A least-squares method for optimal transport using the Monge-Ampère equation

Citation for published version (APA):

Prins, C. R., Thijs Boonkkamp, ten, J. H. M., IJzerman, W. L., & Tukker, T. W. (2015). A least-squares method for optimal transport using the Monge-Ampère equation. *SIAM Journal on Scientific Computing*, 37(6), B937-B961. <https://doi.org/10.1137/140986414>

Document license:
TAVERNE

DOI:
[10.1137/140986414](https://doi.org/10.1137/140986414)

Document status and date:
Published: 01/01/2015

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.

A LEAST-SQUARES METHOD FOR OPTIMAL TRANSPORT USING THE MONGE–AMPÈRE EQUATION*

C. R. PRINS[†], R. BELTMAN[†], J. H. M. TEN THIJE BOONKKAMP[†],
W. L. IJZERMAN[‡], AND T. W. TUKKER[§]

Abstract. In this article we introduce a novel numerical method to solve the problem of optimal transport and the related elliptic Monge–Ampère equation. It is one of the few numerical algorithms capable of solving this problem efficiently with the proper transport boundary condition. The computation time scales well with the grid size and has the additional advantage that the target domain may be nonconvex. We present the method and several numerical experiments.

Key words. optimal mass transport, least-squares, Monge–Ampère equation

AMS subject classifications. 35J66, 35J96, 49K20, 65H04, 65K10, 65N99

DOI. 10.1137/140986414

1. Introduction. In this article, we introduce a novel numerical method to compute the solution of the optimal (mass) transport problem with quadratic cost function in two-dimensional domains. The optimal transport problem with quadratic cost function can be stated as follows. Let $f : \mathcal{X} \rightarrow [0, \infty)$ and $g : \mathcal{Y} \rightarrow (0, \infty)$ be bounded functions denoting (mass) densities with bounded compact supports $\mathcal{X} \subset \mathbb{R}^2$ and $\mathcal{Y} \subset \mathbb{R}^2$. The problem is to find a mapping $\mathbf{m} : \mathcal{X} \mapsto \mathcal{Y}$, minimizing the transportation cost

$$(1.1) \quad \mathcal{C}[\mathbf{m}] = \iint_{\mathcal{X}} |\mathbf{x} - \mathbf{m}(\mathbf{x})|^2 f(\mathbf{x}) \, d\mathbf{x},$$

where $|\cdot|$ denotes the vector 2-norm. In addition, $\mathbf{m}$ must rearrange the density f into the density g , meaning that we require

$$(1.2) \quad \iint_{\mathcal{X}} h(\mathbf{m}(\mathbf{x})) f(\mathbf{x}) \, d\mathbf{x} = \iint_{\mathcal{Y}} h(\mathbf{p}) g(\mathbf{p}) \, d\mathbf{p}$$

for all continuous test functions h [8]. In the classical problem statement [24, p. 1], the density f denotes the height of a pile of soil, and g denotes the depth of an excavation, and the goal is to transport the sand into the excavation with the least amount of work. Practical applications of optimal mass transport are, for example, shape recognition in image processing [2] and mesh generation [9].

An important theorem by Brenier [24, pp. 125–126] states that such an optimal mapping is the (almost everywhere) unique gradient of a convex function. Let ∇u

*Submitted to the journal's Computational Methods in Science and Engineering section September 11, 2014; accepted for publication (in revised form) October 2, 2015; published electronically November 24, 2015. This paper is an adjustment to a chapter of the thesis of the first author, prepared at Eindhoven University of Technology, Eindhoven, The Netherlands. It contains additional performance analysis and new numerical examples.

<http://www.siam.org/journals/sisc/37-6/98641.html>

[†]CASA, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (corienprins@home.nl, r.beltman@tue.nl, j.h.m.tenthijeboonkkamp@tue.nl).

[‡]Philips Lighting, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, Eindhoven, The Netherlands (wilbert.ijzerman@philips.com).

[§]Philips Research, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, Eindhoven, The Netherlands (teus.tukker@philips.com).

denote this gradient. Substituting $\mathbf{m} = \nabla u$ in (1.2), we find by a change of variables

$$(1.3) \quad \det(D^2u) = \frac{f(x, y)}{g(\nabla u(x, y))}, \quad (x, y) \in \mathcal{X}.$$

This equation is known as the Monge–Ampère equation. The accompanying boundary condition is derived from the condition that $\mathbf{m}$ maps $\mathcal{X}$ to $\mathcal{Y}$ and reads [9, 20, 23]

$$(1.4) \quad \nabla u(\partial\mathcal{X}) = \partial\mathcal{Y}.$$

The solution of the optimal transport problem with quadratic cost function defined by (1.1) and (1.2) can be computed by solving the Monge–Ampère equation (1.3) with boundary condition (1.4).

Our interest in optimal transport comes from the field of illumination optics, which concerns the design of lenses and reflectors for use in lighting. In [20] we have shown that the shape of a lens or reflector surface $z = u(x, y)$ redistributing the light from a parallel beam into a given output light distribution is described by a solution of (1.3) with boundary condition (1.4). The function $f(x, y)$ corresponds to the parallel beam of light and the function $g(\nabla u)$ corresponds to the output distribution. In our numerical results we will give an example of this application.

Until recently, very few efficient numerical methods for optimal transport were known. A numerical method for the related Monge–Ampère equation was introduced by Froese, Benamou, and Oberman [3, 4, 9, 10, 11, 12, 16, 17], using finite differences, a wide-stencil scheme, and an innovative discretization of the boundary condition. We implemented and tested this algorithm on several reflector design problems [21]. The method is robust and efficient. However, the method requires that the target domain is convex, which is a problem for practical reflector design. At the end of this article we will compare this method with the least-squares method presented in this paper.

Another popular method to solve the optimal transport problem was developed by Benamou and Brenier [2], using a periodic boundary condition. They add a time dimension to the problem and calculate a continuous evolution of the source density to the target density. This leads to a saddle-point problem which can be solved numerically. The method was further developed by Tannenbaum, Angenent, Haker, Haber, and Rehman [1, 13].

In this article, we introduce a new numerical method for optimal transport, based on the minimization of a least-squares functional. From this mapping we can also calculate the convex solution of the corresponding Monge–Ampère equation. Our new method is inspired by a least-squares method published recently by Caboussat, Glowinski, and Sorensen [5]. Their method numerically solves the Dirichlet problem of the elliptic Monge–Ampère equation, given by

$$(1.5) \quad \det(D^2u) = f(x, y) \quad \text{on } \mathcal{X}, \quad u(x, y) = h(x, y) \quad \text{on } \partial\mathcal{X}.$$

The authors minimize the functional

$$(1.6) \quad J(\phi, \mathbf{P}) = \frac{1}{2} \iint_{\mathcal{X}} \|\mathbf{D}^2\phi - \mathbf{P}\|^2 \, dx \, dy$$

over the set of real symmetric matrices $\mathbf{P}$ such that $\det \mathbf{P} = f(x, y)$ and functions ϕ satisfying the Dirichlet boundary condition. The minimization is performed alternately over $\mathbf{P}$ and ϕ . The minimization over $\mathbf{P}$ with fixed ϕ is a nonlinear problem

which is solved pointwise for each grid point using Newton iteration. The minimization over ϕ with fixed $\mathbf{P}$ comes down to solving a biharmonic equation, which is discretized using mixed finite elements. The function ϕ converges to the convex solution u of (1.5) [5].

Our algorithm differs from the algorithm of Caboussat, Glowinski, and Sorensen in three ways. First, we have an extra term in the functional J to account for the transport boundary condition. Second, we found a method to solve the minimization over $\mathbf{P}$ analytically, instead of using Newton iteration. Third, instead of u we compute the mapping function $\mathbf{m}$. The procedure for $\mathbf{m}$ boils down to solving two separate Poisson problems for the components m_1 and m_2 of $\mathbf{m}$. This is numerically much cheaper than solving a biharmonic equation, and the matrix resulting from the discretization has a much smaller condition number.

The outline of this paper is as follows. In section 2 we introduce the numerical method and give the outline of the algorithm. In each iteration of the algorithm, three minimization problems need to be solved. We elaborate the solutions for these minimization problems in section 3. In section 4 we finalize and summarize our description of the algorithm and show how to calculate the solution u of the Monge–Ampère equation from the optimal transportation mapping $\mathbf{m}$. We give several test results in section 5 and we end with discussion and conclusions in section 6.

2. Numerical method. In this section, we introduce our new numerical method to solve the optimal transport problem with quadratic cost function. The source density $f : \mathcal{X} \rightarrow [0, \infty)$ and target density $g : \mathcal{Y} \rightarrow (0, \infty)$ are such that

$$(2.1) \quad \iint_{\mathcal{X}} f(x, y) \, dx \, dy = \iint_{\mathcal{Y}} g(p, q) \, dp \, dq.$$

Note that this is a requirement on the density functions f and g , in contrast to (1.2), which is a requirement on the mapping $\mathbf{m}$.

We use the fact that the mapping, which solves the optimal transport problem, is equal to the gradient of the convex solution of the Monge–Ampère equation (1.3) with boundary condition (1.4). Therefore, we require that the Jacobi matrix of $\mathbf{m}$, given by

$$(2.2) \quad D\mathbf{m} = \begin{pmatrix} \frac{\partial m_1}{\partial x} & \frac{\partial m_1}{\partial y} \\ \frac{\partial m_2}{\partial x} & \frac{\partial m_2}{\partial y} \end{pmatrix},$$

equals a real symmetric positive semidefinite matrix $\mathbf{P}(x, y)$, satisfying

$$(2.3) \quad \det(\mathbf{P}(x, y)) = \frac{f(x, y)}{g(\mathbf{m}(x, y))}$$

for all $(x, y) \in \mathcal{X}$. Since $\mathbf{P}(x, y)$ is symmetric, $\partial m_1 / \partial y = \partial m_2 / \partial x$, which implies that $\mathbf{m}$ is a conservative vector field and thus the gradient of a function [14, p. 494]. We enforce the equality $D\mathbf{m} = \mathbf{P}$ by minimizing the following functional:

$$(2.4) \quad J_1(\mathbf{m}, \mathbf{P}) = \frac{1}{2} \iint_{\mathcal{X}} \|D\mathbf{m} - \mathbf{P}\|^2 \, dx \, dy.$$

The norm used in this functional is the Fröbenius norm, which is defined as follows. Let $\mathbf{P} : \mathcal{Q}$ denote the Fröbenius inner product of the matrices $\mathbf{P} = (p_{i,j})$ and

$\mathbf{Q} = (q_{i,j})$, defined by

$$(2.5) \quad \mathbf{P} : \mathbf{Q} = \sum_{i,j} p_{i,j} q_{i,j};$$

then the Fröbenius norm is defined as $\|\mathbf{P}\| = \sqrt{\mathbf{P} : \mathbf{P}}$. We minimize the functional in (2.4) over $\mathbf{m}$ and $\mathbf{P}$, where $\mathbf{m}$ comes from the set

$$(2.6) \quad \mathcal{V} = [C^2(\mathcal{X})]^2,$$

which is the set of two-dimensional, twice continuously differentiable vector fields. Let $[C^1(\mathcal{X})]_{\text{spsd}}^{2 \times 2}$ denote all symmetric positive semidefinite 2×2 matrices of $C^1(\mathcal{X})$ -functions. The matrix $\mathbf{P}$ comes from the set

$$(2.7) \quad \mathcal{P}(\mathbf{m}) = \left\{ \mathbf{P} \in [C^1(\mathcal{X})]_{\text{spsd}}^{2 \times 2} \mid \det(\mathbf{P}(x, y)) = \frac{f(x, y)}{g(\mathbf{m}(x, y))} \right\}.$$

The function sets are chosen to match the type of equations that are solved to find $\mathbf{m}$ and $\mathbf{P}$. Because the optimal transport problem has a solution, there exist $\mathbf{P}$ and $\mathbf{m}$ such that $J_I(\mathbf{m}, \mathbf{P}) = 0$. Therefore $D\mathbf{m}$ will be a symmetric positive semidefinite matrix and $\mathbf{m}$ a conservative vector field.

In addition to mass conservation, we require that the mapping $\mathbf{m}$ maps the source $\mathcal{X}$ to the target $\mathcal{Y}$, i.e.,

$$(2.8) \quad \mathbf{m}(\mathcal{X}) = \mathcal{Y}.$$

To this end, we impose the boundary condition (1.4):

$$(2.9) \quad \mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}.$$

This boundary condition is nonstandard and nonlinear and requires special attention. We address the boundary condition by minimizing a second functional:

$$(2.10) \quad J_B(\mathbf{m}, \mathbf{b}) = \frac{1}{2} \oint_{\partial\mathcal{X}} |\mathbf{m} - \mathbf{b}|^2 \, ds.$$

We minimize this functional over $\mathbf{b}$ from the set

$$(2.11) \quad \mathcal{B} = \{\mathbf{b} \in [C^1(\partial\mathcal{X})]^2 \mid \mathbf{b}(\mathbf{x}) \in \partial\mathcal{Y} \quad \forall \mathbf{x} \in \partial\mathcal{X}\}.$$

Again, because the optimal transport problem has a solution, we can find an $\mathbf{m}$ and $\mathbf{b}$ such that $J_B(\mathbf{m}, \mathbf{b}) = 0$, so $\mathbf{m}(\partial\mathcal{X}) \subseteq \partial\mathcal{Y}$. From the continuity of $\mathbf{m}$ follows $\mathbf{m}(\partial\mathcal{X}) = \partial\mathcal{Y}$ [20, p. 94].

We combine the functionals J_I for the interior and J_B for the boundary by a weighted average:

$$(2.12) \quad J(\mathbf{m}, \mathbf{P}, \mathbf{b}) = (1 - \alpha) J_B(\mathbf{m}, \mathbf{b}) + \alpha J_I(\mathbf{m}, \mathbf{P}).$$

The parameter α ($0 < \alpha < 1$) controls the weight of J_I compared to J_B . We calculate the minimizers by repeatedly minimizing over the three sets $\mathcal{V}$, $\mathcal{P}(\mathbf{m})$, and $\mathcal{B}$ separately. We start with an initial guess $\mathbf{m}^0$, which will be specified shortly. Subsequently, we perform the iteration

$$(2.13a) \quad \mathbf{b}^{n+1} = \underset{\mathbf{b} \in \mathcal{B}}{\operatorname{argmin}} J_B(\mathbf{m}^n, \mathbf{b}),$$

$$(2.13b) \quad \mathbf{P}^{n+1} = \underset{\mathbf{P} \in \mathcal{P}(\mathbf{m}^n)}{\operatorname{argmin}} J_I(\mathbf{m}^n, \mathbf{P}),$$

$$(2.13c) \quad \mathbf{m}^{n+1} = \underset{\mathbf{m} \in \mathcal{V}}{\operatorname{argmin}} J(\mathbf{m}, \mathbf{P}^{n+1}, \mathbf{b}^{n+1}).$$

We will elaborate the stopping criterion for this iteration in section 5.1.

We solve the optimal transport problem using a standard rectangular $N_x \times N_y$ grid for some $N_x, N_y \in \mathbb{N}$. Let $[a_{\min}, a_{\max}] \times [b_{\min}, b_{\max}]$ be the smallest bounding box of $\mathcal{X}$ aligned with the axes. The grid is given by

$$(2.14a) \quad x_i = a_{\min} + (i-1)h_x, \quad h_x = \frac{a_{\max} - a_{\min}}{N_x - 1}, \quad i = 1, \dots, N_x,$$

$$(2.14b) \quad y_j = b_{\min} + (j-1)h_y, \quad h_y = \frac{b_{\max} - b_{\min}}{N_y - 1}, \quad j = 1, \dots, N_y.$$

We initialize our minimization by constructing an initial guess $\mathbf{m}^0$ which maps the source region $\mathcal{X}$ to a bounding box of the target region $\mathcal{Y}$. Without loss of generality we assume the source region has a rectangular shape $[a_{\min}, a_{\max}] \times [b_{\min}, b_{\max}]$. The source density function $f(x, y)$ may be zero on part of $\mathcal{X}$. The target density function $g(p, q)$ must be nonzero, and we assume $\mathcal{Y}$ to be simply connected. Let $[c_{\min}, c_{\max}] \times [d_{\min}, d_{\max}] \supset \mathcal{Y}$ be the smallest bounding box of $\mathcal{Y}$ aligned with the coordinates p and q . The initial guess is given by

$$(2.15a) \quad m_1^0 = \frac{x - a_{\min}}{a_{\max} - a_{\min}} c_{\max} + \frac{a_{\max} - x}{a_{\max} - a_{\min}} c_{\min},$$

$$(2.15b) \quad m_2^0 = \frac{y - b_{\min}}{b_{\max} - b_{\min}} d_{\max} + \frac{b_{\max} - y}{b_{\max} - b_{\min}} d_{\min}.$$

Note that the corresponding Jacobi matrix $D\mathbf{m}^0$ is symmetric (in fact diagonal) positive definite. In the next paragraphs, we elaborate each of the three minimization steps (2.13a), (2.13b), and (2.13c).

3. Minimizing procedures for $\mathbf{b}$, $\mathbf{P}$, and $\mathbf{m}$. Using the initial guess $\mathbf{m}^0$, we start the iteration process (2.13). Each iteration consists of three steps, which are described one by one in the following three paragraphs.

3.1. Minimizing procedure for $\mathbf{b}$. In this section we perform the minimization step (2.13a). We assume $\mathbf{m} = \mathbf{m}^n$ given, and we minimize the functional $J_B(\mathbf{m}, \mathbf{b})$ over $\mathbf{b} \in \mathcal{B}$. For simplicity of notation we drop the indices n and $n+1$. The minimization can be performed pointwise, because no derivative of $\mathbf{b}$ with respect to x or y appears in the functional $J_B(\mathbf{m}, \mathbf{b})$. For each grid point $(x_i, y_j) \in \partial\mathcal{X}$, let $\mathbf{m}_{i,j}$ denote the value of the mapping. We perform the minimization

$$(3.1) \quad \min_{\mathbf{b}_{i,j} \in \partial\mathcal{Y}} |\mathbf{m}_{i,j} - \mathbf{b}_{i,j}|^2.$$

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 along the boundary and 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 point to $\mathbf{m}_{i,j}$ on all line segments $(\mathbf{z}_k, \mathbf{z}_{k+1})$.

First we determine for a given point $\mathbf{m}_{i,j}$ and a given line segment $(\mathbf{z}_k, \mathbf{z}_{k+1})$ the projection $\mathbf{m}_{i,j}^k$ of $\mathbf{m}_{i,j}$ on the line through $\mathbf{z}_k$ and $\mathbf{z}_{k+1}$. This projection is given by [14, p. 30]

$$(3.2a) \quad \mathbf{m}_{i,j}^k = \mathbf{z}_k + t_k (\mathbf{z}_{k+1} - \mathbf{z}_k),$$

$$(3.2b) \quad t_k = \frac{(\mathbf{m}_{i,j} - \mathbf{z}_k) \cdot (\mathbf{z}_{k+1} - \mathbf{z}_k)}{|\mathbf{z}_k - \mathbf{z}_{k+1}|^2},$$

see also Figure 1. If $t_k \in [0, 1]$, the projected point is on the line segment, and the nearest point to $\mathbf{m}_{i,j}$ is given by $\mathbf{m}_{i,j}^k$. If $t_k < 0$, the projected point is not on the

FIG. 1. Calculation of the distance of $\mathbf{m}_{i,j}$ to a line segment $(\mathbf{z}_k, \mathbf{z}_{k+1})$. The point $\mathbf{m}_{i,j}^k$ is the projection of $\mathbf{m}_{i,j}$ on the line through $\mathbf{z}_k$ and $\mathbf{z}_{k+1}$. In this case we find $t_k < 0$ and the nearest point on the line segment to $\mathbf{m}_{i,j}$ is given by $\mathbf{b}_{i,j} = \mathbf{z}_k$.

line segment and the nearest point is given by $\mathbf{z}_k$, and if $t_k > 1$, the nearest point is given by $\mathbf{z}_{k+1}$. Thus, the nearest point on the line segment $(\mathbf{z}_k, \mathbf{z}_{k+1})$ is given by

$$(3.3) \quad \widetilde{\mathbf{m}}_{i,j}^k = \mathbf{z}_k + \min(1, \max(0, t_k)) (\mathbf{z}_{k+1} - \mathbf{z}_k).$$

We calculate the nearest point $\mathbf{b}_{i,j}$ on all the line segments by

$$(3.4) \quad \mathbf{b}_{i,j} = \operatorname{argmin}_{\widetilde{\mathbf{m}}_{i,j}^k} |\widetilde{\mathbf{m}}_{i,j}^k - \mathbf{m}_{i,j}|^2.$$

This procedure is repeated for all grid points $\mathbf{x}_{i,j}$ on $\partial\mathcal{X}$.

3.2. Minimizing procedure for $\mathbf{P}$. In this section we perform the minimization step (2.13b). We assume that $\mathbf{m} = \mathbf{m}^n$ is given and minimize $J_I(\mathbf{m}, \mathbf{P})$ defined in (2.4) over the matrices $\mathbf{P} \in \mathcal{P}(\mathbf{m})$. Thus, we require that $\mathbf{P}$ is symmetric and satisfies

$$(3.5) \quad \det(\mathbf{P}) = \frac{f(x, y)}{g(\mathbf{m}(x, y))}.$$

In addition, $\mathbf{P}$ should be positive semidefinite. Recall that the 2×2 matrix $\mathbf{P}$ is positive semidefinite iff

$$(3.6) \quad \operatorname{tr}(\mathbf{P}) \geq 0 \quad \text{and} \quad \det(\mathbf{P}) \geq 0.$$

The second inequality in (3.6) is already satisfied because of (3.5); thus we only have to verify the inequality $\operatorname{tr}(\mathbf{P}) \geq 0$.

Since the integrand of $J_I(\mathbf{m}, \mathbf{P})$ does not contain derivatives of $\mathbf{P}$, the minimization can be carried out pointwise; thus for each $(x_i, y_j) \in \mathcal{X}$ we have to minimize $\|\mathbf{D} - \mathbf{P}\|$, where $\mathbf{D}$ is the central difference approximation of $\mathbf{D}\mathbf{m}$. For ease of presentation we introduce the notation

$$(3.7a) \quad \mathbf{P} = \begin{pmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix}$$

with

$$(3.7b) \quad d_{11} = \delta_x m_1, \quad d_{12} = \delta_y m_1, \quad d_{21} = \delta_x m_2, \quad d_{22} = \delta_y m_2,$$

where δ_x and δ_y are the central difference approximations of $\partial/\partial x$ and $\partial/\partial y$, respectively. Moreover, to avoid crossing grid lines we require $d_{11}, d_{22} > 0$ by imposing

$$(3.8) \quad d_{11} = \max(\delta_x m_1, \varepsilon), \quad d_{22} = \max(\delta_y m_2, \varepsilon)$$

for a threshold value $\varepsilon > 0$. In our computations we chose $\varepsilon = 10^{-8}$. As a consequence, $\text{tr}(\mathbf{D}) > 0$. Note that $\mathbf{D}$ need not be symmetric and $d_{12}, d_{21} < 0$ is possible. Next, we introduce the function

$$(3.9) \quad H(p_{11}, p_{22}, p_{12}) = \frac{1}{2} \|\mathbf{D} - \mathbf{P}\|^2.$$

Thus, for each $(x_i, y_j) \in \mathcal{X}$ we have the following problem:

$$(3.10a) \quad \text{minimize} \quad H(p_{11}, p_{22}, p_{12})$$

$$(3.10b) \quad \text{subject to} \quad p_{11}p_{22} - p_{12}^2 = \frac{f}{g},$$

$$(3.10c) \quad p_{11} + p_{22} \geq 0,$$

where f/g is shorthand notation for $f(x_i, y_j)/g(\mathbf{m}(x_i, y_j))$. This problem can be solved analytically. First, we slightly simplify the minimization problem. Let $\mathbf{D}_S$ be the symmetric part of $\mathbf{D}$, i.e.,

$$(3.11) \quad \mathbf{D}_S = \begin{pmatrix} d_{11} & d_S \\ d_S & d_{22} \end{pmatrix},$$

with $d_S := \frac{1}{2}(d_{12} + d_{21})$. Note that also $\text{tr}(\mathbf{D}_S) > 0$. The new function to be minimized reads

$$(3.12) \quad H_S(p_{11}, p_{22}, p_{12}) = \frac{1}{2} \|\mathbf{D}_S - \mathbf{P}\|^2 = H(p_{11}, p_{22}, p_{12}) - \frac{1}{4}(d_{12} - d_{21})^2.$$

Because $(d_{12} - d_{21})^2$ is constant with respect to 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 $H(p_{11}, p_{22}, p_{12})$, the minimization problem (3.10) is equivalent to

$$(3.13a) \quad \text{minimize} \quad H_S(p_{11}, p_{22}, p_{12})$$

$$(3.13b) \quad \text{subject to} \quad p_{11}p_{22} - p_{12}^2 = \frac{f}{g},$$

$$(3.13c) \quad p_{11} + p_{22} \geq 0.$$

Given d_{11} , d_{22} , d_S , and f/g , we distinguish several cases, with at least one and at most four possible solutions (p_{11}, p_{22}, p_{12}) that minimize $H_S(p_{11}, p_{22}, p_{12})$ and satisfy the nonlinear constraint $p_{11}p_{22} - p_{12}^2 = f/g$, which are (local) minima, maxima, or saddle-points. From these we have to select the ones that give rise to a positive semidefinite matrix $\mathbf{P}$, i.e., satisfy $p_{11} + p_{22} \geq 0$. We will show that this is always possible and refer to these solutions as possible minimizers. Finally, we determine the global minimizer by substituting each possible minimizer in the function $H_S(p_{11}, p_{22}, p_{12})$.

The possible minimizers of (3.13) are critical points of the Lagrangian function $L = L(p_{11}, p_{22}, p_{12}; \lambda)$ defined by

$$(3.14) \quad L(p_{11}, p_{22}, p_{12}; \lambda) = \frac{1}{2} \|\mathbf{D}_S - \mathbf{P}\|^2 + \lambda \left(\det(\mathbf{P}) - \frac{f}{g} \right),$$

where λ is the Lagrange multiplier. If we set all partial derivatives of L to 0, we obtain the following algebraic system:

$$(3.15a) \quad p_{11} + \lambda p_{22} = d_{11},$$

$$(3.15b) \quad \lambda p_{11} + p_{22} = d_{22},$$

$$(3.15c) \quad (1 - \lambda) p_{12} = d_S,$$

$$(3.15d) \quad p_{11} p_{22} - p_{12}^2 = f/g.$$

For given λ , the system (3.15a)–(3.15c) is linear in p_{11} , p_{22} , and p_{12} and is regular if $\lambda \neq \pm 1$. In this case, we can express p_{11} , p_{22} , and p_{12} in terms of λ , substitute the expressions in (3.15d), and derive a quartic equation for λ . We will derive exact solution formulas for this equation. Moreover, from (3.15a)–(3.15b) it is obvious that

$$(3.16) \quad \text{tr}(\mathbf{P}) = \frac{\text{tr}(\mathbf{D}_S)}{\lambda + 1},$$

and consequently, we have to select Lagrange multipliers that satisfy $\lambda > -1$ to enforce $\text{tr}(\mathbf{P}) > 0$. On the other hand, for $\lambda = \pm 1$ the system (3.15a)–(3.15c) is singular. The case $\lambda = -1$ does not occur, since this would imply $d_{11} + d_{22} = 0$, in contradiction with the requirement $\text{tr}(\mathbf{D}_S) > 0$. It remains to investigate the case $\lambda = 1$. From the implication

$$(3.17) \quad \lambda = 1 \Rightarrow [d_{11} = d_{22} \quad \text{and} \quad d_S = 0],$$

we conclude

$$(3.18) \quad [d_{11} \neq d_{22} \quad \text{or} \quad d_S \neq 0] \Rightarrow \lambda \neq 1,$$

and then the system (3.15a)–(3.15c) is regular. On the other hand, if $d_{11} = d_{22}$ and $d_S = 0$ it is possible that $\lambda = 1$, and therefore we have to solve the minimization problem (3.13) in a different manner. Also in this case a solution with $\text{tr}(\mathbf{P}) \geq 0$ is possible.

First we assume that (3.18) applies. Then, we obtain

$$(3.19) \quad p_{11} = \frac{\lambda d_{22} - d_{11}}{\lambda^2 - 1}, \quad p_{22} = \frac{\lambda d_{11} - d_{22}}{\lambda^2 - 1}, \quad p_{12} = \frac{d_S}{1 - \lambda}.$$

Substituting these expressions in (3.15d) gives the quartic equation

$$(3.20a) \quad F(\lambda) = a_4 \lambda^4 + a_2 \lambda^2 + a_1 \lambda + a_0 = 0,$$

with coefficients given by

$$(3.20b) \quad a_4 = f/g \geq 0,$$

$$(3.20c) \quad a_2 = -2f/g - \det(\mathbf{D}_S),$$

$$(3.20d) \quad a_1 = \|\mathbf{D}_S\|^2 \geq 0,$$

$$(3.20e) \quad a_0 = f/g - \det(\mathbf{D}_S).$$

Solution of (3.20a) when $f > 0$. It can be shown that (3.20) has at least two real roots. Moreover, using the relations

$$\|D_S\|^2 = \operatorname{tr}(D_S)^2 - 2 \det(D_S), \quad \operatorname{tr}(D_S)^2 - 4 \det(D_S) \geq 0,$$

we conclude that $F(-1) = -\operatorname{tr}(D_S)^2 < 0$ and $F(1) = \operatorname{tr}(D_S)^2 - 4 \det(D_S) \geq 0$, implying that (3.20) has at least one solution $\lambda > -1$. The solutions of (3.20) are given by

$$(3.21a) \quad \lambda_1 = -\sqrt{\frac{y}{2}} + \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} + \frac{a_1}{2a_4\sqrt{2y}}},$$

$$(3.21b) \quad \lambda_2 = -\sqrt{\frac{y}{2}} - \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} + \frac{a_1}{2a_4\sqrt{2y}}},$$

$$(3.21c) \quad \lambda_3 = \sqrt{\frac{y}{2}} + \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} - \frac{a_1}{2a_4\sqrt{2y}}},$$

$$(3.21d) \quad \lambda_4 = \sqrt{\frac{y}{2}} - \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} - \frac{a_1}{2a_4\sqrt{2y}}},$$

where y is the solution of a cubic equation; for more details see the appendix. The real roots $\lambda > -1$ in (3.21) are substituted in (3.19) and (3.12), yielding the possible minimizers of $H_S(p_{11}, p_{22}, p_{12})$. Note that we have division by 0 in (3.21) if $y = 0$. In the appendix we will show that this happens only when $d_{11} = d_{22} = d_S = 0$, corresponding to the situation (3.17), which we will discuss later.

Solution of (3.20a) when $f = 0$. If $f = 0$, i.e., when the source density is zero, (3.20a) reduces to a quadratic equation because $a_4 = 0$. The roots are given by

$$(3.22) \quad \lambda_{\pm} = \frac{-a_1 \pm \sqrt{a_1^2 - 4a_2a_0}}{2a_2}.$$

The discriminant of this quadratic equation is always positive. We can verify this by substituting the coefficients in (3.20) in the discriminant:

$$(3.23) \quad a_1^2 - 4a_2a_0 = (d_{11}^2 - d_{22}^2)^2 + 4d_S^2(d_{11} + d_{22})^2 \geq 0.$$

Equality is never possible; therefore we always have two distinct real solutions. Also in this case $F(-1) = -\operatorname{tr}(D_S)^2 < 0$ and $F(1) = \operatorname{tr}(D_S)^2 - 4 \det(D_S) \geq 0$; consequently (3.20) has at least one solution $\lambda > -1$.

Solution of (3.20a) when $f = 0$ and $a_2 = 0$. If $f = 0$ and $a_2 = 0$, we have division by zero in (3.22). This occurs when $d_{11}d_{22} - d_S^2 = 0$, and we find from (3.20) that $a_0 = 0$ as well. The only solution to (3.20a) is then given by $\lambda = 0$ and the inequality $\operatorname{tr}(P) \geq 0$ trivially holds. Again the values of p_{11} , p_{22} , and p_{12} are calculated using (3.19), giving one possible minimizer $p_{11} = d_{11}$, $p_{22} = d_{22}$, and $p_{12} = d_S$.

Solution of (3.13) when (3.17) applies. If $d_{11} = d_{22}$ and $d_S = 0$, we cannot perform the step from (3.15) to (3.19) because we have the possibility that $\lambda = 1$. Therefore we determine the minimum of $H_S(p_{11}, p_{22}, p_{12})$ using another method. Using $d_S = 0$ and $d_{22} = d_{11}$, the minimization (3.13) simplifies to

$$(3.24a) \quad \operatorname{argmin}_{(p_{11}, p_{22}, p_{12}) \in \mathbb{R}^3} \frac{1}{2}((p_{11} - d_{11})^2 + 2p_{12}^2 + (p_{22} - d_{11})^2),$$

subject to the conditions

$$(3.24b) \quad p_{11}p_{22} - p_{12}^2 = \frac{f}{g}, \quad p_{11} + p_{22} \geq 0.$$

From the first constraint in (3.24b) it follows that $p_{12}^2 = p_{11}p_{22} - f/g$. Substitution of this expression in (3.24a) gives

$$(3.25) \quad \operatorname{argmin}_{(p_{11}, p_{22})} \frac{1}{2} \left((p_{11} - d_{11})^2 + 2(p_{11}p_{22} - f/g) + (p_{22} - d_{11})^2 \right),$$

where we replaced minimization over $\mathbb{R}^3$ by minimization over $\mathbb{R}^2$ restricted to the domain where $p_{12} = \pm\sqrt{p_{11}p_{22} - f/g}$ is real. The minimizer can be found in the interior of this domain or on the boundary. We find the minimizer in the interior by setting the derivatives with respect to p_{11} and p_{22} to 0. This yields a critical line:

$$(3.26) \quad p_{11} + p_{22} = d_{11}.$$

Next, we need to know which part of this line corresponds to real values of p_{12} . Substituting $p_{11} + p_{22} = d_{11}$ in $p_{11}p_{22} - f/g \geq 0$, we find

$$(3.27) \quad p_{11}^2 - d_{11}p_{11} + \frac{f}{g} \leq 0.$$

This inequality has real solutions p_{11} if the discriminant of the quadratic equation on the left-hand side $d_{11}^2 - 4f/g \geq 0$. Then, the segment of the line corresponding to real values of p_{11} is given by

$$(3.28) \quad \frac{d_{11} - \sqrt{d_{11}^2 - 4f/g}}{2} \leq p_{11} \leq \frac{d_{11} + \sqrt{d_{11}^2 - 4f/g}}{2}.$$

This is a minimizing line segment, and the minimizer may not be unique. For simplicity, we choose p_{11} in the middle of the line segment and find the solutions:

$$(3.29) \quad p_{11} = p_{22} = \frac{d_{11}}{2}, \quad p_{12} = \pm\sqrt{\frac{d_{11}^2}{4} - \frac{f}{g}}.$$

Next, we need to find possible minimizers on the boundary of the domain where p_{12} is real. This boundary is given by $p_{12} = 0$, which is the hyperbola

$$(3.30) \quad p_{11}p_{22} = f/g.$$

Substituting $p_{12} = 0$ and $p_{22} = f/(gp_{11})$, the minimization problem (3.25) reduces to

$$(3.31) \quad \operatorname{argmin}_{p_{11} \in \mathbb{R}} \frac{1}{2} \left((p_{11} - d_{11})^2 + \left(\frac{f}{gp_{11}} - d_{11} \right)^2 \right).$$

We find the critical points by differentiation with respect to p_{11} and subsequent multiplication with p_{11}^3 . This gives the quartic equation

$$(3.32) \quad p_{11}^4 - d_{11}p_{11}^3 + f/g d_{11}p_{11} - f^2/g^2 = 0.$$

This equation can be factored into

$$(3.33) \quad (p_{11}^2 - f/g) (p_{11}^2 - d_{11}p_{11} + f/g) = 0$$

and has four solutions. The first two solutions are given by

$$(3.34) \quad p_{11} = \pm \sqrt{f/g}$$

and are always real. The other two solutions

$$(3.35) \quad p_{11} = \frac{d_{11} \pm \sqrt{d_{11}^2 - 4f/g}}{2}$$

either correspond to the endpoints of the line segment defined by (3.28) or are complex valued. If they correspond to the boundaries of the line segment, they will yield the same value of $H_S(p_{11}, p_{22}, p_{12})$ as the solution (3.29) and can thus be ignored. If they are complex valued, they can be ignored as well. From the first two solutions (3.34), we find the following two possible minimizers:

$$(3.36) \quad p_{11} = p_{22} = \pm \sqrt{f/g}, \quad p_{12} = 0.$$

For the critical points in the interior, given by (3.26), we have $\text{tr}(\mathbf{P}) = \frac{1}{2}\text{tr}(\mathbf{D}_S) \geq 0$, yielding a positive semidefinite matrix $\mathbf{P}$. On the other hand, for the extremes at the boundary, given in (3.36), we have to choose the positive solution in order to guarantee $\text{tr}(\mathbf{P}) \geq 0$.

Summary. The procedure to minimize over $\mathbf{P}$ is as follows. For each grid point (x_i, y_j) , we calculate the value of $f(x_i, y_j)/g(\mathbf{m}(x_i, y_j))$. We determine which case applies, compute all critical points (p_{11}, p_{22}, p_{12}) of the Lagrangian $L(p_{11}, p_{22}, p_{12}; \lambda)$, and select the possible minimizers that give rise to the condition $\text{tr}(\mathbf{P}) \geq 0$. Then, for each (x_i, y_j) , the new $\mathbf{P}$ that gives the smallest value of $H_S(p_{11}, p_{22}, p_{12})$ is chosen from the possible minimizers.

3.3. Minimizing procedure for $\mathbf{m}$. In this section, we perform the minimization step (2.13c). We assume $\mathbf{P}$ and $\mathbf{b}$ are given and minimize $J(\mathbf{m}, \mathbf{P}, \mathbf{b})$ over the functions $\mathbf{m} \in \mathcal{V}$. For ease of notation, we omit the indices n and $n+1$. In contrast to the other two minimization steps, this step cannot be performed pointwise. We minimize $\mathbf{m}$ using calculus of variations. Using the identity

$$(3.37) \quad \|\mathbf{A} + \mathbf{B}\|^2 = \|\mathbf{A}\|^2 + 2\mathbf{A} : \mathbf{B} + \|\mathbf{B}\|^2,$$

we calculate the first variation of J with respect to $\mathbf{m}$ for $\boldsymbol{\eta} \in [C^2(\mathcal{X})]^2$, i.e.,

$$(3.38) \quad \begin{aligned} \delta J(\mathbf{m}, \mathbf{P}, \mathbf{b})[\boldsymbol{\eta}] &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} (J(\mathbf{m} + \epsilon \boldsymbol{\eta}, \mathbf{P}, \mathbf{b}) - J(\mathbf{m}, \mathbf{P}, \mathbf{b})) \\ &= \lim_{\epsilon \rightarrow 0} \left[\frac{\alpha}{2} \iint_{\mathcal{X}} 2(\mathbf{D}\mathbf{m} - \mathbf{P}) : \mathbf{D}\boldsymbol{\eta} + \epsilon \|\mathbf{D}\boldsymbol{\eta}\|^2 \, dx \, dy \right. \\ &\quad \left. + \frac{1-\alpha}{2} \oint_{\partial\mathcal{X}} 2(\mathbf{m} - \mathbf{b}) \cdot \boldsymbol{\eta} + \epsilon |\boldsymbol{\eta}|^2 \, ds \right] \\ &= \alpha \iint_{\mathcal{X}} (\mathbf{D}\mathbf{m} - \mathbf{P}) : \mathbf{D}\boldsymbol{\eta} \, dx \, dy + (1-\alpha) \oint_{\partial\mathcal{X}} (\mathbf{m} - \mathbf{b}) \cdot \boldsymbol{\eta} \, ds. \end{aligned}$$

The minimizer is given by

$$(3.39) \quad \delta J(\mathbf{m}, \mathbf{P}, \mathbf{b})[\boldsymbol{\eta}] = 0 \quad \forall \boldsymbol{\eta} \in [C^2(\mathcal{X})]^2.$$

Introducing the vectors

$$\mathbf{p}_1 = \begin{pmatrix} p_{11} \\ p_{12} \end{pmatrix}, \quad \mathbf{p}_2 = \begin{pmatrix} p_{12} \\ p_{22} \end{pmatrix}, \quad \boldsymbol{\eta} = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix},$$

we can rewrite the first integral of the final expression in (3.38) as follows:

$$(3.40) \quad \iint_{\mathcal{X}} (\mathbf{D}\mathbf{m} - \mathbf{P}) : \mathbf{D}\boldsymbol{\eta} \, dx \, dy = \sum_{k=1}^2 \iint_{\mathcal{X}} (\nabla m_k - \mathbf{p}_k) \cdot \nabla \eta_k \, dx \, dy.$$

Using the vector-scalar product rule [14, p. 144] and Gauss's theorem [14, p. 506] we find for $k = 1, 2$,

$$(3.41) \quad \begin{aligned} & \iint_{\mathcal{X}} (\nabla m_k - \mathbf{p}_k) \cdot \nabla \eta_k \, dx \, dy \\ &= \oint_{\partial\mathcal{X}} (\nabla m_k - \mathbf{p}_k) \cdot \hat{\mathbf{n}} \, \eta_k \, ds - \iint_{\mathcal{X}} (\Delta m_k - \nabla \cdot \mathbf{p}_k) \eta_k \, dx \, dy, \end{aligned}$$

where $\hat{\mathbf{n}}$ is the outward unit normal vector on $\partial\mathcal{X}$. Combining (3.41) with (3.38), (3.39), and (3.40) we obtain

$$(3.42) \quad \begin{aligned} & \sum_{k=1}^2 \left(\oint_{\partial\mathcal{X}} (\alpha (\nabla m_k - \mathbf{p}_k) \cdot \hat{\mathbf{n}} + (1 - \alpha) (m_k - b_k)) \eta_k \, ds \right. \\ & \quad \left. - \alpha \iint_{\mathcal{X}} (\Delta m_k - \nabla \cdot \mathbf{p}_k) \eta_k \, dx \, dy \right) = 0 \quad \forall \boldsymbol{\eta} \in [C^2(\mathcal{X})]^2, \end{aligned}$$

where b_k ($k = 1, 2$) are the components of $\mathbf{b}$. Choosing $\eta_2 = 0$ and applying the fundamental lemma of calculus of variations [6, p. 185] for η_1 , we conclude that m_1 satisfies almost everywhere the Poisson problem

$$(3.43a) \quad \Delta m_1 = \nabla \cdot \mathbf{p}_1, \quad (x, y) \in \mathcal{X},$$

$$(3.43b) \quad (1 - \alpha) m_1 + \alpha \nabla m_1 \cdot \hat{\mathbf{n}} = (1 - \alpha) b_1 + \alpha \mathbf{p}_1 \cdot \hat{\mathbf{n}}, \quad (x, y) \in \partial\mathcal{X}.$$

Similarly, choosing $\eta_1 = 0$ we find

$$(3.44a) \quad \Delta m_2 = \nabla \cdot \mathbf{p}_2, \quad (x, y) \in \mathcal{X},$$

$$(3.44b) \quad (1 - \alpha) m_2 + \alpha \nabla m_2 \cdot \hat{\mathbf{n}} = (1 - \alpha) b_2 + \alpha \mathbf{p}_2 \cdot \hat{\mathbf{n}}, \quad (x, y) \in \partial\mathcal{X}.$$

These are two decoupled Poisson equations with Robin boundary conditions for the two components of $\mathbf{m}$ [15]. These Poisson boundary value problems are solved using standard second order central differences for the first and second order derivatives. These approximations contain points outside the domain when applied at the boundary. We eliminate these points using the discretized Robin boundary condition. The derivatives of p_{11} , p_{22} , and p_{12} are approximated using similar finite difference schemes, and when needed on the boundary, using one-sided second order schemes. The discrete systems are solved using the MATLAB implementation of LU decomposition with full pivoting. The LU decomposition needs to be computed only once, and the Poisson problems can be solved very efficiently during iteration.

4. Computation of u and algorithm summary. The minimization steps given by (2.13a), (2.13b), and (2.13c) are repeated until convergence. Since we are interested in the solution of the Monge–Ampère equation, we compute its solution u from the mapping $\mathbf{m}$, i.e., we calculate u such that $\nabla u = \mathbf{m}$. This u will be the approximate solution of the Monge–Ampère equation (1.3) with boundary condition (1.4). In the ideal situation, $D\mathbf{m} = \mathbf{P}$ and thus $\partial m_1/\partial y = \partial m_2/\partial x$. In this case there exists a function u such that $\nabla u = \mathbf{m}$, because $\mathbf{m}$ is a conservative vector field [14, p. 494]. However, we will most likely not be in this ideal situation; therefore we look for a function which has $\mathbf{m}$ as gradient in the least-squares sense, i.e.,

$$(4.1) \quad u = \operatorname{argmin}_{\psi} I(\psi), \quad I(\psi) = \frac{1}{2} \iint_{\mathcal{X}} |\nabla \psi - \mathbf{m}|^2 \, dx \, dy.$$

We calculate the minimizing function u using calculus of variations. The first variation of the functional (4.1) is given by

$$(4.2) \quad \begin{aligned} \delta I(u)[v] &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} (I(u + \epsilon v) - I(u)) \\ &= \lim_{\epsilon \rightarrow 0} \frac{1}{2} \iint_{\mathcal{X}} \epsilon |\nabla v|^2 + 2 (\nabla u - \mathbf{m}) \cdot \nabla v \, dx \, dy \\ &= \iint_{\mathcal{X}} (\nabla u - \mathbf{m}) \cdot \nabla v \, dx \, dy. \end{aligned}$$

The minimizer is given by

$$(4.3) \quad \delta I(u)[v] = 0 \quad \forall v \in C^2(\mathcal{X}).$$

Let $\hat{\mathbf{n}}$ denote the unit outward normal at the boundary $\partial\mathcal{X}$. Using integration by parts and Gauss's theorem we find

$$(4.4) \quad 0 = \oint_{\partial\mathcal{X}} (\nabla u - \mathbf{m}) \cdot \hat{\mathbf{n}} \, v \, ds - \iint_{\mathcal{X}} (\Delta u - \nabla \cdot \mathbf{m}) v \, dx \, dy \quad \forall v \in C^2(\mathcal{X}).$$

Applying the fundamental lemma of calculus of variations [6, p. 185], we find

$$(4.5a) \quad \Delta u = \nabla \cdot \mathbf{m}, \quad (x, y) \in \mathcal{X},$$

$$(4.5b) \quad \nabla u \cdot \hat{\mathbf{n}} = \mathbf{m} \cdot \hat{\mathbf{n}}, \quad (x, y) \in \partial\mathcal{X}.$$

This is a Neumann problem and has a solution only if the compatibility condition is satisfied. The compatibility condition reads [7]

$$(4.6) \quad \iint_{\mathcal{X}} -\nabla \cdot \mathbf{m} \, dx \, dy + \oint_{\partial\mathcal{X}} \mathbf{m} \cdot \hat{\mathbf{n}} \, ds = 0.$$

By Gauss's theorem we see that it is satisfied indeed. The solution of the Poisson equation with Neumann boundary condition is unique up to a constant. To make the solution unique, we add the constraint

$$(4.7) \quad u(x_i, y_j) = 0$$

for some arbitrarily chosen (x_i, y_j) .

The numerical algorithm is summarized as follows. We discretize the source domain $\mathcal{X}$ and select points on the boundary of the target domain $\mathcal{Y}$. The initial

FIG. 2. “Square to circle” problem: the mapping after 100 iterations on a 101×101 grid. Parameter values are $\alpha = 0.2$ and $N_b = 100$.

guess is given by (2.15). Subsequently, we repeatedly perform the steps given by (2.13a), (2.13b), and (2.13c). The first step is a minimization for $\mathbf{b}(x_i, y_j)$ and is performed pointwise for all grid points $(x_i, y_j) \in \partial\mathcal{X}$ using (3.4). The second step is a minimization procedure for $\mathbf{P}(x_i, y_j)$ and is performed pointwise for all grid points by calculating the global minimizer. The third and last step is a minimization procedure for $\mathbf{m}$ and is performed by solving two Poisson boundary value problems given by (3.43) and (3.44). Finally, after convergence of the iteration (2.13), the function u is computed from $\mathbf{m}$ by solving the Poisson problem (4.5).

5. Numerical results. We examine our algorithm on four test problems: in the first problem we map a square with uniform density into a circle with uniform density, in the second problem we map a circle to a square with uniform density, in the third problem we map a square to target distributions on nonconvex domains, and finally in the fourth problem, we challenge our algorithm to design a special lens mapping a uniform, square parallel beam of light to a projection on a screen of a famous Dutch painting.

5.1. From a square to a circle. In our first test problem, the source domain is given by $\mathcal{X} = [-1, 1]^2$ and the target domain by $\mathcal{Y} = \{(p, q) \in \mathbb{R}^2 \mid p^2 + q^2 \leq 1\}$. We have $f(x, y) = 1/4$ and $g(p, q) = 1/\pi$. We solve the boundary value problem (1.3)–(1.4) for different grid sizes with $\alpha = 0.2$. The resulting mapping after 100 iterations is shown in Figure 2.

To investigate convergence of the iteration procedure (2.13) we introduce the matrix/vector norms

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

for $\mathbf{A} \in [C^1(\mathcal{X})]^{2 \times 2}$, the set of 2×2 matrices of $C^1(\mathcal{X})$ functions, and $\mathbf{a} \in [C^1(\partial\mathcal{X})]^2$, respectively. Let $J_I^n = J_I(\mathbf{m}^n, \mathbf{P}^n)$ and $J_B^n = J_B(\mathbf{m}^n, \mathbf{b}^n)$. Using elementary properties of norms we can show that

$$(5.1a) \quad \left| \sqrt{J_I^{n+1}} - \sqrt{J_I^n} \right| \leq \frac{1}{\sqrt{2}} \left(\|\mathbf{D}\mathbf{m}^{n+1} - \mathbf{D}\mathbf{m}^n\|_{2 \times 2} + \|\mathbf{P}^{n+1} - \mathbf{P}^n\|_{2 \times 2} \right) =: c_I^n,$$

$$(5.1b) \quad \left| \sqrt{J_B^{n+1}} - \sqrt{J_B^n} \right| \leq \frac{1}{\sqrt{2}} \left(\|\mathbf{m}^{n+1} - \mathbf{m}^n\|_2 + \|\mathbf{b}^{n+1} - \mathbf{b}^n\|_2 \right) =: c_B^n,$$

FIG. 3. “Square to circle” problem: convergence history for several grids. Parameter values are $\alpha = 0.2$ and $N_b = 1000$.

i.e., c_I^n is an upper bound for the change in $\sqrt{J_I}$ due to changes in Dm and P . A similar interpretation holds for c_B^n . We have computed J_I , J_B , and the individual error terms in (5.1) for several $N \times N$ grids; see Figure 3. The number of boundary grid points $N_b = 1000$ such that the minimizing procedure for b will not be the limiting factor for accuracy. From this figure we conclude that J_I and J_B stall at certain values, whereas c_I^n and c_B^n keep decreasing for increasing iteration number n . In fact, c_I^n and c_B^n will reach machine epsilon when n is large enough. Therefore, we propose the stopping criterion

$$(5.2) \quad c_I^n \leq 0.1\sqrt{J_I^n} \quad \text{and} \quad c_B^n \leq 0.1\sqrt{J_B^n},$$

implying a relative change in $\sqrt{J_I}$ and $\sqrt{J_B}$ of less than 0.1. The results in Figure 3 are obtained with this criterion. Note that the limit values of J_I and J_B decrease with increasing N and that the number of iterations needed to satisfy (5.2) increases approximately linearly with N .

We also tested the influence of N_b on the convergence of the algorithm. We use a 401×401 grid and plot the values of J_I and J_B as function of the iteration number for different choices of N_b . The results are shown in Figure 4. The value of N_b does not influence the initial convergence of the algorithm, but only the final value

FIG. 4. “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 401×401 grid.

(a) Calculation time for the LU decomposition, the minimization steps (per iteration) for $\mathbf{P}$, $\mathbf{b}$, and $\mathbf{m}$ and the calculation of the function u . We used $N_b = 500$.

(b) Calculation times for the minimizing procedure for $\mathbf{b}$ per iteration as function of the number of boundary points N_b for an 101×101 grid.

FIG. 5. Calculation time as function of N and N_b . The calculations were performed on a laptop with an Intel Core i5 M5202.40 GHz with 4 GB of RAM. The dotted lines are least-squares fits. See also Table 1.

of J after convergence. If N_b is chosen large enough, its effect on this value becomes negligible. Since an increase in N_b does not significantly increase the calculation time, it is recommended to choose N_b sufficiently large.

Furthermore, we analyze the calculation times for the LU decomposition, the three minimization steps performed in each iteration, and the computation of u as a function of N and as a function of the number of boundary points N_b . The results are shown in Figure 5. The calculation times for the LU decomposition and the computation of the function u are approximately 10 times larger than the calculation times for the minimization procedures for $\mathbf{P}$ and $\mathbf{m}$. We expect the calculation times for the minimization procedures for $\mathbf{P}$ and $\mathbf{m}$ (using the LU decomposition) to be linear in the number of grid points and thus quadratic in N , and we expect the calculation time for the minimization procedure for $\mathbf{b}$ to be linear in both N and N_b . The calculation times for the LU decomposition and the linear solver for the calculation of the function u depend on the MATLAB implementation; we expect

TABLE 1

Expected and measured calculation times for the different procedures in the algorithm. For the LU decomposition and the linear solver we have no estimate for the computation time, as it depends strongly on the numerical implementation in MATLAB and the structure of the matrix.

Procedure	Expected calculation time	Relation from fit
LU decomposition		$\propto N^{2.22}$
Minimization for $\mathbf{P}$	$\mathcal{O}(N^2)$	$\propto N^{2.00}$
Minimization for $\mathbf{b}$	$\mathcal{O}(N \cdot N_b)$	$\propto N^{0.94}, \propto N_b^{0.79}$
Minimization for $\mathbf{m}$	$\mathcal{O}(N^2)$	$\propto N^{2.15}$
Linear solver for $\mathbf{u}$		$\propto N^{2.10}$

FIG. 6. “Square to circle” problem: value of J as function of the iteration number for different values of α .

them to be at the very least linear in the number of grid points and thus at least quadratic in N . We performed logarithmic least-squares fits on the data in Figure 5, and the results are presented in Table 1.

The LU decomposition and the computation of $\mathbf{u}$ have to be performed only once. Moreover, the calculation time for $\mathbf{b}$ is rather modest. Therefore, the calculation time per iteration is approximately proportional to N^2 , and consequently, since the number of iterations grows linearly in N , the total calculation time scales roughly with N^3 .

Finally, there is no obvious way to choose an appropriate value for α . Nevertheless, the choice of α strongly influences the performance of the algorithm. We determined an optimal choice for α experimentally. The influence of α on the convergence of the algorithm is shown in Figure 6. A value of α between 0.1 and 0.5 results in good performance. This value is independent of the grid size.

5.2. From a circle to a square. The second test case involves a mapping of a circle to a square. We have $f(x, y) = 1/\pi$ for $x^2 + y^2 \leq 1$ and $f(x, y) = 0$ otherwise. The target domain is the square $[-1, 1]^2$ and we have $g(p, q) = 1/4$; see Figures 7(a), 7(b), and 7(c). Because the target domain is square, we need only four points to model the boundary accurately. In this test case, many more iterations are needed to achieve good convergence. We used 500 iterations and compared the performance of the algorithm for different choices of α . The results are shown in Figure 7(d). Convergence of the mapping was often problematic in the corners. Choosing the right value of α reduces this problem.

5.3. A nonconvex target. In the third problem we test the algorithm on nonconvex target domains. We choose a uniform source distribution on the square $[-1, 1]^2$ with $f(x, y) = 1/4$. The target distribution is also uniform, and the target

FIG. 7. "Circle to square" problem: final mapping after 500 iterations on a 401×401 grid. Convergence is often slow in the corners. Choosing the right value of α reduces this problem.

boundary is defined by

$$(5.3) \quad \rho(\theta) = 1 + C \cos(3\theta), \quad 0 \leq \theta < 2\pi,$$

where $\rho = \rho(\theta)$ is the distance to the origin and where θ is the counterclockwise angle with respect to the p -axis in the (p, q) -plane. We test the algorithm for $C \in \{0.1, 0.2, 0.3, 0.4\}$. We use a 201×201 grid, 1000 points on the boundary, and $\alpha = 0.2$. The mappings after 200 iterations are shown in Figure 8. The value of J after 200 iterations is $4.43 \cdot 10^{-7}$, $8.53 \cdot 10^{-7}$, $5.16 \cdot 10^{-5}$, and $4.82 \cdot 10^{-4}$, respectively. Convergence problems arise for target domains which strongly deviate from a convex shape, but if the shape deviates only moderately, the algorithm performs adequately.

5.4. A Vermeer-lens. Our own interest in the Monge–Ampère equation comes from the field of illumination optics. The aim is to provide numerical algorithms to assist optical engineers in the design of lenses and reflectors for illumination that can transform a parallel beam of light with an arbitrary light distribution into another given arbitrary distribution. Such lenses and reflectors can be used, for example, to create efficient road lighting that illuminates streets uniformly, spots with square, rectangular, or arbitrary-shape output beams, or car headlamps. For this article, we

FIG. 8. "Nonconvex target" problem: mapping after 200 iterations.

FIG. 9. A lens redistributing the light from a parallel beam onto a plane.

will test our algorithm to the limit to design a lens transforming a square uniform parallel beam of light into a target distribution corresponding to a famous Dutch painting.

A schematic drawing of a lens design problem is shown in Figure 9. In this example, we have a parallel beam of light in the z -direction with given radiant exitance

FIG. 10. The painting “Girl with a Pearl Earring” by Johannes Vermeer. Note that the domain of the target distribution $g(p, q)$ is not convex.

$M(x, y)$ [W/m²] and an irradiance on the plane $\mathcal{P}$ given by $L(\xi, \eta)$ [W/m²]. The light from the parallel beam is redistributed by a lens of which the first surface is described by the function $z = u(x, y)$ and the second surface is flat. The goal is to find the lens surface $z = u(x, y)$ that gives for the specified radiant exitance $M(x, y)$ of the parallel beam the desired irradiance $L(\xi, \eta)$ on $\mathcal{P}$. Using conservation of energy and the law of refraction, it can be shown that $u(x, y)$ is the convex solution of the Monge–Ampère equation (1.3) with boundary condition (1.4), where $f(x, y) = M(x, y)$ and the function $g(p, q)$ is as plotted in Figure 10(b) [20, p. 78–88]. The target distribution $g(p, q)$ is a deformation of the painting “Girl with a Pearl Earring” by the famous Dutch painter Johannes Vermeer, as shown in Figure 10. The painting is converted to grayscale and the grayscale values are used as an irradiance pattern. Because the target distribution contains many details and its domain is just not convex, it provides a challenging test for our algorithm.

To avoid division by 0, we increase irradiances of less than 5% of the maximum value to 5% of the maximum value. The target plane is located at distance $d = 100$ from the origin, and the painting has a width of 53.2 and a height of 63 (all in arbitrary units). The resulting target distribution $g(p, q)$ is plotted in Figure 10(b). As source emittance we use a uniform square parallel beam of light, so $f(x, y) = 1/4$ on the domain $[-1, 1]^2$. We calculate the lens surface $z = u(x, y)$ on an 801×801 grid with 1996 points to discretize the boundary (500 at each side of the painting) and use $\alpha = 0.2$. The values of J_I and J_B as the function of the iteration number are plotted in Figure 11(a). It can be seen that the value of J may increase during iteration. This happens during the minimization procedure for $\mathbf{P}$. The reason is that the mapping $\mathbf{m}$ has changed, and thus the set $\mathcal{P}(\mathbf{m})$ has changed; therefore we may have

$$(5.4) \quad \min_{\mathbf{P} \in \mathcal{P}(\mathbf{m}^{n+1})} J_I(\mathbf{m}^{n+1}, \mathbf{P}) > \min_{\mathbf{P} \in \mathcal{P}(\mathbf{m}^n)} J_I(\mathbf{m}^n, \mathbf{P}).$$

The final mapping resulting from the iteration process is shown in Figure 11(b). The target distribution in 10(b) can be clearly recognized. A denser grid corresponds to a higher target density and therefore to lighter colors. The function $u(x, y)$ representing the lens surface, which is computed from this mapping, is shown in Figure 12.

FIG. 11. The mapping for the Vermeer-lens and its convergence history.

FIG. 12. The function $u(x, y)$ representing the lens surface.

We tested the calculated surface using the optical simulation software LightTools [18]. In this program we build a three-dimensional model of the lens, a square surface emitting a parallel beam of light, and a target plane $\mathcal{P}$. The surface emitting the light and the lens in the simulation software are shown in Figure 13. The software shoots 5 million evenly distributed rays randomly from a square plane through the lens and returns the distribution of these rays over the target plane. This distribution is plotted in Figure 14.

The experiment shows that the algorithm can be used to calculate very detailed surfaces. The algorithm is very memory-efficient so the calculation could be performed on a laptop with only 4 GB of RAM. The optical simulation reproduces the painting in great detail. This shows that the algorithm can be used to design advanced lenses for illumination converting arbitrary parallel beams of light into arbitrary target distributions.

6. Discussion and conclusions. We developed a new method to compute the solution of the optimal transport problem and the corresponding convex solution of the

FIG. 13. *Ray-tracing the Vermeer-lens. The screenshot shows the light source (left) and the lens (middle). The target screen is located at a large distance to the right.*

FIG. 14. *The ray-tracing results for the Vermeer-lens.*

Monge–Ampère equation with the transport boundary condition. Numerical methods for optimal transport and the Monge–Ampère equation with transport boundary conditions are very scarce. To the best of our knowledge, the only other methods were published only recently [3, 4, 9, 10, 11, 12, 16, 17]. These methods compute a convex solution of the Monge–Ampère equation using a wide-stencil scheme that enforces convexity.

We implemented this other method as well and published an article with the results [21]. Both algorithms are able to calculate detailed solutions of the Monge–Ampère equation on large grids. The memory requirements of the least-squares algorithm is lower, which allows calculation on larger grids than does the wide-stencil method. The wide-stencil scheme requires fewer iterations but more calculation time per iteration. On larger grids, the least-squares method is somewhat faster, because the calculation time for the different steps during the iterations scale approximately linearly with the number of grid points, and the LU decomposition and linear solve for u need to be performed only once. If $f = 0$ on part of the domain, the wide-stencil scheme performs better: the least-squares algorithm converges slowly for these types of problems. If the target domain is not convex, the least-squares algorithm has a clear

advantage: as long as the domain does not deviate too much from a convex domain, it shows good performance. For the wide-stencil method in [21] we computed solutions for nonconvex target domains by swapping the source and target distributions: we retrieved the solution of the original problem using an additional Legendre–Fenchel transform. This introduces additional calculation time and artifacts in the solution and is possible only if the source domain is convex. It should be noted, however, that also the wide-stencil scheme by Froese [9] can handle mildly nonconvex targets.

The least-squares method has shown good performance on complicated optimal transport problems. The ability to solve transport problems with nonconvex target domains makes our new method a valuable addition to existing methods.

Appendix A. Solution of the quartic equation (3.20a) when $f > 0$. We find the four (possibly complex) solutions of (3.20a) using Ferrari's method [22, p. 32]. The idea is to rewrite the quartic equation as two quadratic equations. We assume $f > 0$ and thus $a_4 > 0$, divide by a_4 , and rewrite the equation to

$$(A.1) \quad \left(\lambda^2 + \frac{a_2}{2a_4} \right)^2 = -\frac{a_1}{a_4} \lambda - \frac{a_0}{a_4} + \left(\frac{a_2}{2a_4} \right)^2.$$

Adding an arbitrary variable y to the left-hand side under the square, and adding the resulting extra term to the right-hand side, we get

$$(A.2) \quad \left(\lambda^2 + \frac{a_2}{2a_4} + y \right)^2 = 2y\lambda^2 - \frac{a_1}{a_4} \lambda - \frac{a_0}{a_4} + \left(\frac{a_2}{2a_4} \right)^2 + \frac{a_2}{a_4} y + y^2.$$

Next, we attempt to write the right-hand side as a perfect square, such that we get the following equation:

$$(A.3) \quad \left(\lambda^2 + \frac{a_2}{2a_4} + y \right)^2 = \left(\sqrt{2y} \lambda - \frac{a_1}{2a_4 \sqrt{2y}} \right)^2.$$

Equating the right-hand sides of (A.2) and (A.3), we find this is possible only if y is a solution of the cubic equation

$$(A.4) \quad y^3 + b_2 y^2 + b_1 y + b_0 = 0$$

with coefficients

$$(A.5) \quad b_2 = \frac{a_2}{a_4}, \quad b_1 = \frac{1}{4} \left(\frac{a_2}{a_4} \right)^2 - \frac{a_0}{a_4}, \quad b_0 = -\frac{1}{8} \left(\frac{a_1}{a_4} \right)^2.$$

One solution for y is given by [19, p. 179]

$$(A.6a) \quad Q = \frac{b_2^2 - 3b_1}{9}, \quad R = \frac{2b_2^3 - 9b_1b_2 + 27b_0}{54},$$

$$(A.6b) \quad A = -\operatorname{sgn}(R) \left(|R| + \sqrt{R^2 - Q^3} \right)^{1/3},$$

$$(A.6c) \quad y = A + \frac{Q}{A} - \frac{b_2}{3}.$$

If $Q = R = 0$, then $A = 0$ and we have division by zero in (A.6c). In this case, it is known that the cubic equation has a triple root given by $y = -b_2/3$. Using the value of y given by (A.6c) or $y = -b_2/3$, we find from (A.3)

$$(A.7) \quad \lambda^2 + \frac{a_2}{2a_4} + y = \pm \left(\sqrt{2y} \lambda - \frac{a_1}{2a_4 \sqrt{2y}} \right).$$

These are two quadratic equations for λ . Solving both, we find the following four roots of the quartic equation:

$$(A.8a) \quad \lambda_1 = -\sqrt{\frac{y}{2}} + \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} + \frac{a_1}{2a_4\sqrt{2y}}},$$

$$(A.8b) \quad \lambda_2 = -\sqrt{\frac{y}{2}} - \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} + \frac{a_1}{2a_4\sqrt{2y}}},$$

$$(A.8c) \quad \lambda_3 = \sqrt{\frac{y}{2}} + \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} - \frac{a_1}{2a_4\sqrt{2y}}},$$

$$(A.8d) \quad \lambda_4 = \sqrt{\frac{y}{2}} - \sqrt{-\frac{y}{2} - \frac{a_2}{2a_4} - \frac{a_1}{2a_4\sqrt{2y}}}.$$

It can be shown that (3.20a) has at least two real roots [20, p. 138]. Furthermore, we have division by 0 in (A.8) if $y = 0$. By substituting $y = 0$ in (A.4), we find that this happens only when $a_1 = 0$, and thus $d_{11} = d_{22} = d_S = 0$, corresponding to $\mathbf{D}_S = \mathbf{O}$, which is treated as a special case.

REFERENCES

- [1] S. ANGENT, S. HAKER, AND A. TANNENBAUM, *Minimizing flows for the Monge–Kantorovich problem*, SIAM J. Math. Anal., 35 (2003), pp. 61–97.
- [2] J. BENAMOU AND Y. BRENIER, *A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem*, Numer. Math, 84 (2000), pp. 375–393.
- [3] J. BENAMOU, B. FROESE, AND A. OBERMAN, *A viscosity solution approach to the Monge–Ampère formulation of the optimal transportation problem*, arXiv:1208.4873.
- [4] J. BENAMOU, B. FROESE, AND A. OBERMAN, *Numerical solution of the optimal transportation problem using the Monge–Ampère equation*, J. Comput. Phys., 260 (2014), pp. 107–126.
- [5] A. CABOUSSAT, R. GLOWINSKI, AND D. SORESENSEN, *A least-squares method for the numerical solution of the Dirichlet problem for the elliptic Monge–Ampère equation in dimension two*, ESAIM Control Optim. Calc. Var., 19 (2013), pp. 780–810.
- [6] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. 1, Wiley, New York, 1989.
- [7] H. ELMAN, D. SILVESTER, AND A. WATHEN, *Finite Elements and Fast Iterative Solvers*, Oxford University Press, New York, 2005.
- [8] L. EVANS, *Partial differential equations and Monge–Ampère mass transfer*, in Current Developments in Mathematics, Cambridge MA (1997), International Press, Boston, 1999, pp. 65–126.
- [9] B. FROESE, *A numerical method for the elliptic Monge–Ampère equation with transport boundary conditions*, SIAM J. Sci. Comput., 34 (2012), pp. A1432–A1459.
- [10] B. FROESE AND A. OBERMAN, *Convergent finite difference solvers for viscosity solutions of the elliptic Monge–Ampère equation in dimensions two and higher*, SIAM J. Numer. Anal., 49 (2011), pp. 1692–1714.
- [11] B. FROESE AND A. OBERMAN, *Fast finite difference solvers for singular solutions of the elliptic Monge–Ampère equation*, J. Comput. Phys., 230 (2011), pp. 818–834.
- [12] B. FROESE AND A. OBERMAN, *Convergent filtered schemes for the Monge–Ampère partial differential equation*, SIAM J. Numer. Anal., 51 (2013), pp. 423–444.
- [13] E. HABER, T. REHMAN, AND A. TANNENBAUM, *An efficient numerical method for the solution of the L_2 optimal mass transfer problem.*, SIAM J. Sci. Comput., 32 (2010), pp. 197–211.
- [14] J. MARSDEN AND A. TROMBA, *Vector Calculus*, 4th ed., W. H. Freeman, New York, 1996.
- [15] R. MATTHEIJ, S. RIENSTRA, AND J. TEN THIJSE BOONKAMP, *Partial Differential Equations*, SIAM, Philadelphia, 2005.
- [16] A. OBERMAN, *Convergent difference schemes for degenerate elliptic and parabolic equations: Hamilton–Jacobi equations and free boundary problems*, SIAM J. Numer. Anal., 44 (2006), pp. 879–895.
- [17] A. OBERMAN, *Wide stencil finite difference schemes for the elliptic Monge–Ampère equation and functions of the eigenvalues of the Hessian*, Discrete Contin. Dyn. Syst. Ser. B, 10 (2008), pp. 221–238.

- [18] ORA, *Ora Lighttools*, <http://optics.synopsys.com/lighttools> (accessed August 19, 2014).
- [19] W. PRESS, B. FLANNERY, S. TEUKOLSKY, AND W. VETTERLING, *Numerical Recipes in Fortran77*, 2nd ed., Cambridge University Press, Cambridge, UK, 1996.
- [20] C. PRINS, *Inverse Methods for Illumination Optics*, Ph.D. thesis, Eindhoven University of Technology, Eindhoven, the Netherlands, 2014.
- [21] C. PRINS, J. TEN THIJE BOONKKAMP, J. VAN ROOSMALEN, T. TUKKER, AND W. IJZERMAN, *A Monge-Ampère solver for free-form reflector design*, SIAM J. Sci. Comput., 36 (2014), pp. B640–B660.
- [22] J. TIGNOL, *Galois's Theory of Algebraic Equations*, Longman Scientific and Technical, Harlow, UK, 1988.
- [23] N. TRUDINGER AND X. WANG, *On the second boundary value problem for Monge-Ampère type equations and optimal transportation*, Ann. Sc. Norm. Super. Pisa Cl. Sci., 8 (2009), pp. 143–174.
- [24] C. VILLANI, *Topics in Optimal Transportation*, AMS, Providence, RI, 2003.