18. Rubinstein, J., Wolansky, G.: Intensity control with a free-form lens. J. Opt. Soc. Am. A **24**(2), 463–469 (2007)
19. Evans, L.C.: Partial differential equations and Monge–Kantorovich mass transfer. In: Current Developments in Mathematics, Cambridge (1997), International Press, Boston, pp. 65–126 (1999)
20. Bouchittè, G., Buttazzo, G., Seppecher, P.: Shape optimization solutions via Monge–Kantorovich equation. C. R. Acad. Sci. Paris Ser. I Math. **324**(10), 1185–1191 (1997)
21. Bouchittè, G., Buttazzo, G.: Characterization of optimal shapes and masses through Monge–Kantorovich equation. J. Eur. Math. Soc. **3**(2), 139–168 (2001)
22. Gangbo, W.: An Introduction to the Mass Transportation Theory and Its Applications. Lecture Notes: School of Mathematics Georgia Institute of Technology Atlanta, USA (2004)
23. Benamou, J.D., Brenier, Y.: A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. Numer. Math. **84**(3), 375–393 (2000)
24. Prins, C.R.: Inverse Methods for Illumination optics. Ph.d. thesis, Eindhoven University of Technology (2014)
25. Glassner, A.S.: An Introduction to Ray Tracing. Academic Press Ltd, London (1991)
26. Born, M., Wolf, E.: Principles of Optics, 5th edn. Pergamon Press, Oxford (1975)
27. Villani, C.: Topics in Optimal Transportation. Graduate Studies in Mathematics, vol. 58. American Mathematical Society, Providence (2003)
28. Ries, H., Rabl, A.: Edge-ray principle of nonimaging optics. J. Opt. Soc. Am. A **11**(10), 2627–2632 (1994)
29. Villani, C.: Optimal Transport Old and New, vol. 338. Springer, Berlin (2009)
30. Caboussat, A., Glowinski, R., Sorensen, D.C.: A least-squares method for the numerical solution of the Dirichlet problem for the elliptic Monge–Ampère equation in dimension two, ESAIM. J. Control Optim. Calc. Var. **19**(3), 780–810 (2013)
31. Glowinski, R.: Variational Methods for the Numerical Solution of Nonlinear Elliptic Problems. SIAM, Philadelphia (2015)
32. Caboussat, A., Glowinski, R., Gourzoulidis, D.: A least-squares/relaxation method for the numerical solution of the three-dimensional elliptic Monge–Ampère equation. J. Sci. Comput **77**(1), 53–78 (2018)
33. Adams, R.A., Essex, C.: A Complete Course Calculus, 8th edn. Pearson, Toronto (2013)
34. Mesterton-Gibbons, M.: A Primer on the Calculus of Variations and Optimal Control Theory, 1st edn. American Mathematical Society, Providence (2009)
35. Mattheij, R.M.M., Rienstra, S.W., ten Thije Boonkkamp, J.H.M.: Partial Differential Equations. Society for Industrial and Applied Mathematics, Philadelphia (2005)
36. Tignol, J.: Galois's Theory of Algebraic Equations. Longman Scientific and Technical, Harlow (1988)