CHEMNITZ UNIVERSITY OF TECHNOLOGY

Preprint

The Numerical Solution of Newton's Problem of Least Resistance

Gerd Wachsmuth*

January 24, 2013
Research Group
Numerical Mathematics
(Partial Differential Equations)

Abstract

In this paper we consider Newton's problem of finding a convex body of least resistance. This problem could equivalently be written as a variational problem over concave functions in $\mathbb{R}^2$.

We propose two different methods for solving it numerically. First, we discretize this problem by writing the concave solution function as a infimum over a finite number of affine functions. The discretized problem could be solved by standard optimization software efficiently.

Second, we conjecture that the optimal body has a certain structure. We exploit this structure and obtain a variational problem in $\mathbb{R}^1$. Deriving its Euler-Lagrange equation yields a program with two unknowns, which can be solved quickly.

1 Introduction

The problem considered in this paper was raised by Newton in the late 17^{th} century. It consists of finding a convex body P with given base $\bar{\Omega}$ and height L, such that the

*Chemnitz University of Technology, Faculty of Mathematics, D–09107 Chemnitz, Germany, gerd.wachsmuth@mathematik.tu-chemnitz.de, http://www.tu-chemnitz.de/mathematik/part_dgl/wachsmuth

resistance induced by the movement through a rare medium is minimal. By describing the body P by a concave function $f : \bar{\Omega} \to \mathbb{R}$,

$$P = \{(x, y, z) \in \mathbb{R}^3 : (x, y) \in \bar{\Omega}, z \in [0, f(x, y)]\},$$

the problem can be written as a variational problem

$$\begin{aligned} &\text{Minimize} \quad J(f) = \int_\Omega \frac{1}{1 + \|\nabla f(x, y)\|^2} \, \mathrm{d}(x, y) \\ &\text{such that} \quad f : \Omega \to [0, L] \text{ is concave.} \end{aligned} \tag{P}$$

Note that the concavity of f implies that f belongs to $W^{1,\infty}_{\mathrm{loc}}(\Omega)$. Hence, the objective is well defined. For a derivation of the objective, we refer to Buttazzo et al. [1995], Buttazzo and Kawohl [1993]. Since the class of bounded convex functions is compact in $W^{1,p}_{\mathrm{loc}}(\Omega)$, the existence of minimizers of (**P**) can be proven, see [Buttazzo et al., 1995, Thm. 2.1].

The case considered by Newton himself and which is dealt with here and in many papers is that Ω the interior of the unit disc

$$\Omega = U_1(0) = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}.$$

Under the assumption that the optimal f is rotational symmetric, Newton was able to give an explicit solution. However, in Brock et al. [1996] it was shown that Newton's radial solution is not a local optimum of (**P**). Hence, the minimizer cannot be rotationally symmetric and the actual shape of the minimizers became an open problem.

In Lachand-Robert and Peletier [2001], the authors restricted the optimal body P to the set of bodies which can be written as a convex hull

$$\operatorname{conv}\big((\partial\Omega \times \{0\}) \cup (N_0 \times \{L\})\big),$$

where $N_0 \subset \mathbb{R}^2$ is the upper face of the body. They showed that the optimal set N_0 is a regular polygon centered at the origin $(0, 0)$. In particular, the optimal P has the symmetry group D_m for some $m \geq 2$. Here, D_m is the dihedral group, which is the symmetry group of the regular, m-sided polygon. The minimizers in this class of bodies have smaller objective values than Newton's radial solutions.

There is only one contribution in the literature which considers a numerical approximation of (**P**), see Lachand-Robert and Oudet [2005]. Their results show that the solution of Lachand-Robert and Peletier [2001] is not optimal for (**P**). Lachand-Robert and Oudet [2005] does not report computational times. However, since their method includes a genetic algorithm, it is to be expected that the method is rather slow. Moreover, in the case $L = 0.4$ they found only a local minimizer, see Section 4 and in particular Table 4.1.

The contribution of the present paper is twofold. In Section 2 we propose an algorithm for the solution of (**P**). We discretize the concave function f by an infimum of a finite number of affine functions. The discretized problem could be efficiently solved by

standard optimization software (by using the full gradient). Altogether, we are able to compute approximations of the minimizers of (**P**) in a few minutes.

By the results obtained in Section 2, we conjecture in Section 3 that the optimal body P belongs to a certain class (for heights L smaller than about 1.4). Using this conjecture, we are able to reduce (**P**) to a one-dimensional variational problem. By deriving the associated Euler-Lagrange equation, we further reduce the problem to a minimization problem with two unknowns. These minimization problems could be solved in a few seconds. Moreover, the results obtained by this method are slightly better than the results obtained in Section 2, which consolidates the conjecture.

In Section 4 we summarize the results.

2 Discretization by the Infimum of Hyperplanes

Let $n \geq 1$ be given. We discretize problem (**P**) by considering only those functions f, which can be written as an infimum over n affine functions. To be precise, let $A \in \mathbb{R}^{n\times 3}$ be the matrix of coefficients. Then $f(A) : \Omega \to \mathbb{R}$ is defined by

$$f(A, x, y) = \inf_{i=1,\dots,n} f_i(A, x, y), \tag{2.1}$$

where the affine functions f_i are given by

$$f_i(A, x, y) = A_{i,1}\, x + A_{i,2}\, y + A_{i,3} \quad \text{for } i = 1, \dots, n. \tag{2.2}$$

The function $f(A)$ is concave by definition. Moreover, problem (**P**) can be approximated by this discretization, see [Lachand-Robert and Oudet, 2005, Lemma 1] for a similar result.

In Section 2.1 we derive formulas for $J(f)$ in terms of the coefficients A. Moreover, we provide the derivatives w.r.t. the coefficients A. We deal with the constraint $f(A, x, y) \in [0, L]$ for all $(x, y) \in \Omega$ in Section 2.2. In Section 2.3 we present a preliminary numerical result. The refinement of a given solution and further improvements of the implementation are addressed in Sections 2.4 and 2.5. Finally, the numerical results are presented in Section 2.6.

2.1 Evaluation of Function Values and Derivatives

In this section we will compute the objective $J(f(A))$ and the derivative $\mathrm{d}J(f(A))/\mathrm{d}A$ in terms of the coefficients A.

Let us denote by

$$D_i(A) = \{(x, y) \in \bar{\Omega} : f_i(A, x, y) \geq f_j(A, x, y) \text{ for all } j = 1, \dots, n\}$$

the dominating region of the function f_i. Often we will suppress the dependence of f, f_i and D_i on A. The function f_i is called *active* if $D_i \neq \emptyset$, and *strictly active* if $\mu_2(D_i) > 0$. Here, we denote by μ_d the d-dimensional Hausdorff measures. For a strictly active function f_i, the set D_i is a curvilinear convex polygon (the intersection of a convex polygon with $\bar{\Omega}$). All straight edges are of the form $D_i \cap D_j$, whereas the non-straight edges are of the form $D_i \cap \partial\Omega$. The computation of D_i is discussed in Section 2.2.

For the differentiability results, we assume that the discretization is not degenerated.

Assumption 2.1. For all $i \neq j$ we have

$$\mu_2(D_i \cap D_j) = 0.$$

Moreover, for pairwise different indices i, j, k we have

$$\mu_1(D_i \cap D_j \cap D_k) = 0.$$

The first part of this assumption is equivalent to requiring that all active functions f_i are distinct. Moreover, it implies

$$J(f) = \sum_{i=1}^{n} \mu_2(D_i) \frac{1}{1 + A_{i,1}^2 + A_{i,2}^2}.$$

In order to compute the derivative of the area of D_i with respect to the coefficients A, we first have a look on the derivative of the area of a polygon P with respect to the coordinates of its vertices. Let a polygon P with vertices $\boldsymbol{p}_i = (x_i, y_i)$ (in counterclockwise orientation), $i = 1, \dots, m$ be given. Using the shoelace formula we find that the area $\mu_2(P)$ is differentiable w.r.t. the coordinates of the vertices and the coordinate-wise derivatives are given by

$$\frac{\mathrm{d}\mu_2(P)}{\mathrm{d}x_i} = \frac{1}{2}(y_{i+1} - y_{i-1}) \quad \text{and} \quad \frac{\mathrm{d}\mu_2(P)}{\mathrm{d}y_i} = \frac{1}{2}(x_{i-1} - x_{i+1}), \tag{2.3}$$

where we used the convention $x_0 = x_m$, $x_1 = x_{m+1}$ and the same for y_i. A simple calculation shows

$$\frac{\mathrm{d}\mu_2(P)}{\mathrm{d}(x_i, y_i)} = \frac{1}{2}\big(\|\boldsymbol{p}_i - \boldsymbol{p}_{i-1}\|\, \boldsymbol{n}_i + \|\boldsymbol{p}_{i+1} - \boldsymbol{p}_i\|\, \boldsymbol{n}_{i+1}\big), \tag{2.4}$$

where $\boldsymbol{n}_i$ is a outer normal vector of the edge $\boldsymbol{e}_i$ between $\boldsymbol{p}_{i-1}$ and $\boldsymbol{p}_i$, see Figure 2.1. In the case of a degenerate vertex $\boldsymbol{p}_i = \boldsymbol{p}_{i+1}$, we have

$$\frac{\mathrm{d}\mu_2(P)}{\mathrm{d}(x_i, y_i)} = \frac{1}{2}\|\boldsymbol{p}_i - \boldsymbol{p}_{i-1}\|\, \boldsymbol{n}_i \quad \text{and} \quad \frac{\mathrm{d}\mu_2(P)}{\mathrm{d}(x_{i+1}, y_{i+1})} = \frac{1}{2}\|\boldsymbol{p}_{i+2} - \boldsymbol{p}_{i+1}\|\, \boldsymbol{n}_{i+2}$$

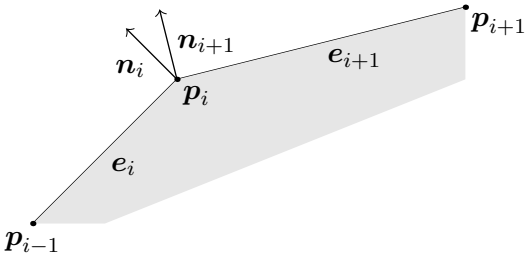

Figure 2.1: Two adjacent edges of a polygon with their normal vectors and incident vertices.

since $\|\boldsymbol{p}_i - \boldsymbol{p}_{i+1}\| = 0$. Note that the degeneracy of the polygon does not induce a non-differentiability.

Finally, we give an expression for the directional derivative of the area. Given variations $\delta\boldsymbol{p}_i$, $i = 1, \dots, m$ of the coordinates of the points $\boldsymbol{p}_i$, the variation of $\mu_2(P)$ can be computed as

$$\delta\mu_2(P) = \frac{1}{2}\sum_{i=1}^{n} \left(\|\boldsymbol{p}_{i+1} - \boldsymbol{p}_i\|\, \boldsymbol{n}_{i+1}^\top (\delta\boldsymbol{p}_i + \delta\boldsymbol{p}_{i+1})\right). \tag{2.5}$$

Note that the sum can be understood edge-wise (the ingredients of each summand are the length of the edge and its outer normal vector) and degenerate edges (i.e., edges with length 0) do not contribute to the derivative.

Formulas (2.4) and (2.5) remain true if the connection of the points $\boldsymbol{p}_{i-1}$ and $\boldsymbol{p}_i$ is not a straight edge, but a part of the boundary $\partial\Omega$. In this case, we have to replace $\boldsymbol{n}_i$ by the outer normal vector of Ω in $\boldsymbol{p}_i$. A (tangential) perturbation $\delta\boldsymbol{p}_i$ of $\boldsymbol{p}_i$ induces the change of the area

$$\delta\mu_2(P) = \frac{1}{2}\,\|\boldsymbol{p}_{i+1} - \boldsymbol{p}_i\|\, \boldsymbol{n}_{i+1}^\top\, \delta\boldsymbol{p}_i.$$

This formula does not contain a contribution related to the boundary arc $(\boldsymbol{p}_{i-1}, \boldsymbol{p}_i)$, since $\delta\boldsymbol{p}_i$ is tangential to the boundary of Ω and hence $\boldsymbol{n}_i^\top \delta\boldsymbol{p}_i = 0$. Hence, in the generalization of (2.5) to this case of a curvilinear polygon, we have to sum only over those indices i such that the edge between $\boldsymbol{p}_{i-1}$ and $\boldsymbol{p}_i$ is a straight edge.

Now we are in the position to prove the key lemma of this section.

Lemma 2.2. The functions $\mu_2(D_i)$ are differentiable w.r.t. A. The partial derivatives are given by

$$\frac{\mathrm{d}\mu_2(D_i)}{\mathrm{d}(A_{j,1}, A_{j,2}, A_{j,3})}(\delta A_{j,1}, \delta A_{j,2}, \delta A_{j,3}) = 0$$

for $i \neq j$ such that $\mu_1(D_i \cap D_j) = 0$ (the dominating regions D_i and D_j are not adjacent). In the case $i \neq j$ and $\mu_1(D_i \cap D_j) > 0$ (the intersection of D_i and D_j is an edge), we

have

$$\frac{\mathrm{d}\mu_2(D_i)}{\mathrm{d}(A_{j,1},A_{j,2},A_{j,3})}(\delta A_{j,1},\delta A_{j,2},\delta A_{j,3}) = \frac{1}{2}\,\|\boldsymbol{p}_{i,j}-\boldsymbol{q}_{i,j}\|\,\frac{(\delta A_{j,1},\delta A_{j,2})\,(\boldsymbol{p}_{i,j}+\boldsymbol{q}_{i,j})+2\,\delta A_{j,3}}{(A_{i,1}-A_{j,1},A_{i,2}-A_{j,2})\,\boldsymbol{n}_{i,j}}$$

where $\boldsymbol{p}_{i,j}$ and $\boldsymbol{q}_{i,j}$ are the vertices of the edge between D_i and D_j and $\boldsymbol{n}_{i,j}$ is the normal vector of the edge pointing towards D_j.

Finally we have

$$\frac{\mathrm{d}\mu_2(D_i)}{\mathrm{d}(A_{i,1},A_{i,2},A_{i,3})}(\delta A_{i,1},\delta A_{i,2},\delta A_{i,3}) = -\sum_{\substack{j\neq i\\ \mu_1(D_i\cap D_j)>0}} \frac{\mathrm{d}\mu_2(D_j)}{\mathrm{d}(A_{i,1},A_{i,2},A_{i,3})}(\delta A_{i,1},\delta A_{i,2},\delta A_{i,3}).$$

Proof. The above reasoning shows that we have to compute the derivative of the normal displacements of the vertices of each edge of D_i.

Let us consider a small perturbations δA of the coefficients.

Let i be a fixed index and let j be another arbitrary index, such that $D_i(A)$ and $D_j(A)$ share a common edge. This edge is a subset of the line

$$\big\{(x,y)\in\mathbb{R}^2 : (A_{i,1}-A_{j,1})\,x+(A_{i,2}-A_{j,2})\,y+(A_{i,3}-A_{j,3})=0\big\}.$$

Let us denote the end points of the edge by $\boldsymbol{p}_j$ and $\boldsymbol{q}_j$, and let $\boldsymbol{n}_j$ be the normal vector of the edge pointing towards D_j (since i is fixed, we suppress the dependence on i in the proof).

Due to the regularity assumption, $D_i(A+\delta A)$ and $D_j(A+\delta A)$ share a common edge after the perturbation δA. By $\delta\boldsymbol{p}_j$ and $\delta\boldsymbol{q}_j$ we denote the perturbation of the end points of the edge. In order to apply (2.5), we have to project $\delta\boldsymbol{p}_j$ and $\delta\boldsymbol{q}_j$ on $\boldsymbol{n}_j$. We use the decompositions $\delta\boldsymbol{p}_j = k^p_{j,1}\,\boldsymbol{n}_j + k^p_{j,2}\,(\boldsymbol{p}_j-\boldsymbol{q}_j)$ and $\delta\boldsymbol{q}_j = k^q_{j,1}\,\boldsymbol{n}_j + k^q_{j,2}\,(\boldsymbol{p}_j-\boldsymbol{q}_j)$. Note that $\boldsymbol{n}_j^\top\,(\boldsymbol{p}_j-\boldsymbol{q}_j)=\boldsymbol{0}$, hence $\boldsymbol{n}_j^\top\,\delta\boldsymbol{p}_j = k^p_{j,1}$. We have

$$\begin{pmatrix}A_{i,1}+\delta A_{i,1}-(A_{j,1}+\delta A_{j,1})\\ A_{i,2}+\delta A_{i,2}-(A_{j,2}+\delta A_{j,2})\end{pmatrix}^\top(\boldsymbol{p}_j+\delta\boldsymbol{p}_j)+A_{i,3}+\delta A_{i,3}-(A_{j,3}+\delta A_{j,3})=0.$$

Ignoring terms of higher order yields

$$\begin{pmatrix}\delta A_{i,1}-\delta A_{j,1}\\ \delta A_{i,2}-\delta A_{j,2}\end{pmatrix}^\top\boldsymbol{p}_j+\begin{pmatrix}A_{i,1}-A_{j,1}\\ A_{i,2}-A_{j,2}\end{pmatrix}^\top\delta\boldsymbol{p}_j+\delta A_{i,3}-\delta A_{j,3}=o(\delta A).$$

Using $(A_{i,1}-A_{j,1},A_{i,2}-A_{j,2})\,(\boldsymbol{p}_j-\boldsymbol{q}_j)=0$, we find

$$\begin{pmatrix}\delta A_{i,1}-\delta A_{j,1}\\ \delta A_{i,2}-\delta A_{j,2}\end{pmatrix}^\top\boldsymbol{p}_j+\begin{pmatrix}A_{i,1}-A_{j,1}\\ A_{i,2}-A_{j,2}\end{pmatrix}^\top(k^p_{j,1}\,\boldsymbol{n})+\delta A_{i,3}-\delta A_{j,3}=o(\delta A).$$

Hence,

$$k^p_{j,1} = -\left(\begin{pmatrix}\delta A_{i,1} - \delta A_{j,1}\\ \delta A_{i,2} - \delta A_{j,2}\end{pmatrix}^\top \boldsymbol{p}_j + \delta A_{i,3} - \delta A_{j,3}\right)\left(\begin{pmatrix}A_{i,1} - A_{j,1}\\ A_{i,2} - A_{j,2}\end{pmatrix}^\top \boldsymbol{n}\right)^{-1} + o(\delta A).$$

A similar formula can be obtained for $k^q_{j,1}$. Finally, (2.5) implies

$$\begin{aligned}\delta\mu_2(D_i) &= \frac{1}{2}\sum_j \|\boldsymbol{p}_j - \boldsymbol{q}_j\|\,(k^p_{j,1} + k^q_{j,1}) + o(\delta A)\\ &= \frac{1}{2}\sum_j \|\boldsymbol{p}_j - \boldsymbol{q}_j\|\,\frac{(\delta A_{j,1} - \delta A_{i,1}, \delta A_{j,2} - \delta A_{i,2})\,(\boldsymbol{p}_j + \boldsymbol{q}_j) + 2\,(\delta A_{j,3} - \delta A_{i,3})}{(A_{i,1} - A_{j,1}, A_{i,2} - A_{j,2})\,\boldsymbol{n}}\\ &\quad + o(\delta A),\end{aligned}$$

where we sum over all j such that D_i and D_j share an edge. The limit $\delta A \to 0$ yields the claim.

Note that also topological changes are possible. Let us consider a vertex $\boldsymbol{p}$, such that $\boldsymbol{p}$ is contained in D_i for $i = 1, \dots, 4$, such that $D_4 = \{\boldsymbol{p}\}$ and $\mu_2(D_i) > 0$ for $i = 1, \dots, 3$, see Figure 2.2. Then there are perturbations, which induce a change in the topology, namely

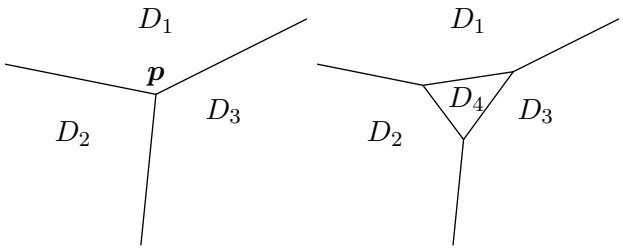

Figure 2.2: Topology change: after a small perturbation, the point $\boldsymbol{p}$ becomes the dominating region D_4.

f_4 becomes strictly active. However, $\mu_2(D_4)$ is only a small-o of the perturbation and the differentiability result remains valid. Another (differentiable) topology change happens if again $\boldsymbol{p}$ is contained in D_i, $i = 1, \dots, 4$ and $\mu_2(D_i) > 0$ for $i = 1, \dots, 4$. Then, after a perturbation, the vertex $\boldsymbol{p}$ may become two vertices and an edge. Note that if D_1 and D_2 share a common edge and if $D_3 = D_1 \cap D_2$ (i.e. D_3 equals this common edge), then $\mu_2(D_3)$ is not differentiable. However, this situation is excluded by Assumption 2.1.

This lemma enables us to prove the main result of this section.

Theorem 2.3. For $f(A)$ given by (2.1), (2.2), we have

$$\frac{\mathrm{d}J(f(A))}{\mathrm{d}(A_{i,1}, A_{i,2}, A_{i,3})} = \mu_2(D_i) \frac{-2}{(1 + A_{i,1}^2 + A_{i,2}^2)^2} \begin{pmatrix} A_{i,1} \\ A_{i,2} \\ 0 \end{pmatrix} + \sum_{j=1}^{n} \frac{\mathrm{d}\mu_2(D_j)}{\mathrm{d}(A_{i,1}, A_{i,2}, A_{i,3})} \frac{1}{1 + A_{j,1}^2 + A_{j,2}^2},$$

where the derivative of $\mu_2(D_i)$ is given in Lemma 2.2.

[Lachand-Robert and Oudet, 2005, Theorem 1] computed the derivative w.r.t. the coefficients $A_{i,3}$ in a slightly more general framework. However, in contrast to their implementation, we also use the derivative information w.r.t. $A_{i,1}$ and $A_{i,2}$. Moreover, we cannot share their opinion of using the full gradient information is of "little advantage", see [Lachand-Robert and Oudet, 2005, p. 372].

2.2 Details of the implementation

In this section we deal with two implementational issues neglected in the previous section: The treatment of the constraints and the computation of the dominating regions D_i as well as their adjacencies.

First we consider the constraint $f(x, y) \leq L$ for all $(x, y) \in \Omega$. We deal with this constraint by enforcing f_1 to be equal to L. In terms of coefficients, we fix

$$A_{1,1} = A_{1,2} = 0 \quad \text{and} \quad A_{1,3} = L.$$

The constraint $f(x, y) \geq 0$ for all $(x, y) \in \Omega$ is equivalent to $f_i(x, y) \geq 0$ for all $i = 1, \dots, n$ and all $(x, y) \in \Omega$. Since f_i attains its minimal value at

$$(x, y) = -\frac{(A_{i,1}, A_{i,2})}{\|(A_{i,1}, A_{i,2})\|},$$

we find this constraint equivalent to

$$A_{i,3} - \sqrt{A_{i,1}^2 + A_{i,2}^2} \geq 0 \tag{2.6}$$

for all $i = 1, \dots, n$.

It remains to discuss the computation of the dominating regions D_i. By definition, we have

$$A_{i,1}\, x + A_{i,2}\, y + A_{i,3} \geq A_{j,1}\, x + A_{j,2}\, y + A_{j,3} \quad \text{for all } (x, y) \in \Omega$$

for all $j = 1, \dots, n$. This is equivalent to

$$(x_i - x)^2 + (y_i - y)^2 - w_i \leq (x_j - x)^2 + (y_j - y)^2 - w_j \quad \text{for all } (x, y) \in \Omega \tag{2.7}$$

for all j, where

$$x_i = A_{i,1}/2, \quad y_i = A_{i,2}/2, \quad w_i = A_{i,3} + x_i^2 + y_i^2,$$

see also [Boissonnat et al., 2006, Section 2.3.3]. Hence, D_i is the region of points (x, y) whose power w.r.t. the weighted point (x_i, y_i, w_i), which is defined by the left-hand side of (2.7), is not greater than the power w.r.t. (x_j, y_j, w_j), $j = 1, \dots, n$. Therefore, the computation of the sets D_i can be transformed to the problem of computing the power diagram of the weighted points (x_i, y_i, w_i). Note that the power diagram is a generalization of the Voronoi diagram (set $w_i = 0$ for all i). Moreover, it is the dual of a weighted Delaunay triangulation (so called regular triangulation). We use the computational geometry library CGAL to compute this power diagram, see in particular Yvinec [2012].

2.3 Preliminary numerical results

Using the results from the previous two sections, we are able to implement a basic algorithm. Starting from a random initial guess A, we use MATLAB's optimization function `fmincon` to compute a minimizer of J. A possible solution for $n = 10$ and $L = 1$ is shown in Figure 2.3. The computational time was about 1 second.[1] Note that only 8 affine functions (including $f_1 \equiv L$) are strictly active, whereas the remaining functions are inactive.

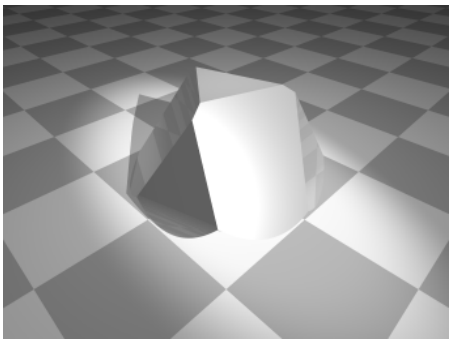

Figure 2.3: Solution with 8 active functions, and height $L = 1$.

[1] All computations were done using a computer with two Intel Xeon Dual Core CPU (4×3.0 GHz) with 16 GB RAM.

2.4 Refinement strategies

Given a solution $A \in \mathbb{R}^{n\times 3}$, one is interested in refining the solution by adding some degrees of freedom (i.e., by adding additional affine functions). In this section we discuss two refinement strategies which have been proven to be successful.

Boundary refinement

In Figure 2.3 we see that f attains local maxima on the boundary at those points, which are incident to an edge. This fact suggests the following refinement strategy.

- Determine all edges (vertices $\boldsymbol{p}_i$, $\boldsymbol{q}_i$) which are adjacent to the boundary. The vertex incident to the boundary is called $\boldsymbol{p}_i$.
- Construct a new affine function g which satisfies

$$g(\boldsymbol{p}_i) = 0, \qquad g((\boldsymbol{p}_i + \boldsymbol{q}_i)/2) = f(A, (\boldsymbol{p}_i + \boldsymbol{q}_i)/2), \qquad g \geq 0 \text{ on } \bar{\Omega}.$$

- Add the coefficients of g to A.

Note that g is unique determined by the three given conditions.

Tangential refinement

In addition to the previous strategy, which only refines near the boundary, we use a second one refining large interior cells D_i:

- Determine all cells D_i whose area $\mu_2(D_i)$ is larger than a given threshold.
- Replace the coefficients of the function f_i in A by the coefficients of two new affine functions, such that the cell D_i is split tangentially (e.g. through its center of mass) after the refinement.

2.5 Further improvements

In this section we describe shortly two further improvements which reduce the overall computational time.

Exploiting the symmetry

It is to be expected from results known from the literature, see Lachand-Robert and Peletier [2001], Lachand-Robert and Oudet [2005], that the solution f of (**P**) has the same symmetry group as a regular m-sided polygon (i.e., the dihedral group D_m). The symmetry parameter m depends on the height L and is monotone decreasing in L, see in particular [Lachand-Robert and Oudet, 2005, Figure 6], [Lachand-Robert and Peletier, 2001, Table 1 and Figure 1]. Given m, we have to carry out the computations only on $1/(2\,m)$ of the circle. For the same accuracy, we also need only $1/(2\,m)$ of the number of affine functions. Hence, with the same number of affine functions (and with a comparable computational time), we can achieve more accurate results, compare Figure 2.4 with Figure 2.3.

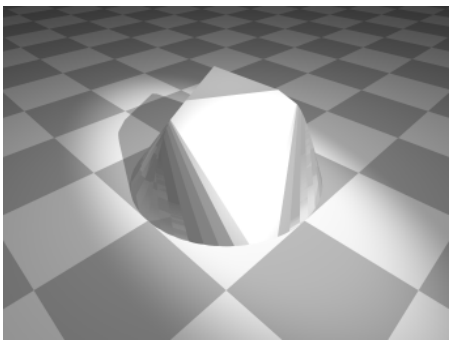

Figure 2.4: Solution with 8 affine functions, symmetry parameter $m = 3$, and height $L = 1$.

Transformation to simple bound constraints

The constraint (2.6) is a nonlinear constraint, which can be handled by typical optimization routines. However, the nonlinearity of (2.6) adds an additional difficulty. Introducing a slack variable $s_i \leq 0$ and substituting $A_{i,3}$ in the objective via

$$A_{i,3} = \sqrt{A_{i,1}^2 + A_{i,2}^2} - s_i$$

yields an optimization problem with simple bound constraints $s_i \leq 0$ and all nonlinearities are hidden in the objective. Note that the slack variable s_i has also a nice geometric interpretation, it is just the negative of the minimum value of f_i on $\partial\Omega$.

2.6 Numerical Results

In this section we present numerical results obtained by using the improvements of the previous section.

In Figure 2.5, the optimal shapes are shown. The computational times were between

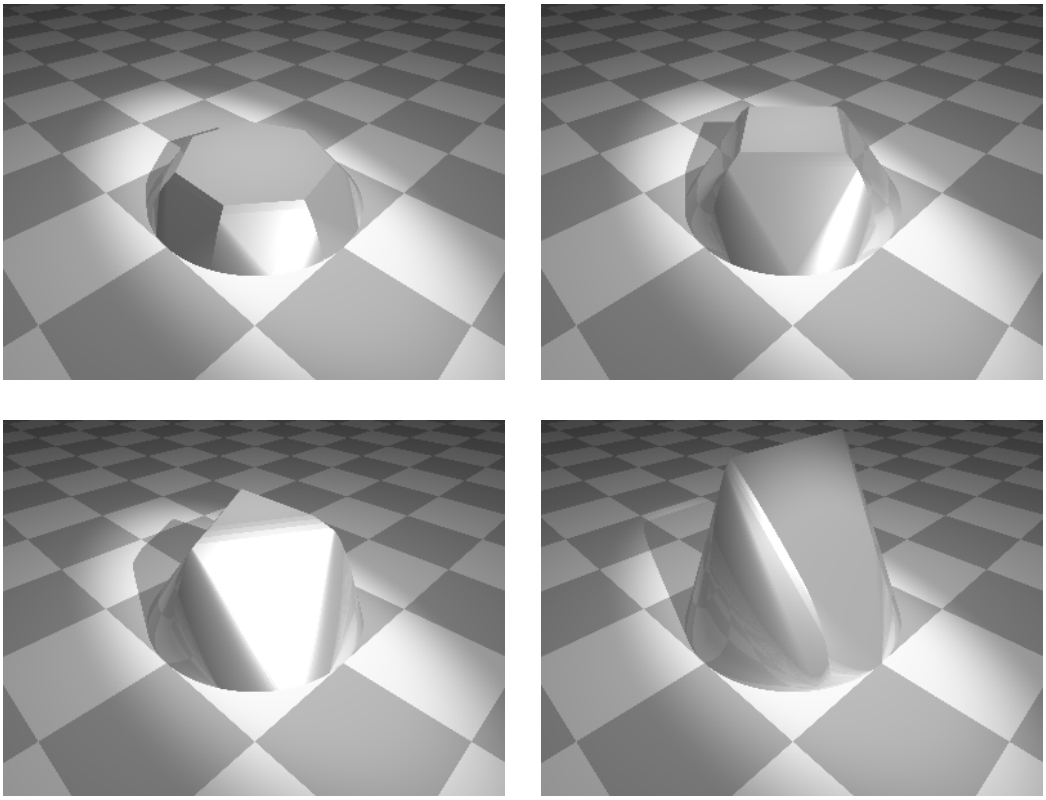

Figure 2.5: Solutions for $L = 0.4$ (top left), $L = 0.7$ (top right), $L = 1.0$ (bottom left) and $L = 1.5$ (bottom right)

2 and 5 minutes and the number of unknowns was below 600. For the values of the objective function we refer to Table 3.2 and for a comparison with results known from the literature, we refer to Section 4.

3 The solution in a set of symmetric and developable bodies

In this section we restrict problem **(P)** to a smaller class of functions. We will see that for a height L smaller than about 1.4, we obtain slightly better result than those of Section 2.6. Hence, it is to be conjectured that the minimizer of **(P)** for $L \in \{0.4, 0.7, 1.0\}$ (or more generally for $L \in (0, \bar{L})$ for some $\bar{L} > 0$) belongs to this class of functions.

In this section it is more appropriate to speak in terms of the body

$$P = \{(x, y, z) \in \mathbb{R}^3 : \ (x, y) \in \bar{\Omega}, \ 0 \le z \le f(x, y)\}$$

instead of the function f.

A similar approach was used in Lachand-Robert and Peletier [2001]. There, the authors considered bodies whose extremal points lie solely on the lower boundary $\partial\Omega \times \{0\}$ or on the upper boundary $N_0 \times \{L\}$ for some $N_0 \subset \mathbb{R}^2$. They showed that N_0 is a regular m-sided polygon (with m depending on L) which is centered at the origin $(0,0)$. However, [Lachand-Robert and Oudet, 2005, Table 1] shows that the minimizers of (**P**) does not belong to this class of functions.

In Figure 2.5 we could see that the optimal bodies (for $L \leq 1$) have the symmetry group D_m, see also Section 2.5. Let us define $\varphi = \pi/m$. Moreover, for $L \leq 1$, the extremal points of these bodies belong to the set

$$\partial\Omega \times \{0\} \cup \{(x,y,z) \in \mathbb{R}^3 : (x,y) = r\left(\cos(2\,i\,\varphi), \sin(2\,i\,\varphi)\right),\ i = 0,\dots,m-1,\ r \in [0,1]\}$$

Hence, the extremal points lie on the boundary $\partial\Omega$ (with height 0) or on the rays with angles $2\,i\,\varphi$, $i = 0,\dots,m-1$.

In order to describe the body P, it is therefore sufficient to choose a concave function $g : [0,1] \to [0,L]$ satisfying $g(0) = L$ and $g(1) = 0$. Then, the body P is given as the convex hull of the points

$$P = \operatorname{conv}\left(\partial\Omega \times \{0\} \cup \{(r\,\cos(2\,i\,\varphi), r\,\sin(2\,i\,\varphi), g(r)),\ i = 0,\dots,m-1,\ r \in [0,1]\}\right).$$

Finally, we could reconstruct the function f by

$$f(x,y) = \sup\{z : (x,y,z) \in P\}.$$

The expected shape of the optimal body P on $1/(2\,m)$ of Ω is depicted in Figure 3.1. The

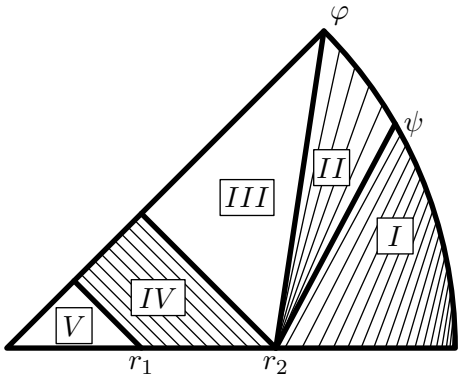

Figure 3.1: Structure of the solution

domain Ω is divided into five regions. In each region it is possible to express f (and P) in terms of the function g, see in particular Section 3.2. On the horizontal axis, the height of P is given by the function g, whereas on the remaining part it is implicitly defined via

the convex hull. The thin lines in this figure represent (the projections to $\mathbb{R}^2$ of) some *tangent lines*, i.e., they are intersections of tangent planes of P with its boundary ∂P. Note that the outer normal of P is constant on these tangent lines. The body P is flat in the regions III and V in Figure 3.1.

3.1 Assumptions

We will fix some assumptions on the optimal g in order to compute J in terms of g.

- $g \in C([0,1])$, g is concave, $g(0) = L$, and $g(1) = 0$,
- there are $0 \le r_1 \le r_2 \le 1$ such that

$$\begin{aligned} g|_{[0,r_1]} &\equiv L, \\ g|_{[r_1,r_2]} &\in C^2([r_1,r_2]), \\ g|_{[r_2,1]} &\in C^2([r_2,1]), \end{aligned}$$

- $\arg\max \frac{g(r)}{1-r\,\cos\varphi} = \{r_2\}$ (this implies that the flat region III will touch the x-axis only at $r = r_2$),
- $g'(r_2-) = \frac{-\cos(\varphi)\,g(r_2)}{1-r_2\,\cos(\varphi)}$ (this implies that regions III and IV will touch tangentially),
- $g'|_{(r_2,T)}$ is strictly monotone decreasing (this implies that the relation $r \leftrightarrow \alpha$ in region I is one-to-one, see Section 3.2).

The last three assumptions are not restrictive. They rather exclude some artificial cases and make the analysis a little bit easier.

3.2 Determination of J in terms of g

In this section we compute the contributions of the regions I–V to the objective J. To this end, we need some longish and tedious geometric computations. Therefore, some of them are sketched and presented in an abbreviated way.

First, we give a formula for the angle $\psi \in [0,\varphi]$. The angle ψ can be characterized as the smallest angle α, such that there is a tangent plane of P passing through the points $(r_2, 0, g(r_2))$ and $\boldsymbol{p}_\alpha = (\cos(\alpha), \sin(\alpha), 0)$.

Let such an angle α be given. The tangent plane of P passing through $\boldsymbol{p}_\alpha$ is given by

$$z = C_\alpha\, \boldsymbol{n}_\alpha^\top\, (\boldsymbol{n}_\alpha - (x,y)^\top),$$

for some $C_\alpha \geq 0$, where $\boldsymbol{n}_\alpha = (\cos(\alpha), \sin(\alpha))$. The point $(r_2, 0, g(r_2))$ lies on this tangent plane, whereas all other points $(r, 0, g(r))$, $r \in [0,1] \setminus \{r_2\}$ lie below this tangent plane. This yields

$$C_\alpha = \frac{g(r_2)}{1 - r_2 \cos(\alpha)}, \quad \text{and} \tag{3.1}$$
$$\frac{g(r)}{1 - r \cos(\alpha)} \leq \frac{g(r_2)}{1 - r_2 \cos(\alpha)} \quad \text{for all } r \in [0,1].$$

Since g is positive, decreasing and concave and $r \mapsto 1/(1 - r\cos(\alpha))$ is positive, increasing and concave, the product function is concave. Therefore, 0 belongs to the superdifferential of $r \mapsto g(r)/(1 - r\cos(\alpha))$ at r_2. Hence,

$$0 \in \frac{[g'(r_2+), g'(r_2-)]}{1 - r_2 \cos(\alpha)} + \frac{g(r_2)\cos(\alpha)}{(1 - r_2\cos(\alpha))^2}.$$

The smallest α, such that this condition is satisfied is characterized by $\psi \in [0, \varphi]$ and

$$\cos(\psi) = \frac{-g'(r_2+)}{g(r_2) - r_2\, g'(r_2+)}.$$

Note that in the case $g'(r_2-) = g'(r_2+)$, we have $\psi = \varphi$.

Region *I*

We show that there is a one-to-one relation between $r \in [r_2, 1]$ and $\alpha(r) \in [0, \psi]$, such that the there is a tangent plane of our body P which touches in the points $(r_2, 0, g(r_2))$ and $\boldsymbol{p}_{\alpha(r)} = (\cos(\alpha(r)), \sin(\alpha(r)), 0)$.

The tangent plane through $\boldsymbol{p}_{\alpha(r)}$ and above $\partial\Omega \times \{0\}$ is of the form

$$z = C_{\alpha(r)}\, \boldsymbol{n}_{\alpha(r)}^\top\, (\boldsymbol{n}_{\alpha(r)} - (x, y)^\top)$$

where $\boldsymbol{n}_{\alpha(r)} = (\cos(\alpha(r)), \sin(\alpha(r)))^\top$ and $C_{\alpha(r)} > 0$. The constant $C_{\alpha(r)}$ and the touching point r are determined by the requirement that the points $(s, 0, g(s))$ lie below this tangent plane:

$$g(s) \leq C_{\alpha(r)}\, \boldsymbol{n}_{\alpha(r)}^\top\, (\boldsymbol{n}_{\alpha(s)} - (r, 0)^\top) \text{ for all } s \in [0,1],$$

where equality holds for $s = r$. This yields

$$r = \operatorname*{arg\,max}_{s \in [0,1]} \frac{g(s)}{1 - s\cos(\alpha(s))},$$
$$C_{\alpha(r)} = \max_{s \in [0,1]} \frac{g(s)}{1 - s\cos(\alpha(s))} = \frac{g(r)}{1 - r\cos(\alpha(r))}.$$

Hence, we have (for $r = r_2$, we use the convention $g'(r) = g'(r_2+)$, and similarly for the second derivative)

$$\cos(\alpha(r)) = \frac{-g'(r)}{g(r) - r\,g'(r)}.$$

In order to compute the contribution to the objective, we consider a small area $\mathrm{d}A$, see Figure 3.3. Using the shoelace formula, we can compute the size of $\mathrm{d}A$. Up to higher

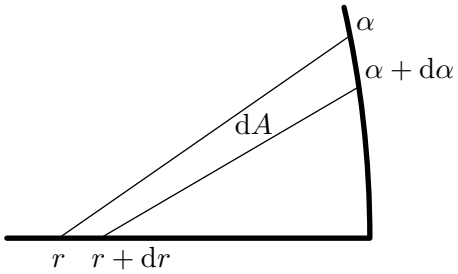

Figure 3.2: A small part $\mathrm{d}A$ of region I.

order terms, we find

$$\mathrm{d}A = \frac{1}{2}\left[\frac{\mathrm{d}\sin(\alpha(r))}{\mathrm{d}r}\left(\cos(\alpha(r)) - r\right) - \sin(\alpha(r))\left(\frac{\mathrm{d}\cos(\alpha(r))}{\mathrm{d}r} + 1\right)\right]\mathrm{d}r.$$

A simple calculation shows

$$\begin{aligned}
\sin(\alpha(r)) &= \sqrt{1 - \cos(\alpha(r))^2},\\
\frac{\mathrm{d}\cos(\alpha(r))}{\mathrm{d}r} &= \frac{-g''(r)\,g(r)}{(g(r) - r\,g'(r))^2},\\
\frac{\mathrm{d}\sin(\alpha(r))}{\mathrm{d}r} &= -\cot(\alpha(r))\,\frac{\mathrm{d}\cos(\alpha(r))}{\mathrm{d}r}.
\end{aligned}$$

The integrand on $\mathrm{d}A$ is given by (note that the gradient of f equals $C_{\alpha(r)}\,\boldsymbol{n}_{\alpha(r)}$ on $\mathrm{d}A$)

$$\frac{1}{1 + (C_{\alpha(r)})^2} = \frac{1}{1 + (g(r) - r\,g'(r))^2}$$

Finally, we arrive at

$$J_1 = \frac{1}{2}\int_{r_2}^{1} \frac{\frac{\mathrm{d}\sin(\alpha(r))}{\mathrm{d}r}\left(\cos(\alpha(r)) - r\right) - \sin(\alpha(r))\left(\frac{\mathrm{d}\cos(\alpha(r))}{\mathrm{d}r} + 1\right)}{1 + \left(g(r) - r\,g'(r)\right)^2}\mathrm{d}r$$

Since g is assumed to be twice continuously differentiable on $[r_2, 1]$, it satisfies the associated Euler-Lagrange-Equation (ELE). Since g'' enters the integrand (affine) linearly, the ELE is an ordinary differential equation of second order.

Region *II*

Let an angle $\alpha \in [\psi, \varphi]$ be given. In order to compute the contribution to the objective,

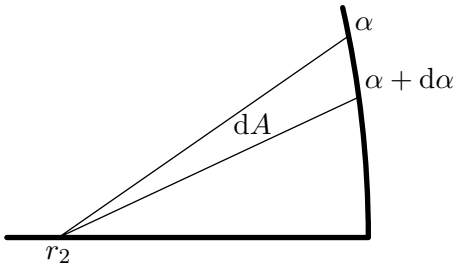

Figure 3.3: A small part $\mathrm{d}A$ of region II.

we consider a small part of the area $\mathrm{d}A$. We find (up to higher order terms) $\mathrm{d}A = (1 - r_2 \cos(\alpha))/2$. The integrand can be obtained similarly to region I (see also (3.1)) and we arrive at

$$J_2 = \frac{1}{2} \int_{\psi}^{\varphi} \frac{1 - r_2 \cos(\alpha)}{1 + (g(r_2)/(1 - r_2 \cos(\alpha))^2} \, \mathrm{d}\alpha$$

Region *III*

In this region, the function f is affine. Hence, the gradient is constant and the contribution to the objective can be computed as

$$J_3 = \frac{1}{2} \frac{r_2 \sin(\varphi) (1 - r_2 \cos(\varphi))}{1 + \big(g(r_2)/(1 - r_2 \cos(\varphi))\big)^2}.$$

Region *IV*

On this region, the body P is given as the convex hull of the points

$$\{(r, 0, g(r)) : r \in [r_1, r_2]\} \cup \{(r \cos(2\varphi), r \sin(2\varphi), g(r) : r \in [r_1, r_2]\}.$$

Similar to regions I and II, one can express the contribution to the objective as an integral over $r \in [r_1, r_2]$. One obtains

$$J_4 = \frac{1}{2} \sin(2\varphi) \cos^2(\varphi) \int_{r_1}^{r_2} \frac{r}{\cos^2(\varphi) + g'(r)^2} \, \mathrm{d}r$$

Similar as for region I, we know that the optimal function g satisfies the associated Euler-Lagrange-Equation (ELE). Again, the ELE is an ordinary differential equation of second order.

Region V

On this region, the gradient of f equals 0 and we obtain

$$J_5 = \frac{1}{2} r_1^2 \cos(\varphi) \sin(\varphi).$$

3.3 Algorithm

Using the results from the previous section, we find

$$J(f(P)) = 2\,m\,(J_1 + J_2 + J_3 + J_4 + J_5).$$

In order to find the optimal g, we propose the following strategy.

(i) Choose two parameters $g'(1) < 0$ and $R_2 \in [0, 1]$.

(ii) Integrate (backwards in time) the ELE from region I, until

$$r = R_2 \quad \text{or} \quad \frac{-g'(r)}{g(r) - r\,g'(r)} = \cos(\varphi)$$

is satisfied.

(iii) Set $r_2 = r$ and $g'(r_2-) = \frac{-\cos(\varphi)\,g(r_2)}{1-r_2\,\cos(\varphi)}$.

(iv) Integrate (backwards in time) the ELE from region IV until

$$r = 0 \quad \text{or} \quad f(r) = L$$

is satisfied and set $r_1 = r$.

(v) Compute the value of the objective

$$J(f(P)) = 2\,m\,(J_1 + J_2 + J_3 + J_4 + J_5).$$

Hence, the optimization problem is reduced to a problem with two real parameters $g'(1)$ and R_2. Furthermore, in the cases $L \leq 0.9$ or $L \geq 1.3$, the integration of the ELE from region I in step (ii) is stopped due to the second condition and hence, the parameter R_2 is not used. Hence, we have to optimize only w.r.t. $g'(1)$ in these cases.

The evaluation of J in terms of the two parameters $g'(1)$ and R_2 takes only about 0.05 seconds. Hence, the optimization w.r.t. these parameters can be done easily, e.g. via a simple bisection.

L	m	$g'(-1)$	R_2	J
1.0	3	-3.380	0.638	1.137751
0.7	4	-2.581		1.545461
0.4	4	-1.928		2.100331

Table 3.1: Optimized values for $g'(-1)$ and R_2 for different values of the height L.

3.4 Numerical results

In this section we present the results obtained by the algorithm described in Section 3.3. The optimized values for $g'(-1)$ and R_2 are shown in Table 3.1. The associated bodies are shown in Figure 3.4. Finally, we compare the values of the objective obtained in [Lachand-Robert and Oudet, 2005, Table 1] with those obtained of Section 2 and Section 3 in Table 3.2.

L	literature	Section 2	Section 3
1.5	0.7012	0.699923	
1.0	1.1379	1.137781	1.137751
0.7	1.5457	1.545487	1.545461
0.4	2.1006	2.099645	2.099606

Table 3.2: Comparison of the optimal values of J, which are known from [Lachand-Robert and Oudet, 2005, Table 1] (left column), obtained in Section 2 (middle column) and in Section 3 (right column).

4 Summary

In this paper we proposed two different numerical methods for Newton's problem of finding a concave function of least resistance.

In Section 2 we discretized a concave function by the infimum of a finite number of affine functions. Using the gradient of the objective w.r.t. the coefficients of the affine functions in an optimization solver together with reasonable refinement strategies yields an efficient algorithm. In a couple of minutes we are able to obtain slightly better results than Lachand-Robert and Oudet [2005]. Note that Lachand-Robert and Oudet [2005] does not report computational times. However, since their method includes a genetic algorithm, it is to be expected that their computational times are much larger. Moreover, for $L = 0.4$ we found that a body with symmetry group D_7 is better than the solution of Lachand-Robert and Oudet [2005], which has the symmetry group D_6.

From the results obtained in Section 2 we conjecture that the optimal body (for $L \in (0, L_0)$, for some $L_0 > 0$) belongs to a certain class, see Section 3. Exploiting the

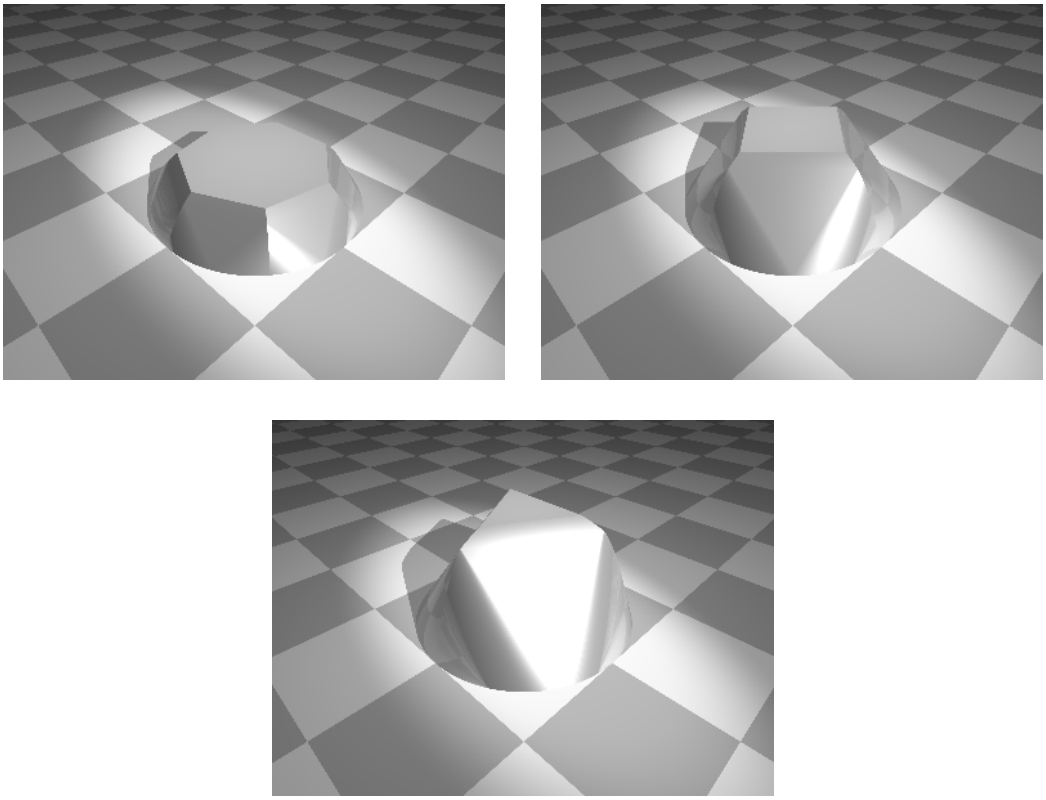

Figure 3.4: The optimal shapes determined by the algorithm of Section 3, for $L = 0.4$ (top left), $L = 0.7$ (top right), and $L = 1.0$ (bottom).

structure of this class of functions, Newton's problem could be reduced to a variational problem in $\mathbb{R}^1$. We are able to set up an algorithm which computes the optimal body in a few seconds. Moreover, the objective values are a little bit smaller than those obtained in Section 2. This consolidates the conjecture that the optimal body indeed belongs to the considered class of functions.

Further numerical experiments suggest that the optimal body belongs to the class considered in Section 3 if and only if this body has the symmetry group D_n with $n \geq 3$. Moreover, the transition of the symmetry group from D_2 to D_3 happens for L between 1.4 and 1.5.

More results for various heights L are given in Table 4.1. For $m = 2$ and $L = 1.4$ we obtain the objective value 0.773696 using the algorithm of Section 2, which is worse than the value obtained with $m = 3$. Similar to the results in Lachand-Robert and Peletier [2001], the optimal value of m seems to be monotone decreasing in dependence of L.

$L\backslash m$	3	4	5	6	7	8
1.5	0.6999493					
1.4	**0.7677364**	0.7792766				
1.3	**0.8441426**	0.8544164				
1.2	**0.9303614**	0.9387926				
1.1	**1.027729**	1.033598				
1.0	**1.137751**	1.140152				
0.9	1.261989	**1.259905**	1.265067			
0.8	1.402272	**1.394439**	1.397488			
0.7	1.560453	**1.545461**	1.545781			
0.6		1.714798	**1.711683**	1.712854		
0.5			1.897056	**1.896107**	1.896892	
0.4				2.100331	**2.099606**	2.099854

Table 4.1: Comparison of the optimal values of J for different values of the height L and the symmetry m obtained by the algorithm described in Section 3. Bold values highlight the optimal symmetry for given height L. Blank entries means that the value was not computed.

Hence, the symmetry D_m is optimal for $\hat{M}_m \le L \le \hat{M}_{m+1}$, with $\hat{M}_m$ given by

$$\begin{aligned} 1.4 \le \hat{M}_3 \le 1.5, \qquad & M_3 \approx 1.179535875, \\ 0.9 \le \hat{M}_4 \le 1.0, \qquad & M_4 \approx 0.754344515, \\ 0.6 \le \hat{M}_5 \le 0.7, \qquad & M_5 \approx 0.561232469, \\ 0.5 \le \hat{M}_6 \le 0.6, \qquad & M_6 \approx 0.447571675, \\ 0.4 \le \hat{M}_7 \le 0.5, \qquad & M_7 \approx 0.372163842. \end{aligned}$$

Note that the values $\hat{M}_m$ are significantly larger than M_m, which have same meaning for the solution in the class of developable functions, see [Lachand-Robert and Peletier, 2001, Table 1].

The structure of the solution in the case $m = 2$ remains an open question.

References

J.-D. Boissonnat, C. Wormser, and M. Yvinec. Curved voronoi diagrams. In J.-D. Boissonnat and M. Teillaud, editors, *Effective Computational Geometry for Curves and Surfaces*, pages 67–116. 2006. doi: 10.1007/978-3-540-33259-6_2.

F. Brock, V. Ferone, and B. Kawohl. A symmetry problem in the calculus of variations. *Calculus of Variations and Partial Differential Equations*, 4(6):593–599, 1996. doi: 10.1007/BF01261764.

G. Buttazzo and B. Kawohl. On Newton's problem of minimal resistance. *The Mathematical Intelligencer*, 15(4):7–12, 1993. doi: 10.1007/BF03024318.

G. Buttazzo, V. Ferone, and B. Kawohl. Minimum problems over sets of concave functions and related questions. *Mathematische Nachrichten*, 173:71–89, 1995. doi: 10.1002/mana.19951730106.

T. Lachand-Robert and É. Oudet. Minimizing within convex bodies using a convex hull method. *SIAM Journal on Optimization*, 16(2):368–379 (electronic), 2005. ISSN 1052-6234. doi: 10.1137/040608039.

T. Lachand-Robert and M. A. Peletier. Newton's problem of the body of minimal resistance in the class of convex developable functions. *Mathematische Nachrichten*, 226: 153–176, 2001.

CGAL. Computational Geometry Algorithms Library. http://www.cgal.org.

M. Yvinec. 2D triangulations. In *CGAL User and Reference Manual*. CGAL Editorial Board, 4.1 edition, 2012. http://www.cgal.org/Manual/4.1/doc_html/cgal_manual/packages.html#Pkg:Triangulation2.