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

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computer Science

CASA-Report 14-29
September 2014

A least-squares method for optimal transport using the monge-ampère equation

by

C.R. Prins, J.H.M. ten Thije Boonkkamp, W.L. IJzerman, T.W. Tukker

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

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

C. R. PRINS*, 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 mass 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 boundary conditions. It scales well with the grid size and has the additional advantage that the target domain may be non-convex. We present the method and several numerical experiments.

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 mass transport problem with quadratic cost function can be stated as follows. Let $f : \mathcal{X} \to [0, \infty)$ and $g : \mathcal{Y} \to (0, \infty)$ 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 $\boldsymbol{m} : \mathcal{X} \mapsto \mathcal{Y}$, minimizing the transportation cost

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

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

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

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, p.125-126] states that such an optimal mapping is the (almost everywhere) unique gradient of a convex function. Let ∇u denote this gradient. Substituting $\boldsymbol{m} = \nabla u$ in (1.2), we find by a change of variables

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

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

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

The solution of the optimal transportation 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).

*CASA, Eindhoven University of Technology, PO Box 513 5600 MB Eindhoven, The Netherlands
†Philips Lighting, Eindhoven, The Netherlands
‡Philips Research, Eindhoven, The Netherlands

Our interest in optimal mass 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 an certain 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 from this application.

Up to recently, very few efficient numerical methods for optimal mass transport were known. A numerical method for the related Monge-Ampère equation was introduced by Froese, Benamou and Oberman [3, 4, 9–12, 16, 17], using finite differences, a wide-stencil scheme and an innovative discretizaton of the boundary condition. We implemented and tested this algorithm [21] on several reflector design problems. 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 mass transport problem was developed by Benamou & 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 mass 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 et al. [5]. Their method numerically solves the Dirichlet problem of the elliptic Monge-Ampère equation, given by

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

The authors minimize the functional

$$J(\phi, \boldsymbol{P}) = \frac{1}{2} \iint_{\mathcal{X}} \left|\left|D^2\phi - \boldsymbol{P}\right|\right|^2 \, \mathrm{d}x \, \mathrm{d}y, \tag{1.6}$$

over the set of real symmetric matrices $\boldsymbol{P}$ such that $\det \boldsymbol{P} = f(x, y)$ and functions ϕ satisfying the Dirichlet boundary condition. The minimization is performed alternatingly over $\boldsymbol{P}$ and ϕ. The minimization over $\boldsymbol{P}$ with fixed ϕ is a nonlinear problem which is solved pointwise for each gridpoint using Newton iteration. The minimization over ϕ with fixed $\boldsymbol{P}$ 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 et al. in three ways. First, we have an extra term in the functional J to account for the transport boundary condition. Secondly, we found a method to solve the minimization over $\boldsymbol{P}$ analytically, instead of using Newton iteration. Thirdly, instead of u we compute the mapping function $\boldsymbol{m}$. The procedure for $\boldsymbol{m}$ boils down to solving two separate Poisson problems for the components m_1 and m_2 of $\boldsymbol{m}$, each iteration. 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 $\boldsymbol{m}$. We give several test results in Section 5 which we discuss in Section 6.

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

$$\iint_{\mathcal{X}} f(x, y) \, \mathrm{d}x \, \mathrm{d}y = \iint_{\mathcal{Y}} g(p, q) \, \mathrm{d}p \, \mathrm{d}q. \tag{2.1}$$

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 $\boldsymbol{m}$.

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

$$D\boldsymbol{m} = \begin{pmatrix} \frac{\partial m_1}{\partial x} & \frac{\partial m_1}{\partial y} \\ \frac{\partial m_2}{\partial x} & \frac{\partial m_2}{\partial y} \end{pmatrix}, \tag{2.2}$$

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

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

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

$$J_{\mathrm{I}}(\boldsymbol{m}, \boldsymbol{P}) = \frac{1}{2} \iint_{\mathcal{X}} ||D\boldsymbol{m} - \boldsymbol{P}||^2 \, \mathrm{d}x \, \mathrm{d}y. \tag{2.4}$$

The norm used in this functional is called the Fröbenius norm. Let $\boldsymbol{P} : \boldsymbol{Q}$ denote the Fröbenius inner product of the matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$, defined by

$$\boldsymbol{P} : \boldsymbol{Q} = \sum_{i,j} (\boldsymbol{P})_{i,j} (\boldsymbol{Q})_{i,j}, \tag{2.5}$$

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

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

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

$$\mathcal{P}(\boldsymbol{m}) = \left\{ \boldsymbol{P} \in [C^1(\mathcal{X})]^{2\times 2}_{\mathrm{Sym}} \,\middle|\, \det\left(\boldsymbol{P}(x, y)\right) = \frac{f(x, y)}{g(\boldsymbol{m}(x, y))} \right\}. \tag{2.7}$$

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

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

$$\boldsymbol{m}(\mathcal{X}) = \mathcal{Y}. \tag{2.8}$$

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

$$\boldsymbol{m}(\partial\mathcal{X}) = \partial\mathcal{Y}. \tag{2.9}$$

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

$$J_{\mathrm{B}}(\boldsymbol{m}, \boldsymbol{b}) = \frac{1}{2} \oint_{\partial\mathcal{X}} |\boldsymbol{m} - \boldsymbol{b}|^2 \, \mathrm{d}s. \tag{2.10}$$

We minimize this function over $\boldsymbol{b}$ from the set

$$\mathcal{B} = \left\{ \boldsymbol{b} \in [C(\partial\mathcal{X})]^2 \;\middle|\; \boldsymbol{b}(\boldsymbol{x}) \in \partial\mathcal{Y} \quad \forall \boldsymbol{x} \in \partial\mathcal{X} \right\}. \tag{2.11}$$

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

We combine the functional J_{I} for the interior and J_{B} for the boundary by a weighted average:

$$J(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b}) = (1 - \alpha)\, J_{\mathrm{B}}(\boldsymbol{m}, \boldsymbol{b}) + \alpha\, J_{\mathrm{I}}(\boldsymbol{m}, \boldsymbol{P}). \tag{2.12}$$

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

$$\boldsymbol{b}^{n+1} = \operatorname*{arg\,min}_{\boldsymbol{b} \in \mathcal{B}} J_{\mathrm{B}}(\boldsymbol{m}^n, \boldsymbol{b}), \tag{2.13a}$$

$$\boldsymbol{P}^{n+1} = \operatorname*{arg\,min}_{\boldsymbol{P} \in \mathcal{P}(\boldsymbol{m}^n)} J_{\mathrm{I}}(\boldsymbol{m}^n, \boldsymbol{P}), \tag{2.13b}$$

$$\boldsymbol{m}^{n+1} = \operatorname*{arg\,min}_{\boldsymbol{m} \in \mathcal{V}} J(\boldsymbol{m}, \boldsymbol{P}^{n+1}, \boldsymbol{b}^{n+1}). \tag{2.13c}$$

This procedure is continued the value of $J(\boldsymbol{m}^n, \boldsymbol{P}^n, \boldsymbol{b}^n)$ stalls.

We solve the optimal mass 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

$$x_i = a_{\min} + (i-1)\, h_x, \qquad h_x = \frac{a_{\max} - a_{\min}}{N_x - 1}, \;\; i = 1, \ldots, N_x, \tag{2.14a}$$

$$y_j = b_{\min} + (j-1)\, h_y, \qquad h_y = \frac{b_{\max} - b_{\min}}{N_y - 1}, \;\; j = 1, \ldots, N_y. \tag{2.14b}$$

We initialize our minimization by constructing an initial guess $\boldsymbol{m}^0$ which maps the source area $\mathcal{X}$ to a bounding box of the target area $\mathcal{Y}$. Without loss of generality we assume the source area 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

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

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

In the next paragraphs, we elaborate each of the three minimizations (2.13a), (2.13b) and (2.13c).

3. Minimizing procedures for $\boldsymbol{b}$, $\boldsymbol{P}$ and $\boldsymbol{m}$. Using the initial guess $\boldsymbol{m}^0$, 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 $\boldsymbol{b}$. In this section we perform the minimization step (2.13a). We assume $\boldsymbol{m}$ fixed, and minimize the functional $J_{\mathrm{B}}(\boldsymbol{m}, \boldsymbol{b})$ over $\boldsymbol{b} \in \mathcal{B}$. For simplicity of notation we drop the indices n and $n+1$. The minimization can be performed point-wise, because no derivative of $\boldsymbol{b}$ with respect to x or y appears in the functional $J_{\mathrm{B}}(\boldsymbol{m}, \boldsymbol{b})$. For each gridpoint $(x_i, y_j) \in \partial\mathcal{X}$, let $\boldsymbol{m}_{i,j}$ denote the value of the mapping. We perform the minimization

$$\min_{\boldsymbol{b}_{i,j} \in \partial\mathcal{Y}} |\boldsymbol{m}_{i,j} - \boldsymbol{b}_{i,j}|^2 . \tag{3.1}$$

We discretize the boundary of $\mathcal{Y}$ using points $\boldsymbol{z}_k \in \partial\mathcal{Y}$, $(k = 1, \ldots, N_{\mathrm{b}})$ with increasing index along the boundary, and define $\boldsymbol{z}_{N_{\mathrm{b}}+1} = \boldsymbol{z}_1$. We connect adjacent points by line segments $(\boldsymbol{z}_k, \boldsymbol{z}_{k+1})$ and determine the closest point to $\boldsymbol{m}_{i,j}$ to all of the line segments $(\boldsymbol{z}_k, \boldsymbol{z}_{k+1})$.

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

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

$$\boldsymbol{m}_{i,j}^k = \boldsymbol{z}_k + t_k \, (\boldsymbol{z}_{k+1} - \boldsymbol{z}_k), \tag{3.2b}$$

see also Figure 3.1. If $t_k \in [0, 1]$, the projected point is on the line segment, and the nearest point to $\boldsymbol{m}_{i,j}$ is given by $\boldsymbol{m}_{i,j}^k$. If $t_k < 0$, the projected point is not on the line segment and the nearest point is given by $\boldsymbol{z}_k$, and if $t_k > 1$, the nearest point is given by $\boldsymbol{z}_{k+1}$. Thus, the nearest point on the line segment $(\boldsymbol{z}_k, \boldsymbol{z}_{k+1})$ is given by

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

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

$$\boldsymbol{b}_{i,j} = \operatorname*{arg\,min}_{\widetilde{\boldsymbol{m}}_{i,j}^k} \left\{ \left| \widetilde{\boldsymbol{m}}_{i,j}^k - \boldsymbol{m}_{i,j} \right|^2 \right\}. \tag{3.4}$$

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

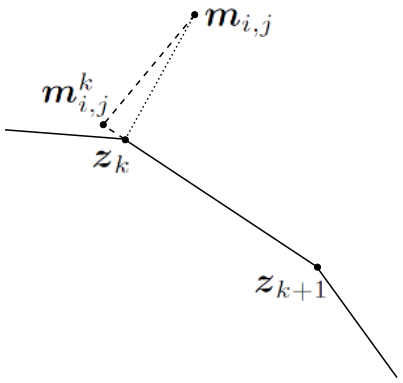


Figure 3.1: Calculation of the distance of $\boldsymbol{m}_{i,j}$ to a line segment $(\boldsymbol{z}_k, \boldsymbol{z}_{k+1})$. The point $\boldsymbol{m}_{i,j}^k$ is the projection of $\boldsymbol{m}_{i,j}$ on the line through $\boldsymbol{z}_k$ and $\boldsymbol{z}_{k+1}$. In this case we find $t_k < 0$ and the nearest point on the line segment to $\boldsymbol{m}_{i,j}$ is given by $\boldsymbol{b}_{i,j} = \boldsymbol{z}_k$.

3.2. Minimizing procedure for $\boldsymbol{P}$. In this section we perform the minimization step (2.13b). We assume $\boldsymbol{m}$ fixed and minimize $J_{\mathrm{I}}(\boldsymbol{m}, \boldsymbol{P})$ defined in (2.4) over the matrices $\boldsymbol{P} \in \mathcal{P}(\boldsymbol{m})$ under the condition

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

Since the integrand of $J_{\mathrm{I}}(\boldsymbol{m}, \boldsymbol{P})$ does not contain derivatives of $\boldsymbol{P}$, the minimization can be performed pointwise. Define

$$d_{11} = \frac{\partial m_1}{\partial x},\ d_{12} = \frac{\partial m_1}{\partial y},\ d_{21} = \frac{\partial m_2}{\partial x},\ d_{22} = \frac{\partial m_2}{\partial y}, \tag{3.6}$$

and

$$\boldsymbol{D} = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix}, \qquad \boldsymbol{P} = \begin{pmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{pmatrix}. \tag{3.7}$$

Subsequently, define the function

$$H(p_{11}, p_{12}, p_{22}) = \frac{1}{2} \left|\left| \boldsymbol{P} - \boldsymbol{D} \right|\right|^2. \tag{3.8}$$

We have for each $(x, y) \in \mathcal{X}$ a quadratic minimization problem:

$$\underset{(p_{11},p_{12},p_{22})\in\mathbb{R}^3}{\arg\min} \left\{ H(p_{11}, p_{12}, p_{22}) \;\middle|\; p_{11}\, p_{22} - p_{12}^2 = \frac{f(x,y)}{g\left(\boldsymbol{m}(x,y)\right)} \right\}. \tag{3.9}$$

This problem can be solved analytically. First we slightly simplify the minimization problem. Let $d_{\mathrm{S}} := \frac{1}{2}\left(d_{12} + d_{21}\right)$, and define

$$\boldsymbol{D}_{\mathrm{S}} = \begin{pmatrix} d_{11} & d_{\mathrm{S}} \\ d_{\mathrm{S}} & d_{22} \end{pmatrix}. \tag{3.10}$$

Subsequently we introduce

$$H_{\mathrm{S}}(p_{11}, p_{12}, p_{22}) = \frac{1}{2} \left|\left| \boldsymbol{P} - \boldsymbol{D}_{\mathrm{S}} \right|\right|^2 = H(p_{11}, p_{12}, p_{22}) - \frac{1}{4}(d_{12} - d_{21})^2. \tag{3.11}$$

Because $(d_{12} - d_{21})^2$ is constant with respect to p_{11}, p_{12} and p_{22}, and because we are only interested in the minimizer (p_{11}, p_{12}, p_{22}) and not in the value of $H(p_{11}, p_{12}, p_{22})$, the minimization problem (3.9) is equivalent to

$$\underset{(p_{11},p_{12},p_{22})\in\mathbb{R}^3}{\arg\min} \left\{ H_{\mathrm{S}}(p_{11}, p_{12}, p_{22}) \;\middle|\; p_{11}\, p_{22} - p_{12}^2 = \frac{f(x,y)}{g\left(\boldsymbol{m}(x,y)\right)} \right\}. \tag{3.12}$$

Given d_{11}, d_{S}, d_{22} and f/g, we distinguish five different cases. In each case, we find at least one and at most four critical points $(p_{11}, p_{12}, p_{22}) \in \mathbb{R}^3$ of (3.12), which we call possible minimizers, satisfying the nonlinear constraint $p_{11}\, p_{22} - p_{12}^2 = f/g$. These possible minimizers are local minima, maxima or saddle points of the minimization problem (3.12) and the global minimizer is one of these. We determine the global minimizer by substituting each possible minimizer in the function $H_{\mathrm{S}}(p_{11}, p_{12}, p_{22})$. More details are given in [20].

The possible minimizers of (3.12) are given by the critical points of the Lagrange function:

$$\Lambda(p_{11}, p_{12}, p_{22}, \lambda) = \frac{1}{2}\, ||\boldsymbol{P} - \boldsymbol{D}_{\mathrm{S}}||^2 + \lambda \left(\det\left(\boldsymbol{P}\right) - \frac{f}{g} \right), \tag{3.13}$$

where λ is the Lagrange multiplier. Setting the partial derivatives with respect to p_{11}, p_{12}, p_{22} and λ to 0 we find the critical points. This gives the following equations

$$p_{11} + \lambda\, p_{22} = d_{11}, \tag{3.14a}$$
$$(1 - \lambda)\, p_{12} = d_{\mathrm{S}}, \tag{3.14b}$$
$$\lambda\, p_{11} + p_{22} = d_{22}, \tag{3.14c}$$
$$p_{11}\, p_{22} - p_{12}^2 = \frac{f}{g}. \tag{3.14d}$$

The system (3.14a) - (3.14c) is linear in p_{11}, p_{12} and p_{22}, and may be inverted if $\lambda \neq \pm 1$. From the two implications

$$\lambda = 1 \Rightarrow \left[d_{11} = d_{22} \text{ and } d_{\mathrm{S}} = 0 \right], \tag{3.15a}$$
$$\lambda = -1 \Rightarrow \left[d_{11} = -d_{22} \right], \tag{3.15b}$$

we conclude

$$\left[d_{11} \neq d_{22} \text{ or } d_{\mathrm{S}} \neq 0 \right] \Rightarrow \lambda \neq 1, \tag{3.16a}$$
$$\left[d_{11} \neq -d_{22} \right] \Rightarrow \lambda \neq -1. \tag{3.16b}$$

Therefore, if (3.16a) and (3.16b) apply, we can safely assume that $\lambda \neq \pm 1$ and invert (3.14). If $[d_{11} = d_{22}$ and $d_{\mathrm{S}} = 0]$ or $[d_{11} = -d_{22}]$, we will use a different method to solve the minimization problem (3.12).

First we will assume that (3.16a) and (3.16b) apply. We find:

$$p_{11} = \frac{\lambda\, d_{22} - d_{11}}{\lambda^2 - 1}, \tag{3.17a}$$
$$p_{12} = \frac{d_{\mathrm{S}}}{1 - \lambda}, \tag{3.17b}$$
$$p_{22} = \frac{\lambda\, d_{11} - d_{22}}{\lambda^2 - 1}. \tag{3.17c}$$

Substituting these expressions in (3.14d) gives the quartic equation

$$a_4\lambda^4 + a_2\lambda^2 + a_1\lambda + a_0 = 0, \tag{3.18}$$

with coefficients given by

$$a_4 = f/g \geq 0, \tag{3.19a}$$
$$a_2 = -2\,f/g - \det(\boldsymbol{D}_{\mathrm{S}}), \tag{3.19b}$$
$$a_1 = ||\boldsymbol{D}_{\mathrm{S}}||^2 \geq 0, \tag{3.19c}$$
$$a_0 = f/g - \det(\boldsymbol{D}_{\mathrm{S}}). \tag{3.19d}$$

First we will show how to solve (3.18), subsequently we give the minimizers when $[d_{11} = d_{22}$ and $d_{\mathrm{S}} = 0]$ or $[d_{11} = -d_{22}]$.

Solution of (3.18) **when** $f > 0$**.** We find the four (possibly complex) solutions of equation (3.18) using Ferrari's method [22, p.32]. The idea is to rewrite the quartic equation as two quadratic equations. First we assume $f > 0$ and thus $a_4 > 0$, divide by a_4, and rewrite the equation to

$$\left(\lambda^2 + \frac{a_2}{2\,a_4}\right)^2 = -\frac{a_1}{a_4}\lambda - \frac{a_0}{a_4} + \left(\frac{a_2}{2\,a_4}\right)^2. \tag{3.20}$$

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

$$\left(\lambda^2 + \frac{a_2}{2\,a_4} + y\right)^2 = 2y\lambda^2 - \frac{a_1}{a_4}\lambda - \frac{a_0}{a_4} + \left(\frac{a_2}{2\,a_4}\right)^2 + \frac{a_2}{a_4}y + y^2. \tag{3.21}$$

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

$$\left(\lambda^2 + \frac{a_2}{2\,a_4} + y\right)^2 = \left(\sqrt{2\,y}\lambda - \frac{a_1}{2\,a_4\,\sqrt{2\,y}}\right)^2. \tag{3.22}$$

Equating the right-hand sides of (3.21) and (3.22), we find this is only possible if y is a solution of the cubic equation

$$y^3 + b_2\,y^2 + b_1\,y + b_0 = 0, \tag{3.23}$$

with coefficients

$$b_2 = \frac{a_2}{a_4}, \quad b_1 = \frac{1}{4}\left(\frac{a_2}{a_4}\right)^2 - \frac{a_0}{a_4}, \quad b_0 = -\frac{1}{8}\left(\frac{a_1}{a_4}\right)^2. \tag{3.24}$$

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

$$Q = \frac{b_2^2 - 3\,b_1}{9}, \quad R = \frac{2\,b_2^3 - 9\,b_1\,b_2 + 27\,b_0}{54}, \tag{3.25a}$$
$$A = -\mathrm{sgn}(R)\left(|R| + \sqrt{R^2 - Q^3}\right)^{1/3}, \tag{3.25b}$$
$$y = A + \frac{Q}{A} - \frac{b_2}{3}. \tag{3.25c}$$

If $Q = R = 0$, then $A = 0$ and we have division by zero in (3.25c). In this case, it is known that the cubic equation has a triple root given by $y = -\frac{b_2}{3}$. Using the value of y given by (3.25c) or $y = -\frac{b_2}{3}$, we find from (3.22)

$$\lambda^2 + \frac{a_2}{2\,a_4} + y = \pm\left(\sqrt{2\,y}\lambda - \frac{a_1}{2\,a_4\,\sqrt{2\,y}}\right). \tag{3.26}$$

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

$$\lambda_1 = -\sqrt{\frac{y}{2}} + \sqrt{-\frac{y}{2} - \frac{a_2}{2\,a_4} + \frac{a_1}{2\,a_4\sqrt{2\,y}}}, \tag{3.27a}$$

$$\lambda_2 = -\sqrt{\frac{y}{2}} - \sqrt{-\frac{y}{2} - \frac{a_2}{2\,a_4} + \frac{a_1}{2\,a_4\sqrt{2\,y}}}, \tag{3.27b}$$

$$\lambda_3 = \sqrt{\frac{y}{2}} + \sqrt{-\frac{y}{2} - \frac{a_2}{2\,a_4} - \frac{a_1}{2\,a_4\sqrt{2\,y}}}, \tag{3.27c}$$

$$\lambda_4 = \sqrt{\frac{y}{2}} - \sqrt{-\frac{y}{2} - \frac{a_2}{2\,a_4} - \frac{a_1}{2\,a_4\sqrt{2\,y}}}. \tag{3.27d}$$

It can be shown that (3.18) has at least two real roots [20, p.138]. Furthermore, we have division by 0 in (3.27) if $y = 0$. By substituting $y = 0$ in (3.23), we find that this only happens when $a_1 = 0$, and thus in case $d_{11} = d_{22} = d_{\mathrm{S}} = 0$, which corresponds to the situation (3.15a) which we will discuss later. The real roots of the quartic equation given by (3.27) are substituted in (3.17) and (3.8), yielding at least two and at most four critical points of the Lagrangian $\Lambda(p_{11}, p_{12}, p_{22})$ given by (3.13), and thus at least two and at most four possible minimizers of (3.12).

Solution of (3.18) **when** $f = 0$**.** If $f = 0$, i.e., when the source density is zero, (3.18) reduces to a quadratic equation because $a_4 = 0$. The roots are given by

$$\lambda = \frac{-a_1 \pm \sqrt{a_1^2 - 4\,a_2\,a_0}}{2\,a_2}. \tag{3.28}$$

The discriminant of this quadratic equation is always positive. We can verify this by substutiting (3.19) in the discriminant:

$$a_1^2 - 4\,a_2\,a_0 = \left(d_{11}^2 - d_{22}^2\right)^2 + 4\,d_{\mathrm{S}}^2\,(d_{11} + d_{22})^2 \geq 0. \tag{3.29}$$

Equality is only possible in the two exceptional cases described below, therefore we always have two distinct real solutions. Substituting these values of λ in (3.17), we find two possible minimizers.

Solution of (3.18) **when** $f = 0$ **and** $a_2 = 0$**.** If $f = 0$ and $a_2 = 0$, we have division by zero in (3.28). This occurs when $d_{11}d_{22} - d_{\mathrm{S}}^2 = 0$, and we find from (3.19) that $a_0 = 0$ as well. The only solution to (3.18) is then given by $\lambda = 0$. Again the values of p_{11}, p_{12} and p_{22} are calculated using (3.17), giving one possible minimizer $p_{11} = d_{11}$, $p_{12} = d_{\mathrm{S}}$ and $p_{22} = d_{22}$.

Solution of (3.12) **when** (3.15a) **applies.** If $d_{11} = d_{22}$ and $d_{\mathrm{S}} = 0$, we cannot perform the step from (3.14) to (3.17) because we have the possibility that $\lambda = 1$.

Therefore we determine the minimum of $H(p_{11}, p_{12}, p_{22})$ using another method. Using $d_{\mathrm{S}} = 0$ and $d_{22} = d_{11}$, the minimization (3.12) simplifies to

$$\underset{(p_{11},p_{12},p_{22})\in\mathbb{R}^3}{\arg\min} \frac{1}{2}\left((p_{11}-d_{11})^2 + 2\,p_{12}^2 + (p_{22}-d_{11})^2\right), \tag{3.30}$$

under the condition

$$p_{11}p_{22} - p_{12}^2 = \frac{f}{g}. \tag{3.31}$$

From the constraint it follows that $p_{12}^2 = p_{11}p_{22} - f/g$. Substitution in (3.30) gives

$$\underset{(p_{11},p_{22})}{\arg\min}\frac{1}{2}\left((p_{11}-d_{11})^2 + 2\,(p_{11}p_{22} - f/g) + (p_{22}-d_{11})^2\right), \tag{3.32}$$

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:

$$p_{11} + p_{22} = d_{11}. \tag{3.33}$$

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

$$p_{11}^2 - d_{11}\,p_{11} + \frac{f}{g} \leq 0. \tag{3.34}$$

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

$$\frac{d_{11} - \sqrt{d_{11}^2 - 4\,f/g}}{2} \leq p_{11} \leq \frac{d_{11} + \sqrt{d_{11}^2 - 4\,f/g}}{2}. \tag{3.35}$$

This is a minimizing line segment, and the the minimizer may not be unique. For simplicity, we choose p_{11} in the middle of the line segment, and find two vectors (p_{11}, p_{12}, p_{22}) given by

$$p_{11} = p_{22} = \frac{d_{11}}{2}, \qquad p_{12} = \pm\sqrt{\frac{d_{11}^2}{4} - \frac{f}{g}}. \tag{3.36}$$

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 an hyperbola:

$$p_{11}p_{22} = f/g. \tag{3.37}$$

Using $p_{12} = 0$ and $p_{22} = \frac{f/g}{p_{11}}$, (3.32) reduces to

$$\underset{p_{11}\in\mathbb{R}}{\arg\min}\frac{1}{2}\left((p_{11}-d_{11})^2 + \left(\frac{f}{g\,p_{11}} - d_{11}\right)^2\right). \tag{3.38}$$

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

$$p_{11}^4 - d_{11}\, p_{11}^3 + f/g\, d_{11}\, p_{11} - f^2/g^2 = 0. \tag{3.39}$$

This quartic equation can be factored into

$$\left(p_{11}^2 - f/g\right)\left(p_{11}^2 - d_{11}p_{11} + f/g\right) = 0, \tag{3.40}$$

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

$$p_{11} = \pm\sqrt{f/g}, \tag{3.41}$$

and are always real. The other two solutions

$$p_{11} = \frac{d_{11} \pm \sqrt{d_{11}^2 - 4\, f/g}}{2} \tag{3.42}$$

either correspond to the endpoints of the line segment defined by (3.35), or are complex valued. If they correspond to the boundaries of the line segment, they will yield the same value of $H_{\mathrm{S}}(p_{11}, p_{12}, p_{22})$ as the solution (3.36) and can thus be ignored. If they are complex valued, they can be ignored as well. From the first two solutions (3.41), we find the following two possible minimizers:

$$p_{11} = p_{22} = \pm\sqrt{f/g}, \qquad p_{12} = 0. \tag{3.43}$$

Solution of (3.12) **when** (3.15b) **applies.** If $d_{11} = -d_{22}$, we also cannot perform the step from (3.14) to (3.17) because we have the possibility that $\lambda = -1$. We determine the minimizers using a different method. We find from (3.14a) and (3.14b):

$$p_{22} = p_{11} - d_{11}, \tag{3.44a}$$
$$p_{12} = d_{\mathrm{S}}/2. \tag{3.44b}$$

Substituting this in (3.14d) we find

$$p_{11}^2 - d_{11}p_{11} - d_{\mathrm{S}}^2/4 - f/g = 0. \tag{3.45}$$

Solving for p_{11} we find two solutions:

$$p_{11} = \frac{d_{11}}{2} \pm \frac{\sqrt{d_{11}^2 + 4\, f/g + d_{\mathrm{S}}^2}}{2}, \tag{3.46a}$$
$$p_{12} = \frac{d_{\mathrm{S}}}{2}, \tag{3.46b}$$
$$p_{22} = -\frac{d_{11}}{2} \pm \frac{\sqrt{d_{11}^2 + 4\, f/g + d_{\mathrm{S}}^2}}{2}, \tag{3.46c}$$

which are always real.

Summary. The procedure to minimize over $\boldsymbol{P}$ is as follows: for each grid point (x_i, y_j), we calculate the value of $f(x_i, y_j)/g(\boldsymbol{m}(x_i, y_j))$. We determine which of the five cases applies, and find at least one and at most four possible minimizers (p_{11}, p_{12}, p_{22}). For each i, j, the new $\boldsymbol{P}$ is constructed from the 3-tuple which gives the smallest value of $H(p_{11}, p_{12}, p_{22})$.

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

$$||\boldsymbol{A} + \boldsymbol{B}||^2 = ||\boldsymbol{A}||^2 + 2\,\boldsymbol{A} : \boldsymbol{B} + ||\boldsymbol{B}||^2\,, \tag{3.47}$$

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

$$\begin{aligned} \delta J(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b})\,[\boldsymbol{\eta}] = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[J(\boldsymbol{m} + \epsilon\boldsymbol{\eta}, \boldsymbol{P}, \boldsymbol{b}) - J(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b})\right] = \\ \lim_{\epsilon \to 0} \left[\frac{\alpha}{2} \iint_{\mathcal{X}} 2\,(D\boldsymbol{m} - \boldsymbol{P}) : D\boldsymbol{\eta} + \epsilon\, ||D\boldsymbol{\eta}||^2 \,\mathrm{d}x\,\mathrm{d}y + \right. \\ \left. \frac{1-\alpha}{2} \oint_{\partial\mathcal{X}} 2\,(\boldsymbol{m} - \boldsymbol{b}) \cdot \boldsymbol{\eta} + \epsilon |\boldsymbol{\eta}|^2 \,\mathrm{d}s \right] = \\ \alpha \iint_{\mathcal{X}} (D\boldsymbol{m} - \boldsymbol{P}) : D\boldsymbol{\eta}\,\mathrm{d}x\,\mathrm{d}y + (1-\alpha) \oint_{\partial\mathcal{X}} (\boldsymbol{m} - \boldsymbol{b}) \cdot \boldsymbol{\eta}\,\mathrm{d}s. \end{aligned} \tag{3.48}$$

The minimizer is given by

$$\delta J(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b})\,[\boldsymbol{\eta}] = 0, \quad \forall \boldsymbol{\eta} \in \left[C^2(\mathcal{X})\right]^2. \tag{3.49}$$

Let

$$\boldsymbol{p}_1 = \begin{pmatrix} p_{11} \\ p_{12} \end{pmatrix},\ \boldsymbol{p}_2 = \begin{pmatrix} p_{12} \\ p_{22} \end{pmatrix},\ \boldsymbol{P} = [\,\boldsymbol{p}_1\ \ \boldsymbol{p}_2\,],\ \boldsymbol{\eta} = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix},\ \boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

We rewrite the first integral of the final expression in (3.48) as follows:

$$\iint_{\mathcal{X}} (D\boldsymbol{m} - \boldsymbol{P}) : D\boldsymbol{\eta}\,\mathrm{d}x\,\mathrm{d}y = \sum_{k=1}^{2} \iint_{\mathcal{X}} (\nabla m_k - \boldsymbol{p}_k) \cdot \nabla \eta_k \,\mathrm{d}x\,\mathrm{d}y. \tag{3.50}$$

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

$$\begin{aligned} \iint_{\mathcal{X}} (\nabla m_k - \boldsymbol{p}_k) \cdot \nabla \eta_k \,\mathrm{d}x\,\mathrm{d}y = \\ \oint_{\partial\mathcal{X}} (\nabla m_k - \boldsymbol{p}_k) \cdot \hat{\boldsymbol{n}}\, \eta_k \,\mathrm{d}s - \iint_{\mathcal{X}} (\Delta m_k - \nabla \cdot \boldsymbol{p}_k)\, \eta_k \,\mathrm{d}x\,\mathrm{d}y, \end{aligned} \tag{3.51}$$

where $\hat{\boldsymbol{n}}$ is the outward-pointing unit normal vector on $\partial\mathcal{X}$. Combining (3.51) with (3.48), (3.49) and (3.50) we find

$$\begin{aligned} \sum_{k=1}^{2} \oint_{\partial\mathcal{X}} \Big(\left[\alpha\,(\nabla m_k - \boldsymbol{p}_k) \cdot \hat{\boldsymbol{n}} + (1-\alpha)\,(m_k - b_k)\right] \eta_k \\ - \alpha \iint_{\mathcal{X}} (\Delta m_k - \nabla \cdot \boldsymbol{p}_k)\, \eta_k \,\mathrm{d}x\,\mathrm{d}y = 0 \quad \forall \boldsymbol{\eta} \in \left[C^2(\mathcal{X})\right]^2. \end{aligned} \tag{3.52}$$

Choosing $\eta_2 = 0$ and applying the fundamental lemma of the Calculus of Variations [6, p.185] for η_1, we have almost everywhere

$$\Delta m_1 = \nabla \cdot \boldsymbol{p}_1 \qquad (x, y) \in \mathcal{X}, \tag{3.53a}$$

$$(1-\alpha)\, m_1 + \alpha\, \nabla m_1 \cdot \hat{\boldsymbol{n}} = (1-\alpha)\, b_1 + \alpha\, \boldsymbol{p}_1 \cdot \hat{\boldsymbol{n}} \qquad (x, y) \in \partial\mathcal{X}. \tag{3.53b}$$

Similarly, choosing $\eta_1 = 0$ we find

$$\Delta m_2 = \nabla \cdot \boldsymbol{p}_2 \qquad (x,y) \in \mathcal{X}, \tag{3.54a}$$

$$(1-\alpha)\, m_2 + \alpha\, \nabla m_2 \cdot \hat{\boldsymbol{n}} = (1-\alpha)\, b_2 + \alpha\, \boldsymbol{p}_2 \cdot \hat{\boldsymbol{n}} \qquad (x,y) \in \partial\mathcal{X}. \tag{3.54b}$$

These are two decoupled Poisson equations with Robin boundary conditions for the two components of $\boldsymbol{m}$ [15]. These Poisson equations are solved using standard second order accurate central finite 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_{12} and p_{22} are approximated using similar finite difference schemes, and when needed on the boundary, using one-sided second-order schemes. The discretized Poisson equations are solved using the Matlab-implementation of LU-decomposition with full pivoting. The LU-decomposition only needs to be computed once, and the Poisson equations can be solved very efficiently during the iterations.

4. Computation of u and algorithm summary. The minimization steps given by (2.13a), (2.13b) and (2.13c) are repeated until $J(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b})$ is no longer decreasing. Then we stop the iteration. If we are interested in the solution of the Monge-Ampère equation, we can compute its solution u from the mapping $\boldsymbol{m}$, i.e., we calculate u such that $\nabla u = \boldsymbol{m}$. This u will be the approximate solution of the Monge-Ampère equation (1.3) with boundary condition (1.4). In the ideal situation, $D\boldsymbol{m} = \boldsymbol{P}$ and thus $\frac{\partial m_1}{\partial y} = \frac{\partial m_2}{\partial x}$. In this case there exists a function u such that $\nabla u = \boldsymbol{m}$, because $\boldsymbol{m}$ 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 $\boldsymbol{m}$ as gradient in a least-squares sense, i.e.,

$$u = \arg\min_{\psi} I(\psi), \qquad I(\psi) = \frac{1}{2} \iint_{\mathcal{X}} |\nabla\psi - \boldsymbol{m}|^2 \, \mathrm{d}x\, \mathrm{d}y. \tag{4.1}$$

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

$$\begin{aligned} \delta I(u)[v] &= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(I(u + \epsilon v) - I(u) \right) \\ &= \lim_{\epsilon \to 0} \frac{1}{2} \iint_{\mathcal{X}} \epsilon \, |\nabla v|^2 + 2 \, (\nabla u - \boldsymbol{m}) \cdot \nabla v \, \mathrm{d}x \, \mathrm{d}y \\ &= \iint_{\mathcal{X}} (\nabla u - \boldsymbol{m}) \cdot \nabla v \, \mathrm{d}x \, \mathrm{d}y. \end{aligned} \tag{4.2}$$

The minimizer is given by

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

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

$$0 = \oint_{\partial\mathcal{X}} (\nabla u - \boldsymbol{m}) \cdot \hat{\boldsymbol{n}} \, v \, \mathrm{d}s - \iint_{\mathcal{X}} (\Delta u - \nabla \cdot \boldsymbol{m}) \, v \, \mathrm{d}x \, \mathrm{d}y, \ \forall v \in C^2(\mathcal{X}). \tag{4.4}$$

Applying the fundamental lemma of Calculus of Variations [6, p.185], we find

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

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

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

$$\iint_{\mathcal{X}} -\nabla \cdot \boldsymbol{m} \,\mathrm{d}x\,\mathrm{d}y + \oint_{\partial\mathcal{X}} \boldsymbol{m} \cdot \hat{\boldsymbol{n}} \,\mathrm{d}s = 0. \tag{4.6}$$

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

$$u(x_i, y_j) = 0, \tag{4.7}$$

for some arbitrarily chosen i and 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 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 $\boldsymbol{b}(x_i, y_j)$, 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 $\boldsymbol{P}(x_i, y_j)$, and is performed pointwise for all gridpoints by calculating the global minimizer. The third and last step is a minimization procedure for $\boldsymbol{m}$, and is performed by solving two Poisson boundary value problems given by (3.53) and (3.54). The three steps are repeated until the functional $J(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{b})$ defined by (2.12) has decreased sufficiently. If needed, the function u is computed from $\boldsymbol{m}$ by solving the Poisson problem (4.5).

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

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 $\mathcal{Y} = \left\{(p, q) \in \mathbb{R}^2 \,\middle|\, p^2 + q^2 \leq 1\right\}$. We have $f(x, y) = 1/4$ and $g(p, q) = 1/\pi$. We solve the boundary value problem (1.3) and (1.4) for different grid sizes with $\alpha = 0.2$. The resulting mapping after 100 iterations is shown in Figure 5.1.

Subsequently we test the algorithm with different grid constants. We use an $N \times N$ grid with $N \in \{51, 61, 71, 81, 91, 101, 151, 201, 251, 301, 351, 401\}$. The number of boundary points $N_\mathrm{b} = 1000$, such that it will not be the limiting factor in accuracy. We calculate the mapping on the $N \times N$ grids using 400 iterations. The value of J as function of the iteration number for the different grid sizes is shown in Figure 5.2a. The lower lines correspond to larger grids. Surprisingly, the grid size does not influence the convergence speed in the initial iterations. It does, however, determine at which value the convergence stalls. The final value of J after 400 iterations is plotted in Figure 5.2b.

We analyzed the calculation time for the LU decomposition, the three minimization steps performed in each iteration and the computation of $\boldsymbol{u}$ for $N \times N$ grids as function N, and as function of the number of boundary points N_b. The results are shown in Figure 5.3. The calculation times for the LU decomposition and the computation of the function u are clearly the largest, however, they only have to be

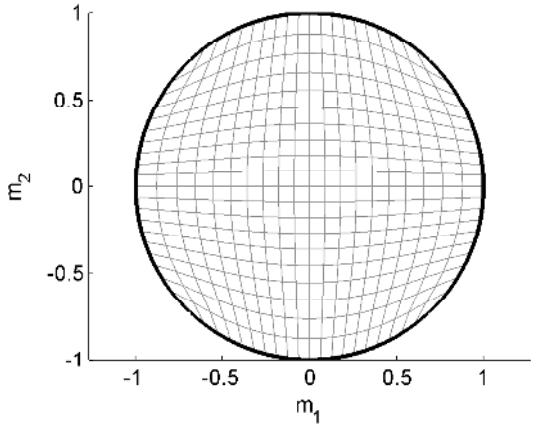
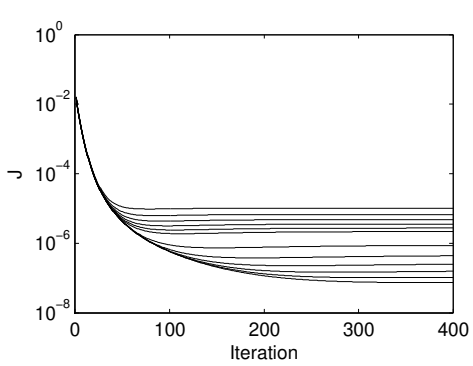
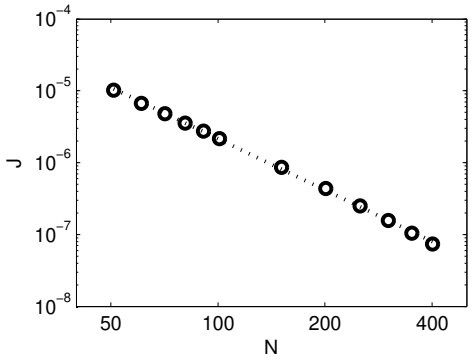


Figure 5.1: The mapping for the 'square to a circle' problem after 100 iterations on a 101×101 grid, 100 points on the boundary of the target domain and $\alpha = 0.2$.



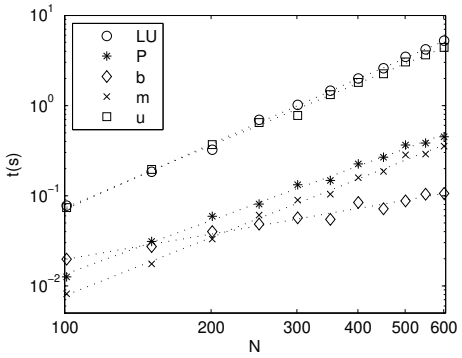
(a) The value of J as function of the iteration number for the different grid sizes. The lower lines correspond to larger values of N.



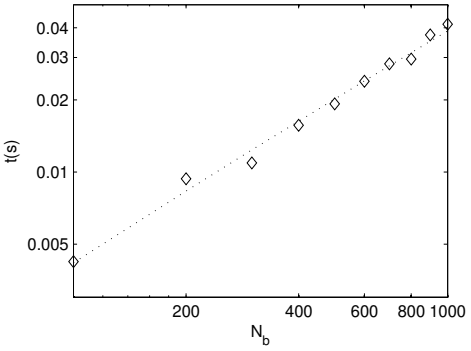
(b) Value of J after 400 iterations as function of N on a loglog scale. The fit is a logarithmic linear least-squares, we find that the value of J is proportional to $N^{-2.37}$.

Figure 5.2: Calculation of the mapping for different grid sizes with $N_\mathrm{b} = 1000$, $\alpha = 0.2$ and 400 iterations. The grids have dimensions $N \times N$ with $N \in \{51, 61, 71, 81, 91, 101, 151, 201, 251, 301, 351, 401\}$.

performed once. We expect the calculation times for the minimization procedures for $\boldsymbol{P}$ and $\boldsymbol{m}$ (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 $\boldsymbol{b}$ to be linear in both N and N_b. The calculation times for the LU decomposition and the linear solve for the calculation of the function u depend on the Matlab implementation, we expect 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 5.3, the results are shown in Table 5.1.



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



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

Figure 5.3: Calculation time as function of N and N_b. The calculations were performed on a laptop with an Intel Core i5 M520 2.40 GHz with 4 GB of RAM. The dotted lines are least-squares fits. See also Table 5.1.

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

Table 5.1: Expected and measured calculation times for the different procedures in the algorithm.

We also tested the influence of N_b on the convergence rate 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 5.4. Again, the value of N_b does not influence the initial convergence of the algorithm, but only the final value of J after convergence. If N_b is chosen large enough, its effect on the final value of J becomes negligible. Because an increase in N_b does not significantly increase the calculation time, it is recommended to choose N_b sufficiently large.

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 5.5. 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 $[-1, 1]^2 \in \mathbb{R}^2$ and we have $g(p, q) = 1/4$. Because the

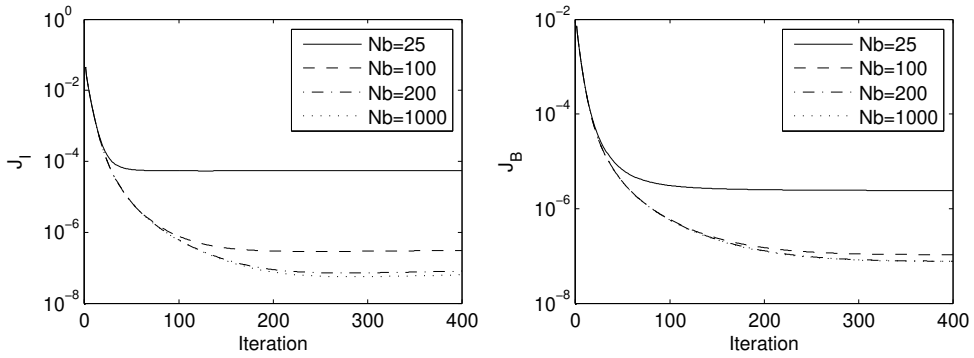


Figure 5.4: Value of J_{I} and J_{B} as function of the iteration number for different values of N_{b} on a 401×401 grid.

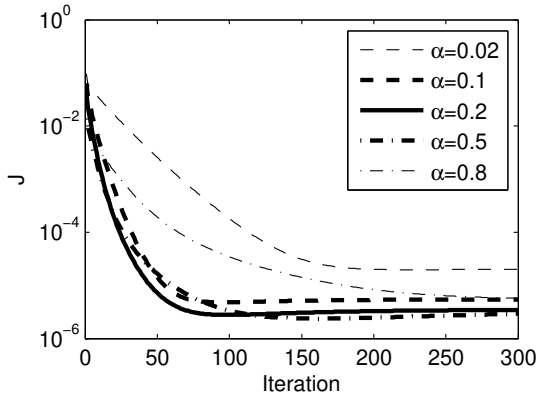


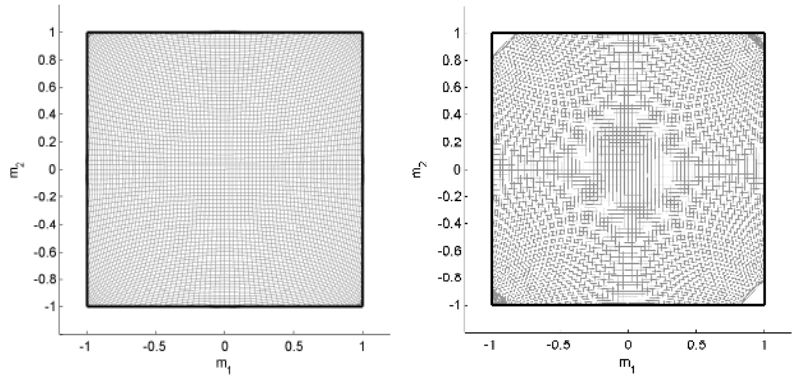
Figure 5.5: Value of J as function of the iteration number for mapping a square to a circle with different values of α.

target domain is square, we only need 4 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 5.6d. Convergence of the mapping was often problematic in the corners. Choosing the right value of α reduces this problem.

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

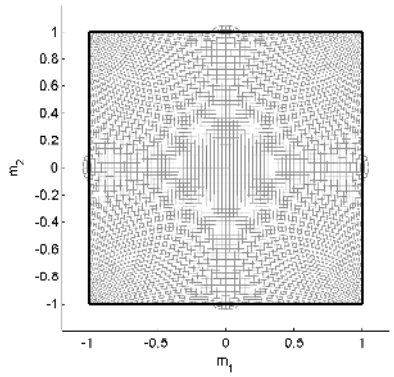
$$\rho(\theta) = 1 + C\cos(3\,\theta), \tag{5.1}$$

where θ is the counter-clockwise 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 5.7. The value of J after 200 iterations is $4.43 \cdot 10^{-7}$, $8.53 \cdot 10^{-7}$, $5.16 \cdot 10^{-5}$

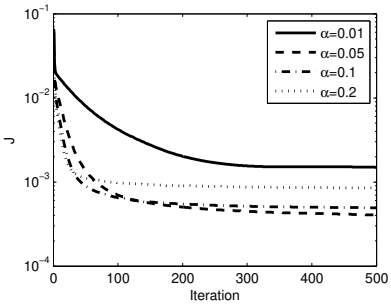


(a) $\alpha = 0.05$

(b) $\alpha = 0.01$



(c) $\alpha = 0.2$

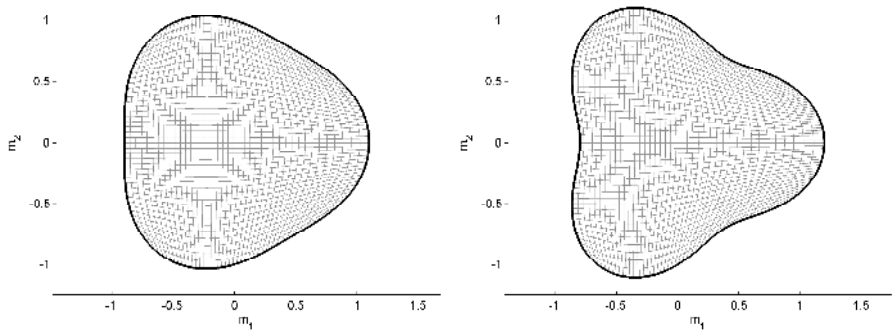


(d) Value of J as function of the iteration number with different values of α for mapping a circle to a square.

Figure 5.6: Final mapping after 500 iterations on a 401×401 grid for the second test case. Convergence is often slow in the corners. Choosing the right value of α reduces this problem.

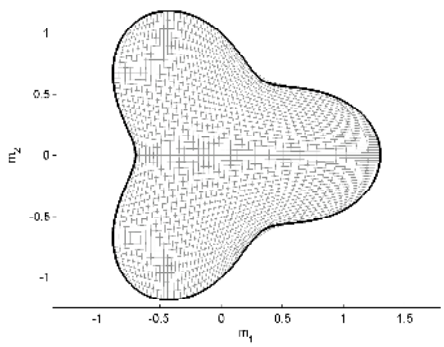
and $4.82 \cdot 10^{-4}$, respectively. Convergence problems arise for target domains which strongly deviate from a convex shape, but if the shape only deviates moderately, the algorithm performs adequately.

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 illuminating the streets uniformly, spots with square, rectangular or arbitrary-shaped output beams, or car headlamps. For this article, we 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.

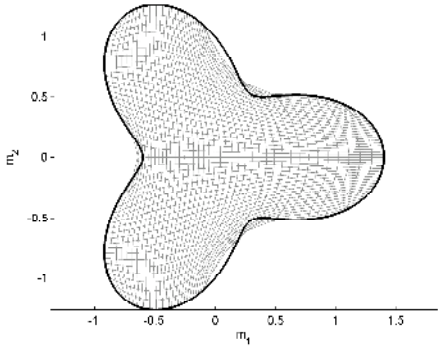


(a) $C = 0.1$

(b) $C = 0.2$



(c) $C = 0.3$



(d) $C = 0.4$

Figure 5.7: Mapping after 200 iterations for target domains with boundary defined by (5.1).

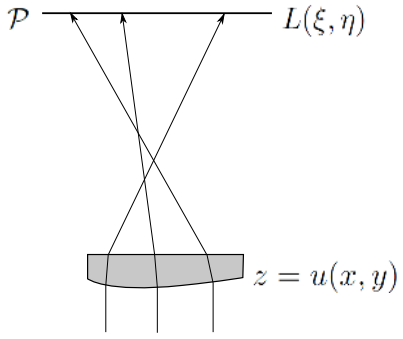
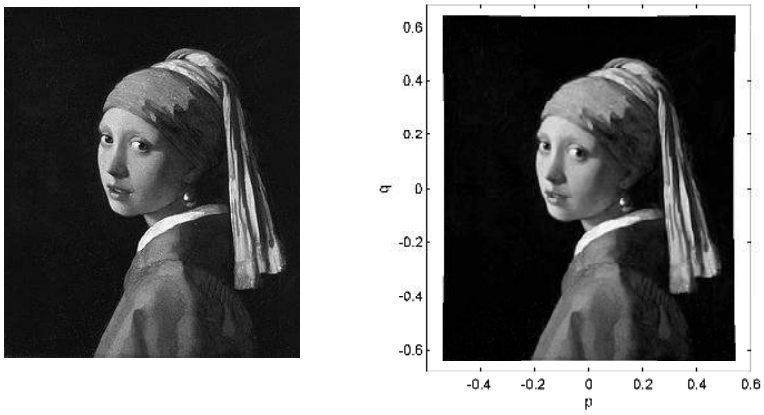


Figure 5.8: A lens redistributing the light from a parallel beam onto a plane.



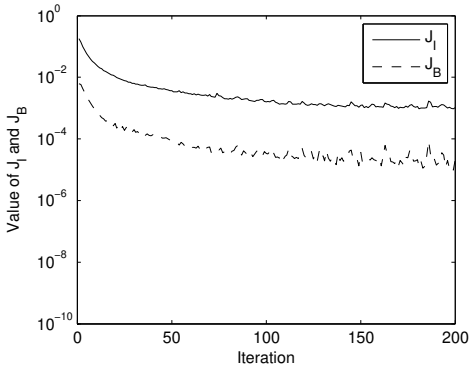
(a) The original painting

(b) The corresponding target distribution $g(p, q)$.

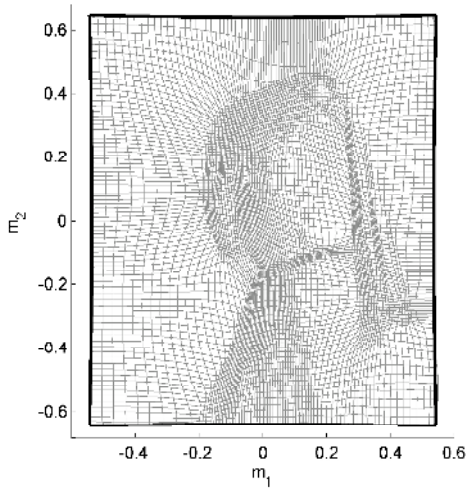
Figure 5.9: The painting "Girl with a pearl earring" by Johannes Vermeer. Note that the domain of the target distribution $g(p, q)$ is not convex.

A schematic drawing of a lens design problem is shown in Figure 5.8. In this example, we have a parallel beam of light in the z-direction with given radiant exitance $M(x, y)$ [W/m^2], and an irradiance on the plane $\mathcal{P}$ given by $L(x, y)$ [W/m^2]. 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 $z = u(x, y)$ which gives for the specified radiant exitance $M(x, y)$ of the parallel beam the specified irradiance $L(x, y)$ 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 5.9b [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 5.9. The painting is converted to grayscale and the grayscale values are used as 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 5.9b. 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 a 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 function of the iteration number are plotted in Figure 5.10a. It can be seen that the value of J may increase during the iterations. This happens during the minimization procedure for $\boldsymbol{P}$. The reason is that the mapping $\boldsymbol{m}$ has changed, and thus the set $\mathcal{P}(\boldsymbol{m})$ has changed, therefore we



(a) Values of J_{I} and J_{B} for the Vermeer-lens as function of the iteration number.



(b) The final mapping for the Vermeer-lens.

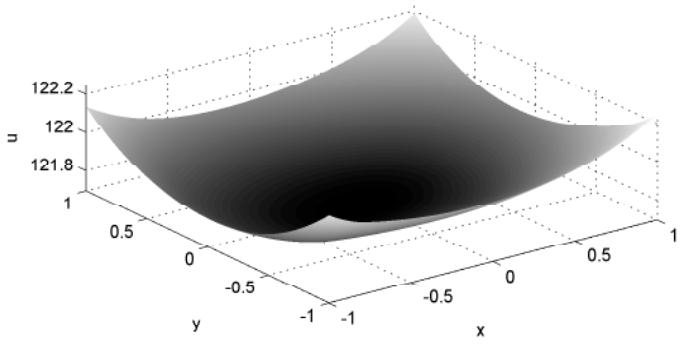


Figure 5.10: The function $u(x, y)$ representing the lens-surface.

may have

$$\min_{\boldsymbol{P}\in\mathcal{P}(\boldsymbol{m}^{n+1})} J_{\mathrm{I}}(\boldsymbol{m}^{n+1}, \boldsymbol{P}) > \min_{\boldsymbol{P}\in\mathcal{P}(\boldsymbol{m}^{n})} J_{\mathrm{I}}(\boldsymbol{m}^{n}, \boldsymbol{P}). \tag{5.2}$$

The final mapping resulting from the iteration process is shown in Figure 5.10b. The target distribution 5.9b can clearly be 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 5.10.

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 5.11. The software shoots 5 million evenly distributed rays randomly from a square plane through the

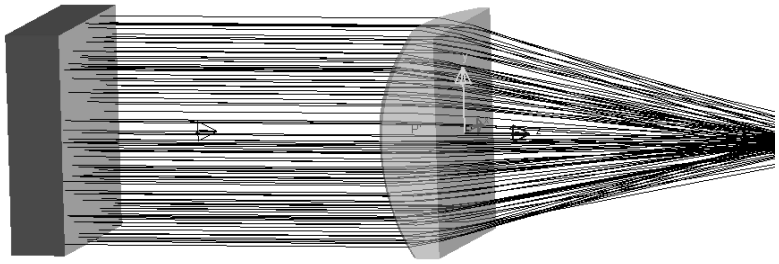


Figure 5.11: Ray-tracing the Vermeer lens. The screenshot shows the light source (left) and the lens (middle). The target screen is located at a large distance to the right.

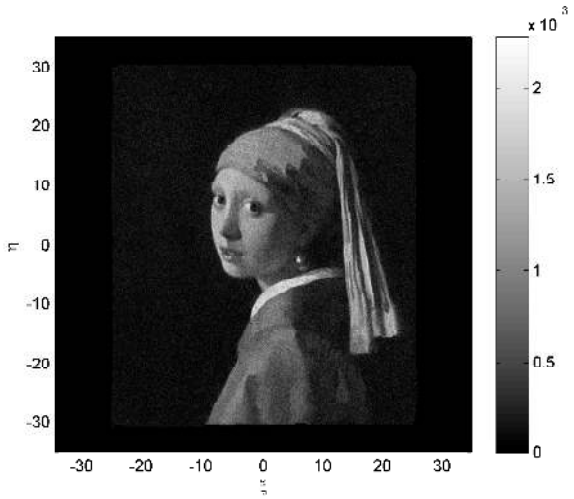


Figure 5.12: The ray-tracing results for the Vermeer lens.

lens and returns the distribution of these rays over the target plane. This distribution is plotted in Figure 5.12.

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 optimal mass transport and the corresponding convex solution of the Monge-Ampère equation with the transport boundary condition. Numerical methods for optimal mass 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, 9–12, 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 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 linear with the number of grid points, and the LU decomposition and linear solve for u only need to be performed once. If $f = 0$ on part of the domain, the wide-stencil scheme performs better: the least-squares algorithm converges slowly for these type 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 we computed solutions for non-convex target domains by swapping the source and target distributions: we retrieved the solution of the original problem using an additional Legendre-Fenchel transform. However, this introduces additional calculation time, artifacts in the solution, and is only possible if the source domain is convex [21].

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

REFERENCES

[1] S. Angenent, 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.* Preprint available at `http://arxiv.org/abs/1208.4873`.

[4] ———, *Numerical solution of the optimal transportation problem using the Monge-Ampère equation*, Journal of Computational Physics, 260 (2014), pp. 107–126.

[5] A. Caboussat, R. Glowinski, and D. Sorensen, *A least-squares method for the numerical solution of the Dirichlet problem for the elliptic Monge-Ampère equation in dimension two.*, ESAIM: Control, Optimisation and Calculus of Variations, 19 (2013), pp. 780–810.

[6] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, vol. 1, John Wiley & Sons, 1989 ed., 1953.

[7] H. Elman, D. Silvester, and A. Wathen, *Finite elements and fast iterative solvers.*, Oxford University Press, 2005.

[8] L. Evans, *Partial differential equations and Monge-Kantorovich mass transfer.* 2001.

[9] B. Froese, *A numerical method for the elliptic Monge-Ampère equation with transport boundary conditions*, SIAM Journal on Scientific Computing, 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 Journal of Numerical Analysis, 49 (2011), pp. 1692–1714.

[11] ———, *Fast finite difference solvers for singular solutions of the elliptic Monge-Ampère equation*, Journal of Computational Physics, 230 (2011), pp. 818–834.

[12] ———, *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*, W.H.Freeman and Company, 4th ed., 1996.

[15] R. Mattheij, S. Rienstra, and J. ten Thije Boonkkamp, *Partial Differential Equations*, SIAM, 2005.

[16] A. Oberman, *Convergent difference schemes for degenerate elliptic and parabolic equations:*

Hamilton-Jacobi equations and free boundary problems, SIAM Journal of Numerical Analysis, 44 (2006), pp. 879–895.

[17] ———, *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, (2008).

[18] ORA, *Ora lighttools product website*. `http://optics.synopsys.com/lighttools/`. Accessed August 19, 2014.

[19] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical Recipes in Fortran77*, Cambridge University Press, 2nd ed., 1996.

[20] C. Prins, *Inverse methods for illumination optics*, PhD thesis, Eindhoven University of Technology, 2014.

[21] C. Prins, J. ten Thije Boonkkamp, J. v. Roosmalen, T. Tukker, and W. IJzerman, *A Monge-Ampère solver for free-form reflector design*, SIAM J. Sci. Comput., 36 (2014).

[22] J. Tignol, *Galois's theory of algebraic equations*, Longman Scientific & Technical, 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*, American Mathematical Society, 2003.

PREVIOUS PUBLICATIONS IN THIS SERIES:

Number	Author(s)	Title	Month
14-25	F.A. Radu J.M. Nordbotten I.S. Pop K. Kumar	A robust linearization scheme for finite volume based discretizations for simulation of two-phase flow in porous media	Sept. '14
14-26	X. Cao I.S. Pop	Uniqueness of weak solutions for a pseudo-parabolic equation modeling two phase flow in porous media	Sept. '14
14-27	M.E. Hochstenbach L. Reichel G. Rodriguez	Regularization parameter determination for discrete ill-posed problems	Sept. '14
14-28	X. Cao I.S. Pop	Two-phase porous media flows with dynamic capillary effects and hysteresis: uniqueness of weak solutions	Sept. '14
14-29	C.R. Prins J.H.M. ten Thije Boonkkamp W.L. IJzerman T.W. Tukker	A least-squares method for optimal transport using the monge-ampère equation	Sept. '14

Ontwerp: de Tantes,
Tobias Baanders, CWI