UCLouvain

École polytechnique de Louvain

Newton's problem

A smooth minimum method

Author: **Nizar Bousselmi**
Supervisor: **François Glineur**
Readers: **Yurii Nesterov, Loïc Van Hoorebeeck**
Academic year 2020-2021
Master [120] in Mathematical Engineering

Acknowledgements

I would like to thank Prof. François Glineur for his precious and insightful help and guidance.

I also thank my colleague Andres for his careful review.

Abstract

Newton's problem consists in minimizing the resistance of a fluid on a body. It has been proposed by Sir Isaac Newton in 1685 and has been widely studied since. Yet, the optimal solution is still unknown. Many numerical and analytical methods were tested and some of them provided excellent results.

In this thesis, we propose a new method to solve Newton's problem based on the analysis of the previous work of the literature. It uses the smooth minimum function, namely, the *LogSumExp* function. Its smoothness allows in some sense to easily evaluate and optimize the resistance of the body. However, the method suffers from technical difficulties.

At the end, the method provides bodies with similar properties than the bodies obtained in the literature. In particular, the best resistances of the literature are around 0.4% better than our results obtained with the smooth method.

Contents

Chapter 1

Introduction

1.1 Presentation of the problem

In 1685 [1], Sir Isaac Newton proposed its minimal resistance problem. Given a body moving through a homogeneous fluid, find the shape of the body that minimizes its resistance with the fluid. Even if it sounds as a typical fluid mechanics problem, the additional hypotheses make the problem interesting and challenging from an optimization point of view.

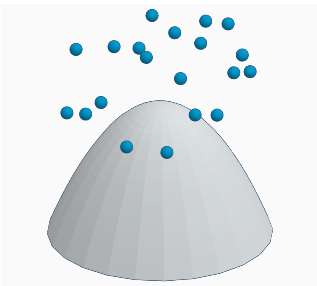


Fig. 1.1 Body moving through a fluid.

1.1.1 Hypotheses of the problem

It is necessary to add several assumptions to have a well-posed problem.

Assumption 1 The fluid is composed of particles that hit the body exactly once with perfectly elastic shocks. These hits are the only interactions between the body and the fluid, i.e., we neglect all other possible fluid mechanics effects like tangential friction, vorticity or turbulence.

Assumption 2 The body has a circular basis and moves orthogonally to this basis.

Assumption 3 The body has a fixed maximal height.

Assumption 4 The body must be a convex set.

We will shortly describe the resistance of the body. For the moment, it is already sufficient to know that, roughly speaking, a body with steep faces has a low resistance.

Some comments can be made about the assumptions. Assumption 1 makes the problem not physically realistic in general. Yet, it is still realistic for an ideal gas and a body at low speed [2]. Assumption 2 is just practical. Assumption 3 allows to have a well-posed problem since without a height limit, we can have a body infinitely high, with faces infinitely steep and resistance tending to zero.

Convexity assumption Assumption 4 is the most important and is the core of the problem, both in terms of difficulty and interest. Moreover, without convexity, there are bodies which make no sense with the given physical assumptions.

An illustrative example, from [3], is a body described by a sine function with a fixed angular frequency ω. Fig. 1.2 is such a body with $\omega = 10$.

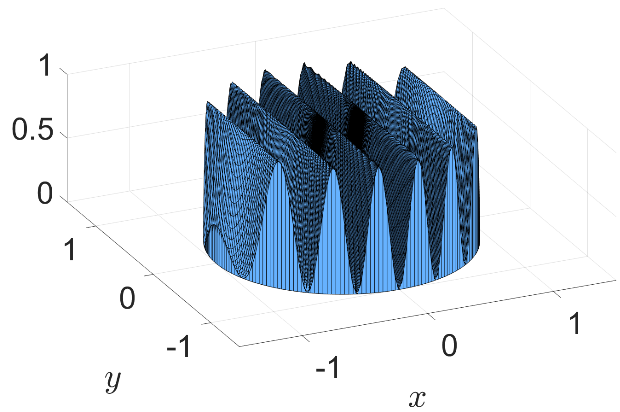


Fig. 1.2 Body described by a sine.

Increasing the frequency ω will make the body steeper and steeper everywhere and its resistance will tend to zero. Actually, such a body would strongly suffer from the accumulation of the particles on its holes. However, we stated from Assumption 1 that the only interaction of the fluid on the body was the impact of the particles. In reality, it would clearly not be the case with the body in Fig. 1.2. The particles would hit the body more than once as it is illustrated in Fig. 1.3.

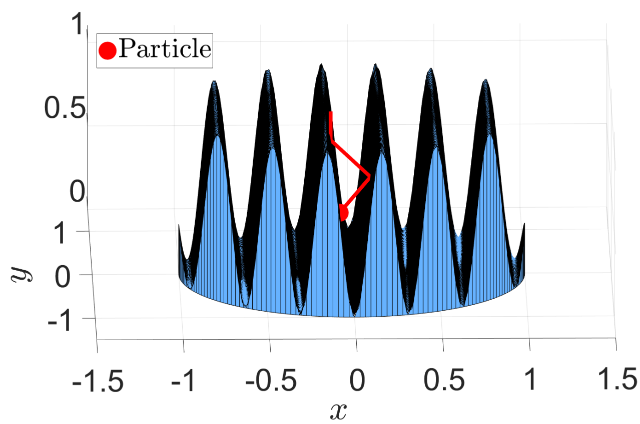


Fig. 1.3 Particle hitting more than once a non-convex body.

The convex body hypothesis guarantees that the particles hit the body once and never interact anymore with the body. However, as we will see, the constraint of convexity is challenging to manage.

Nevertheless, we can mention that imposing the convexity is not the unique way to ensure that the particles hit the body only once. Actually, it is a rather restrictive condition. A «single-impact assumption» has been derived [4] and this case is also studied in the literature [5, 6, 7].

1.1.2 Formulation of the problem

We write the problem in the optimization framework. It requires to introduce some notations and conventions to describe the body.

The basis of the body is the centred unit disk

$$\Omega = \left\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\right\} \subseteq \mathbb{R}^2. \tag{1.1}$$

The body goes upward and its maximal height is noted M where $M \geq 0$ is a parameter. By convention, the body lies between the $z = 0$ and $z = M$ planes.

Formal description of the body The body is described by two equivalent and convenient ways.

- the function u, the shape of the body

$$u : \Omega \to \mathbb{R} : (x, y) \to u(x, y) \tag{1.2}$$

- the set $U \subset \mathbb{R}^3$ of all the points (x, y, z) of the body

$$U = \left\{(x, y, z) \in \mathbb{R}^3 \mid 0 \leq z \leq u(x, y),\ (x, y) \in \Omega\right\} \tag{1.3}$$

Importantly, we must observe that the convexity of the body means that the set U is a convex set and that the function u is a concave function.

Proposition 1. *The set U is a convex set if and only if the function u is a concave function.*

We define the sets of concave functions u and convex sets U.

- set of concave functions u

$$\mathcal{C} = \{u : \Omega \to \mathbb{R} \mid u \text{ concave}\} \tag{1.4}$$

- set of convex sets U

$$\mathcal{S} = \{U \subset \mathbb{R}^3 \mid U \text{ convex}\} \tag{1.5}$$

As a body can be equivalently described by a function u or a set U, there exists a bijection between the sets $\mathcal{C}$ and $\mathcal{S}$. In this work, we will equivalently use one description or the other. Fig. 1.4 is an illustration of the function u and the sets U and Ω for some arbitrary body. We can indeed observe that u is a concave function and that U is a convex set.

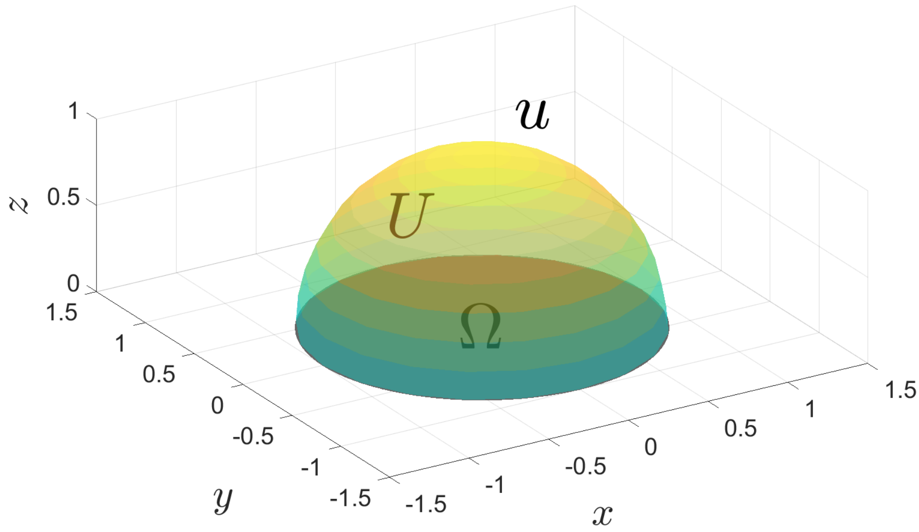


Fig. 1.4 Function u and sets U and Ω.

Expression of the resistance The objective function of the problem is the resistance of the fluid on the body. The particles hit the body vertically, once, and perfectly elastically. The pressure of the fluid under these assumptions is prescribed by the Newtonian sine-squared pressure law [3]. It states that the pressure p of the particle on the body is proportional to the sine-squared of the slope s of the body.

Given our description of the body with a function u, the pressure p on the point $(x, y) \in \Omega$ is

$$p = \sin^2(s) = \frac{1}{1 + ||\nabla u(x, y)||^2} \tag{1.6}$$

and the total resistance $\mathcal{F}$ of all the particles over the whole domain Ω is

$$\mathcal{F}(u) = \int_\Omega \frac{1}{1 + ||\nabla u(x, y)||^2} \mathrm{d}x\mathrm{d}y. \tag{1.7}$$

We will sometimes use the function $f(t) = \frac{1}{1+t^2}$ to express the resistance as

$$\mathcal{F}(u) = \int_\Omega f\big(||\nabla u||\big) \mathrm{d}x\mathrm{d}y. \tag{1.8}$$

Constraints We already mentioned the convexity constraint, we do not develop it anymore. Concerning the maximal height of the body, we write

$$0 \leq u(x, y) \leq M \quad \forall (x, y) \in \Omega. \tag{1.9}$$

We will often abbreviate the notation (1.9) by $0 \leq u \leq M$.

Optimization problem The whole optimization problem can be stated as follows

$$\begin{aligned} \min_{u \in \mathcal{C}} \quad & \int_\Omega \frac{1}{1 + ||\nabla u||^2} \mathrm{d}x\mathrm{d}y \\ \text{s.t.} \quad & 0 \leq u(x, y) \leq M, \quad \forall (x, y) \in \Omega. \end{aligned} \tag{1.10}$$

At first sight, there are two practical difficulties to solve problem (1.10).

- The decision variable of the optimization problem (1.10) is the function u. If we do not restrict the set of admissible functions $\mathcal{C}$, u has infinitely many parameters. Moreover, the functional $\mathcal{F}$ is not convex [3].
- The constraint of convexity of the body cannot be written in a practical way. In general, we must ensure that the function u is concave, namely, $\forall t \in [0,1]$ and $\forall (x_1, y_1), (x_2, y_2) \in \Omega$
$$u\big(t(x_1,y_1) + (1-t)(x_2,y_2)\big) \geq tu(x_1,y_1) + (1-t)u(x_2,y_2). \tag{1.11}$$

Therefore, it is primary to find a suitable subset $X \subset \mathcal{C}$ of concave functions with a finite number of parameters to solve the Newton's problem in practice.

1.1.3 Radially symmetric solution

Problem (1.10) expressed in polar coordinates is

$$\begin{aligned} \min_{u\in\mathcal{C}} \quad & \int_\Omega \frac{r}{1+||\nabla u||^2} \mathrm{d}r\mathrm{d}\theta \\ \text{s.t.} \quad & 0 \leq u(r,\theta) \leq M, \quad \forall (r,\theta) \in [0,1]\times[0,2\pi]. \end{aligned} \tag{1.12}$$

We can remark that problem (1.12) is symmetric with respect to the variable θ. This radial symmetry may suggest that the optimal solution is also radially symmetric and that we are just looking for the radial profile $u(r)$ for $r \in [0,1]$ that defines the body for all θ, i.e.,

$$u(r,\theta) = u(r) \qquad \forall (r,\theta) \in [0,1]\times[0,2\pi]. \tag{1.13}$$

In this case, the Newton's problem in polar coordinates (1.12) becomes

$$\begin{aligned} \min_{u:[0,1]\to\mathbb{R}} \quad & 2\pi \int_0^1 \frac{r}{1+|u'(r)|^2} \mathrm{d}r \\ \text{s.t.} \quad & u \text{ decreasing}, \\ & u(0) = M, \\ & u(1) = 0. \end{aligned} \tag{1.14}$$

It was the case studied by Newton. He found the optimal radially symmetric body. This solution can be expressed [3] in a parametric form. Defining the function

$$g(t) = \frac{1}{(1+t^2)^2}\bigg(-\frac{7}{4} + \frac{3}{4}t^4 + t^2 - \log(t)\bigg) \tag{1.15}$$

and the parameters $T = g^{-1}(M)$ and $r_0 = \frac{4T}{(1+T^2)^2}$, the profile $u(r)$ of the optimal radially symmetric Newton's solution parametrized by $t \in [1,T]$ is

$$r(t) = \frac{r_0}{4t}(1+t^2)^2 \tag{1.16}$$

$$u(t) = M - \frac{r_0}{4}\bigg(-\frac{7}{4} + \frac{3}{4}t^4 + t^2 - \log(t)\bigg). \tag{1.17}$$

However, it occurs that this solution is not optimal in general. There exist bodies with better resistance $\mathcal{F}$ that the optimal radially symmetric body found by Newton. We show the Newton's solution and other bodies in the next section.

1.1.4 Examples of bodies

We propose to compute the resistance $\mathcal{F}$ of simple bodies, namely, the cylinder, the cone and the half-sphere in the general case of height M. For each body, we provide its radial profile $u(r)$, the squared derivative of this profile $|u'(r)|^2$ and its resistance $\mathcal{F}$. The latter is computed thanks to its expression in the radial case , i.e.,

$$\mathcal{F}(u) = 2\pi \int_0^1 \frac{r}{1+|u'(r)|^2} \mathrm{d}r. \tag{1.18}$$

Resistance of a cylinder

$$u_{cyl}(r) = M \qquad \forall r \in [0,1] \tag{1.19}$$
$$|u'_{cyl}(r)|^2 = 0 \qquad \forall r \in [0,1] \tag{1.20}$$
$$\mathcal{F}(u_{cyl}) = 2\pi \int_0^1 \frac{r}{1+0} \mathrm{d}r = \pi. \tag{1.21}$$

Resistance of a cone

$$u_{cone}(r) = M(1-r) \qquad \forall r \in [0,1] \tag{1.22}$$
$$|u'_{cone}(r)|^2 = M^2 \qquad \forall r \in [0,1] \tag{1.23}$$
$$\mathcal{F}(u_{cone}) = 2\pi \int_0^1 \frac{r}{1+M^2} \mathrm{d}r = \frac{\pi}{1+M^2}. \tag{1.24}$$

Observe that the cone is an example of body with resistance $\mathcal{F}$ tending to zero when M increases infinitely.

Resistance of a stretched half-sphere of height M and base radius 1

$$u_{h\text{-}s}(r) = M\sqrt{1-r^2} \qquad \forall r \in [0,1] \tag{1.25}$$
$$|u'_{h\text{-}s}(r)|^2 = \frac{M^2 r^2}{1-r^2} \qquad \forall r \in [0,1] \tag{1.26}$$
$$\mathcal{F}(u_{h\text{-}s}) = 2\pi \int_0^1 \frac{r}{1+\frac{M^2 r^2}{1-r^2}} \mathrm{d}r = \pi \frac{2M^2 \log(M) - M^2 + 1}{M^2-1}. \tag{1.27}$$

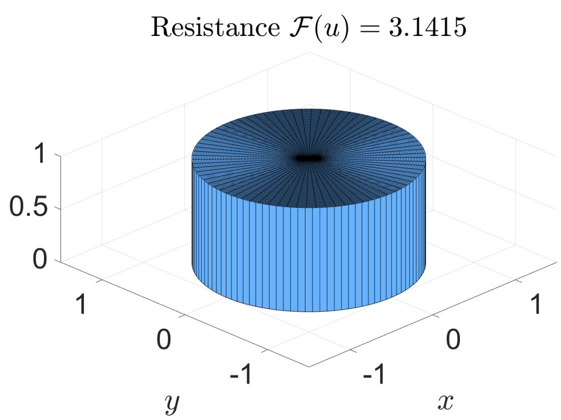
The resistance of the half-sphere of height $M = 1$ is $\mathcal{F}(u_{h\text{-}s}) = \frac{\pi}{2}$.

Table 1.1 contains the three resistances computed.

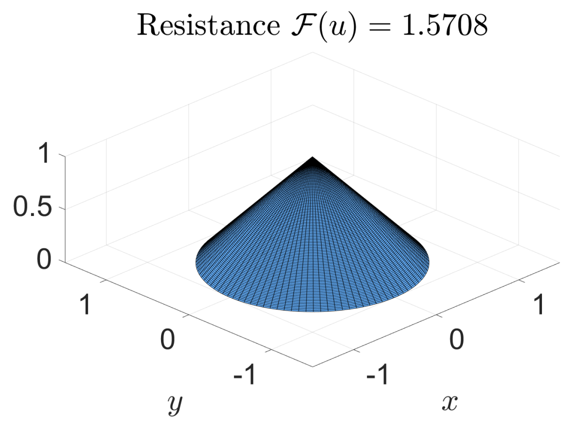
Table 1.1 Resistances $\mathcal{F}$ of the cylinder, cone and half-sphere for general height M.

	Cylinder	Cone	Half-sphere
Resistance $\mathcal{F}$	π	$\frac{\pi}{1+M^2}$	$\pi\frac{2M^2\log(M)-M^2+1}{M^2-1}$

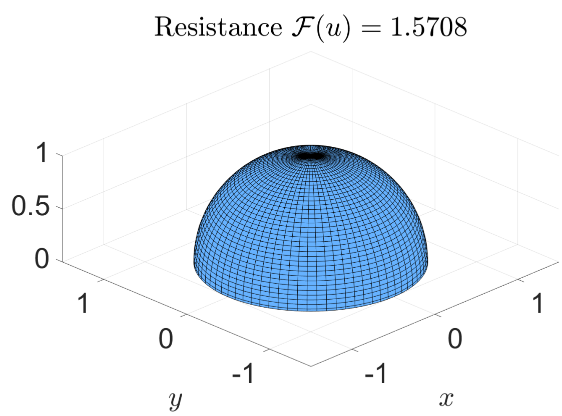
Now, in order to illustrate, Fig. 1.5 and Fig. 1.6 provide examples of different bodies with height $M = 1$ and their associated resistance.



(a) Cylinder.

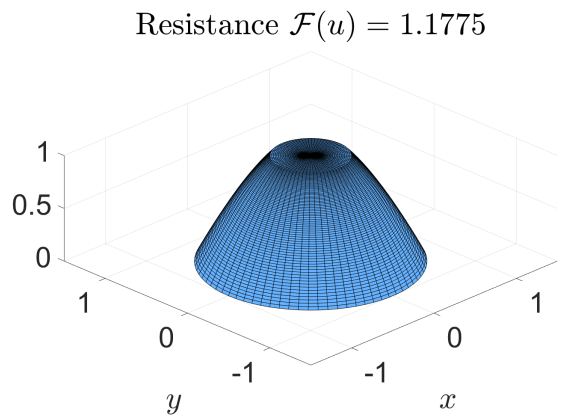


(b) Cone.

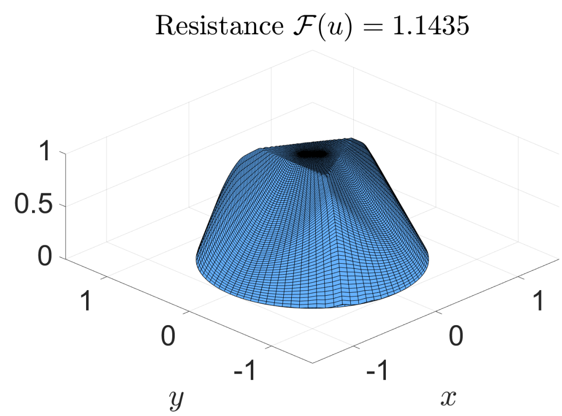


(c) Half-sphere.

Fig. 1.5 Examples of admissible bodies (1).



(a) Radially symmetric solution obtained by Newton.



(b) Triangular body.

Fig. 1.6 Examples of admissible bodies (2).

The cylinder, Fig. 1.5a, is the worst possible convex body. It does not use any of the available height to produce steepness.

The radially symmetric Newton's solution, Fig. 1.6a, has already a quite surprising shape. It is flat from the center up to a radius r_0 and then it is almost conical from r_0 up to the radius 1. At first sight, we could think that the cone, Fig. 1.5b, is better since a steeper face yields a lower resistance. However, by loosing some steepness before r_0, we earn much more steepness after r_0 such that we end up with the Newton's solution better than the cone.

Again surprisingly, the optimal radially symmetric Newton's solution is not optimal in general. The body in Fig. 1.6b has a better resistance. Its shape is also peculiar and we will study it later.

1.2 Related problems

As mentioned, one of the crucial characteristics of the Newton's problem is the constraint of convexity of the body. There exist several problems with such property [8, 9]. We mention the Cheeger's problem.

Cheeger's problem [10, 11] Given a bounded set $\Omega \subset \mathbb{R}^n$, find the subset $U \subseteq \Omega$ that minimizes the quotient of the perimeter of U by its surface. The perimeter of U is formally defined by the Caccioppoli definition [12]. It has been shown [13] that the optimal set U is convex if the domain Ω is convex. Moreover, the problem can be formulated as the following optimization problem [14]

$$\begin{aligned} \min_{u} \quad & \frac{\int_\Omega ||\nabla u|| \mathrm{d}V}{\int_\Omega |u| \mathrm{d}V} \\ \text{s.t.} \quad & u \text{ concave.} \end{aligned} \tag{1.28}$$

Fig. 1.7 is an illustration, from [9], of the outer sets Ω and inner sets U.

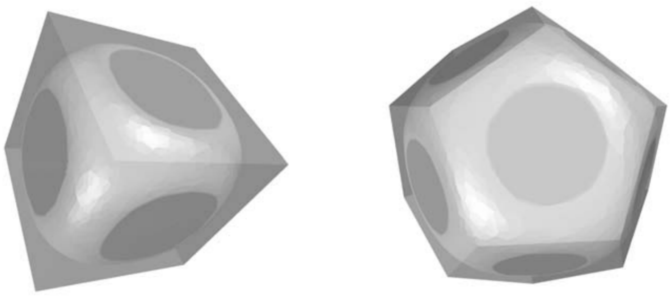


Fig. 1.7 Cheeger's sets (from [9]).

In general, a method designed to solve Newton's problem (1.10) should be adaptable to solve other problems with the convexity constraint like the Cheeger's problem (1.28). It is mainly the objective function that differs from one problem to another. As such problem deals with bodies U, their objective function always involves the integral of some function $f(u)$ over the domain Ω.

Therefore, even if the optimal solution of the Newton's problem has not a wide range of practical applications, finding a technique to solve a problem with the convexity constraint allows to solve several other problems.

1.3 Plan of the thesis

In Chapter 2, we provide a review of the literature. More precisely, we present some theoretical results and then study the different numerical methods proposed and try to isolate their strengths and weaknesses. In Chapter 3, we present our new method to tackle the problem. In Chapter 4, we cover some practical aspects of our method. In Chapter 5, we present the numerical results obtained and compare them to the literature.

List of Figures

List of Tables

Chapter 2

State of the art

Currently, the Newton's problem is open. Yet, there exist numerous theoretical results. We present some of them, namely, the existence, non-uniqueness and non-strictly concavity of the optimal body u^*, a property on the gradient ∇u^* of the optimal body and the decrease of the optimal resistance $\mathcal{F}^*$ with the height M of the body. Moreover, we present a known and interesting conjecture and a practical observation.

Afterwards, we analyze the different methods already tested on Newton's problem. These methods mix geometric, analytic and numerical approaches. Inevitably, tackling the problem requires to restrict the admissible domain $\mathcal{C}$ of functions u to a sub-domain of functions that can be parametrized by a finite number of variables and such that the concavity is effortlessly guaranteed. The stake is to reduce the domain to a type of functions easily handled while keeping a sufficiently large set of functions. We present the idea, the modelization and the numerical solutions for each method proposed in the literature. At the end, we analyze the strengths and weaknesses of these methods.

2.1 Theoretical results

Theorem 1 (Existence of a solution u^* [16]). *There exists an optimal solution $u^* \in \mathcal{C}$ to the Newton's problem* (1.10)*.*

It is not clear at first sight if the minimizer of the functional $\mathcal{F}(u)$ is reached for some admissible function $u \in \mathcal{C}$. Theorem 1 states that it is the case, the minimizer $u^* \in \mathcal{C}$ exists.

Theorem 2 (Non-uniqueness of the solution u^*). *There are infinitely many optimal solutions u^* to the Newton's problem* (1.10)*.*

Since the solution u^* is not radially symmetric, almost any rotation of u^* provides a different body with the exact same resistance as u^*.

Theorem 3 (Non-strictly concavity of the solution u^* [17]). *Let an open set $\omega \subseteq \Omega \subseteq \mathbb{R}^2$ such that u^* is twice continuously differentiable on ω, if u^* solves the Newton's problem* (1.10)*, then, $det(\nabla^2 u^*) = 0$ everywhere on ω, i.e., $\nabla^2 u^*$ has a zero eigenvalue everywhere on ω.*

Note that a concave function u with a zero eigenvalue is not always non-strictly concave. However by simplicity, we will refer to Theorem 3 as the *non-strictly concavity* of the solution.

Remark also that the theorem assumes the twice differentiability of the solution. It does not indicate anything on the behavior of the optimal solution where it is not twice differentiable. Moreover, there is no proof or suggestion of the global twice differentiability of the optimal solution.

Theorem 3 states that the optimal body u^* is locally developable. We propose three different and equivalent interpretations of a developable body.

- Geometrically, a developable body is defined by a developable surface. A developable surface is a surface where each point of the surface belongs to some straight lines, called generatrices. These generatrices entirely belong to the surface. Moreover, all the points of a generatrices share the same tangent plane.
- Analytically, a developable surface has a zero Gaussian curvature. The Gaussian curvature K of a function $u(x, y)$ on the point (x, y) is the determinant of the Hessian of u on this point, i.e.,
$$K = \det\left(\nabla^2 u(x, y)\right) = 0. \tag{2.1}$$
- Intuitively, a developable surface can roll without sliding on its tangent plane and the contact is made on a single line. Also, we can flatten a developable surface on a 2d sheet without distortion.

For instance, the cylinder and the cone, Fig. 1.5a and Fig. 1.5b, have a zero Gaussian curvature whereas the half-sphere, Fig. 1.5c, has a non-zero Gaussian curvature. Moreover, the radially symmetric solution of Newton, Fig. 1.6a, is strictly concave on its lateral face. We can observe it in Fig. 2.1 that shows the radial profile $u(r)$ of the Newton's solution (in blue) and a tangent line (in red). Newton's solution could definitely not be optimal.

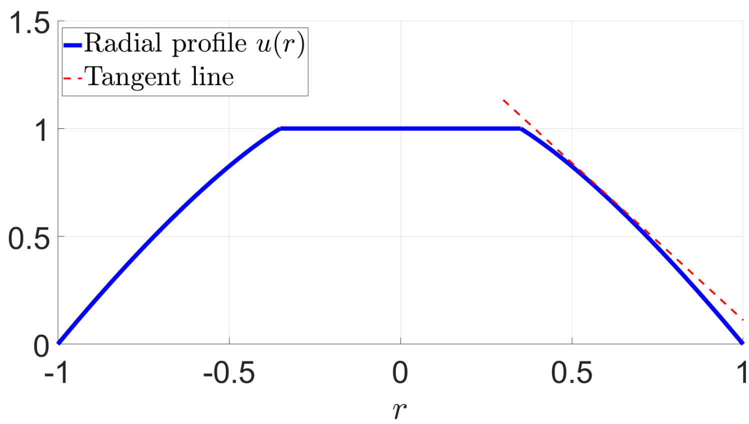


Fig. 2.1 Radial profile $u(r)$ of the Newton's solution.

As the authors enlightened, this non-strictly concavity implies that the body U^* contains a segment in ω. However, Theorem 3 does not oblige the local linearity of u^*.

Theorem 4 (Excluded values of the gradient of the solution ∇u^* [4])**.** *If u^* solves Newton's problem* (1.10)*, then,*

$$||\nabla u^*(x, y)|| \notin (0, 1) \qquad \forall (x, y) \in \Omega. \tag{2.2}$$

Bodies with a top flat face with $\nabla u = 0$, like the Newton's solution in Fig. 1.6a, often arise in the literature. We already discussed in Sect. 1.1.4 that it is interesting to have a lower slope, $\nabla u = 0$, on some points in order to have a higher slope, $\nabla u \geq 1$, on other points. Moreover, note that Theorem 4 is due to the concaveness of the function $f(t) = \frac{1}{1+t^2}$ around $t = 0$. It «attracts» t to 0. Fig. 2.2 is the function $f(t)$ (in black) and its interval of concaveness (in red) that is $\left[-\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right]$.

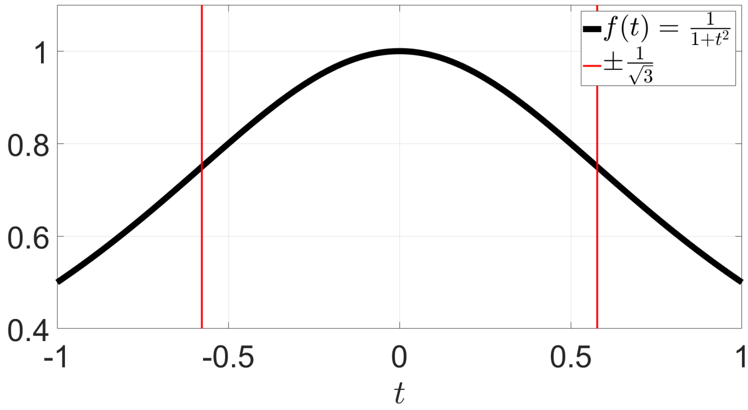


Fig. 2.2 Function $f(t) = \frac{1}{1+t^2}$.

Theorem 5. *The optimal resistance $\mathcal{F}^*$ is a decreasing monotone function of the height M of the body.*

Consider the optimal body u^*_M of height M and its resistance $\mathcal{F}(u^*_M) = \mathcal{F}^*_M$. Increasing the height up to $M + \epsilon \geq M$ allows to vertically scale the body u^*_M and therefore obtaining a body $\tilde{u} = \frac{M+\epsilon}{M} u^*_M$ with a strictly lower resistance $\mathcal{F}(\tilde{u})$. Note that the body $\tilde{u}$ is not optimal for the height $M + \epsilon$, i.e.,

$$\mathcal{F}^*_{M+\epsilon} < \mathcal{F}(\tilde{u}) < \mathcal{F}^*_M. \tag{2.3}$$

We discuss the non-optimality of the scaled body $\tilde{u}$ later with a conjecture.
Fig. 2.3 is the current best known values [15] of resistance $\mathcal{F}_M$ (blue dots) for different heights M.

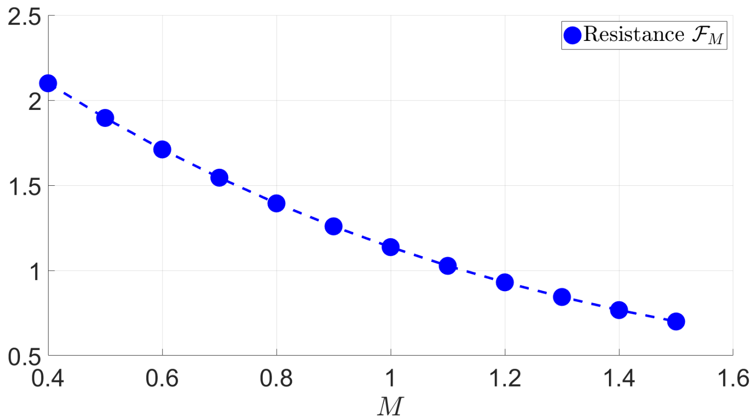


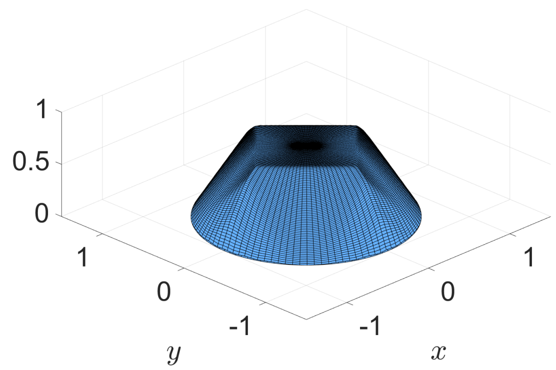
Fig. 2.3 Evolution of the best known resistance $\mathcal{F}_M$ as a function of the height M.

Note that we trivially know that $\mathcal{F}_0^* = \pi \approx 3.1416$ since the only admissible body of height $M = 0$ is the cylinder of height zero. Also, we already mentioned the case of infinite height M, i.e., $\mathcal{F}_\infty^* = 0$.

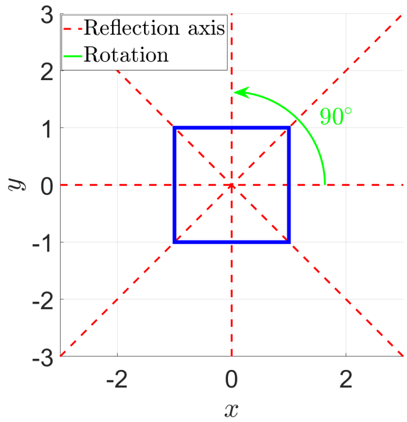
Conjecture 1 (m-symmetry)**.** *The optimal solution u^* shares the same isometries, i.e., reflections and rotations, that the regular polygon of m sides where m depends on the height M of the body.*

Parameter M prescribes the maximal height of the body. Actually, the optimal body does not seem to simply scale when the parameter M varies. The situation is much more sophisticated. It occurs that the solutions obtained in the literature naturally exhibit the same symmetry as some polygons. Moreover, the number of sides m of these polygons depends on the height M. We will refer as m for the symmetry parameter of a body u whenever it possesses such symmetry.

Fig. 2.4 is a body (left) having the same $m = 4$ reflections and $m = 4$ rotations as a square (right). The group of symmetries of a regular polygon is called a dihedral group.



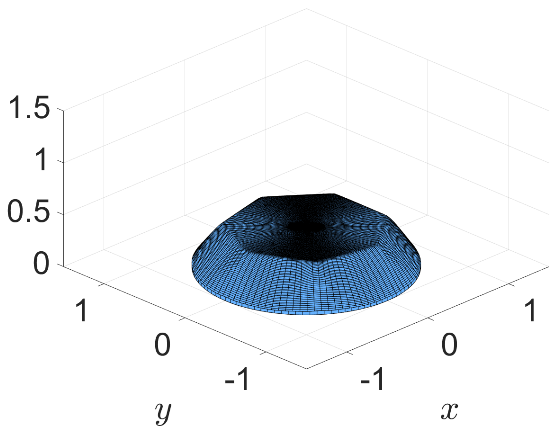
(a) Example of body with symmetry parameter $m = 4$.



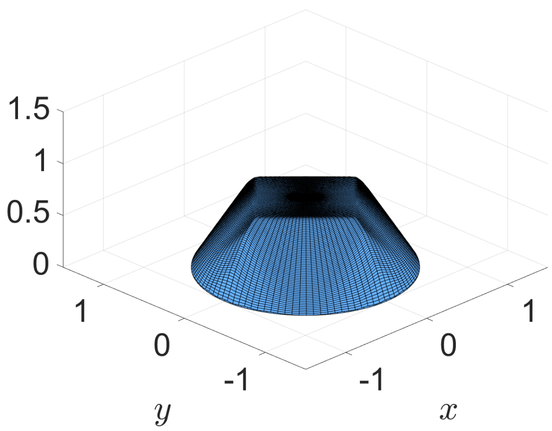
(b) Square and its isometries.

Fig. 2.4 Body and square and their isometries.

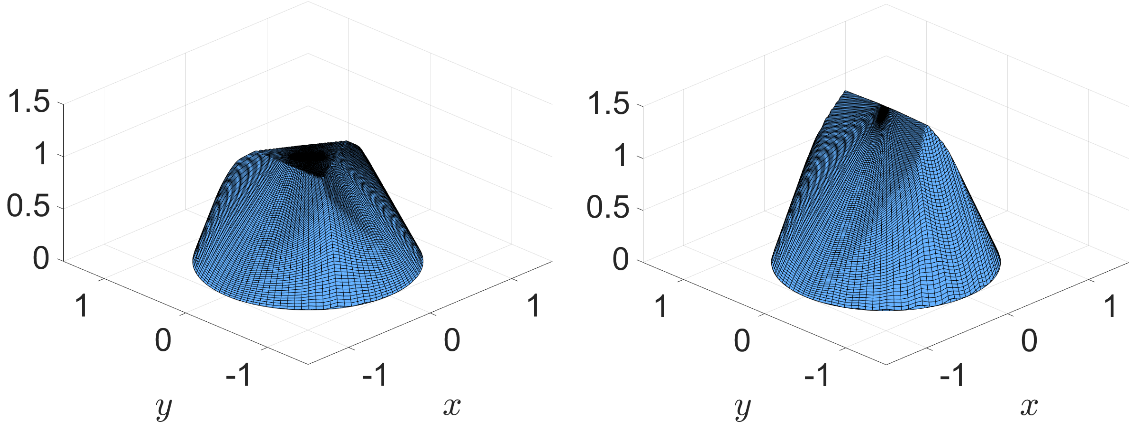
Moreover, Figs. 2.5 are sketches of the shapes of the current best bodies for different heights M. Observe that the symmetry parameter m changes. Note that $M \in \mathbb{R}$ is a continuous variable whereas $m \in \mathbb{N}$ is discrete.



(a) $M = 0.4$ and $m = 7$.



(b) $M = 0.7$ and $m = 4$.



(c) $M = 1.0$ and $m = 3$. **(d)** $M = 1.5$ and $m = 2$.

Fig. 2.5 Sketches of the shapes of the current optimal bodies for different heights M (m is the symmetry parameter).

Observing the solutions of the literature yields the following remarks.

- Decreasing the height M increases the number of sides m and in particular

$$\lim_{M \to 0} m = \infty. \tag{2.4}$$

- Increasing the height M decreases the number of sides m and in particular

$$\lim_{M \to \infty} m = 2. \tag{2.5}$$

  A polygon of $m = 2$ sides is a segment as illustrated in Fig. 2.5d.

Observation 1 (Local minimum)**.** *The Newton's problem* (1.10) *has local minimums.*

There are no theoretical results stating that there exist local minimums. However, it appears in practice that the methods of the literature and ours got stuck into local minimums. It will be useful to have a procedure to avoid them.

2.2 Developable body

In 2001 [18], T. Lachand–Robert and M. A. Peletier proposed to construct the body U as the convex hull of some set N_0 and the disk Ω.

2.2.1 Idea

Instead of considering any convex body, they only worked in a certain class of convex bodies. These bodies are the convex hull of the disk Ω and a convex bounded set $N_0 \subset \mathbb{R}^2$, included in the $z = M$ plane, to find. Formally, the new class of bodies is

$$\mathcal{S}_d = \Big\{ U_d \subseteq \mathbb{R}^3 \mid U_d = conv\big((\Omega \times \{0\}) \cup (N_0 \times \{M\})\big),\ N_0 \text{ convex and bounded} \Big\} \tag{2.6}$$

where $conv(X)$ is the convex hull of the set X.

By definition of the convex hull, such bodies $U_d \in \mathcal{S}_d$ are convex sets and therefore $\mathcal{S}_d \subseteq \mathcal{S}$. The associated function are noted U_d and the set of functions $\mathcal{C}_d \subseteq \mathcal{C}$.

The bodies $U_d \in \mathcal{S}_d$ are called developable bodies as discussed in Sect. 2.1. Indeed, these sets are globally developable. Theorem 3 motivates the choice of the class $\mathcal{S}_d$. Actually, Theorem 3 states that the optimal body U^* is developable by parts. Here, the bodies $U_d \in \mathcal{S}_d$ are globally developable which is much stronger.

2.2.2 Modelization and resolution

At this point, the body $U_d \in \mathcal{S}_d$ is entirely described by the choice of the convex set N_0. Fig. 2.6 is an example of a possible developable body where N_0 is an ellipse.

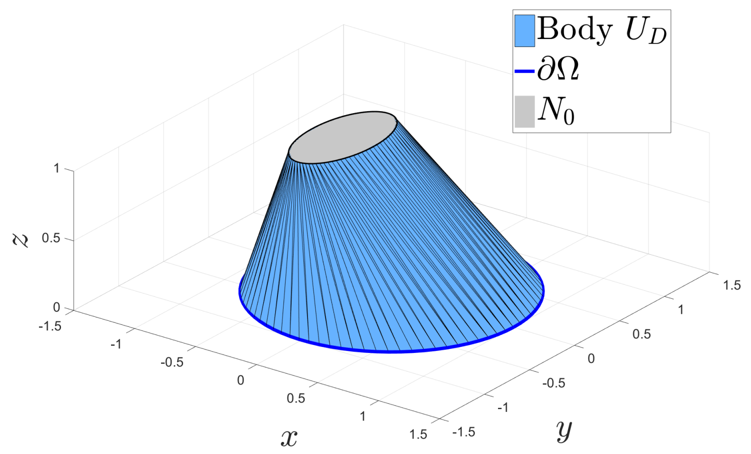


Fig. 2.6 Example of developable body where N_0 is an ellipse.

Observe that the body proposed in Fig. 2.6 is indeed composed by a developable surface. The generatrices are represented (thin black lines) on the lateral face of the body.

The remaining work of the authors consists to find and characterize the set N_0 that provides the optimal developable body $U_d^* \in \mathcal{S}_d$.

Management of the constraints The convexity constraint on U_d is naturally guaranteed by definition of a convex hull.

The constraint of maximal height M is also directly satisfied by definition of the developable body U_d. Indeed, the convex hull of the set $\Omega \times \{0\}$, lying into the plane $z = 0$, and the set $N_0 \times \{M\}$, lying into the plane $z = M$, always has a height of exactly M.

Evaluation of the resistance The authors introduce some notations to clarify the computations. In particular, we note the function $\rho(\theta) : [0, 2\pi] \to [0, 1]$ that gives the distance between the point $(0, 0)$ and the frontier of the set N_0 at the polar coordinate θ. Each convex and bounded set N_0 is described by one and only one function ρ via

$$N_0 = \Big\{(r, \theta) \mid r \leq \rho(\theta)\Big\}. \tag{2.7}$$

Therefore, U_d is entirely characterized by the function ρ. Fig. 2.7 is an illustration of the distance ρ (in blue) for some set N_0 (in red).

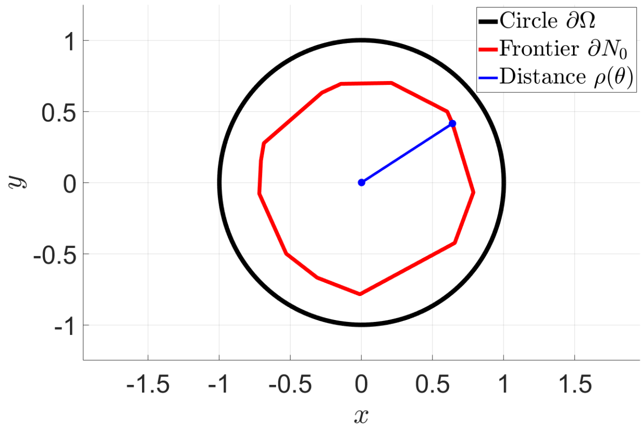


Fig. 2.7 Illustration of the distance $\rho(\theta)$.

By construction of the body U_d, the gradient of U_d is zero on N_0 and constant outside N_0.

$$||\nabla U_d(x,y)|| = \begin{cases} 0, & \text{if } (x,y) \in N_0 \\ \frac{M}{1-\rho(\theta)}, & \text{if } (x,y) \in \Omega \setminus N_0 \end{cases} \tag{2.8}$$

where $1-\rho(\theta)$ is the distance between N_0 and $\partial\Omega$.

Therefore, noting $f(t) = \frac{1}{1+t^2}$, in particular $f(0) = 1$, the resistance is

$$\mathcal{F}(U_d) = \mathcal{F}(\rho) = \int_\Omega f\big(||\nabla U_d(x,y)||\big)\mathrm{d}x\mathrm{d}y \tag{2.9}$$

$$= \int_{N_0(\rho)} f(0)\mathrm{d}x\mathrm{d}y + \int_{\Omega\setminus N_0(\rho)} f\left(\frac{M}{1-\rho(\theta)}\right)\mathrm{d}x\mathrm{d}y \tag{2.10}$$

$$= \underbrace{Area(N_0(\rho))}_{flat\ part} + \underbrace{\int_{\Omega\setminus N_0(\rho)} f\left(\frac{M}{1-\rho(\theta)}\right)\mathrm{d}x\mathrm{d}y}_{lateral\ face}. \tag{2.11}$$

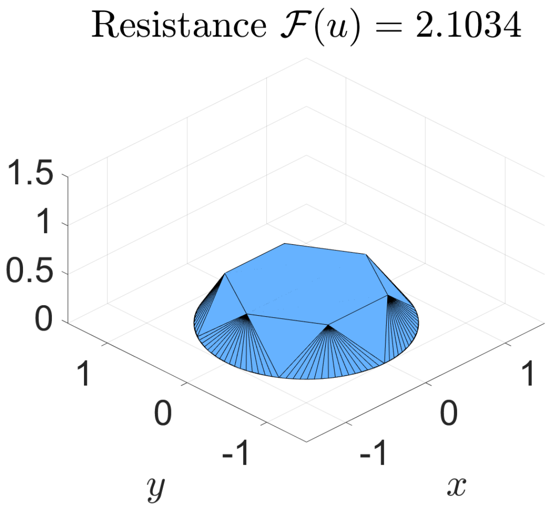
We note $N_0(\rho)$ to emphasise that the set N_0 depends on the choice of the function $\rho(\theta)$. Also, recall that $\rho(\theta) = \rho(\theta(x,y))$ is a function of x and y. The authors computed the resistance $\mathcal{F}(\rho)$ analytically.

Note that the first term of the expression (2.11) is the resistance due to the flat part of the body whereas the second term is the resistance due to the lateral face of the body. It is the already mentioned trade-off between «saving steepness» in the flat part and «using steepness» to have a low resistance in the lateral face. Indeed, increasing the values of the function $\rho(\theta)$ increases the area of the flat part but decreases the value of the integral in the lateral face in (2.11).

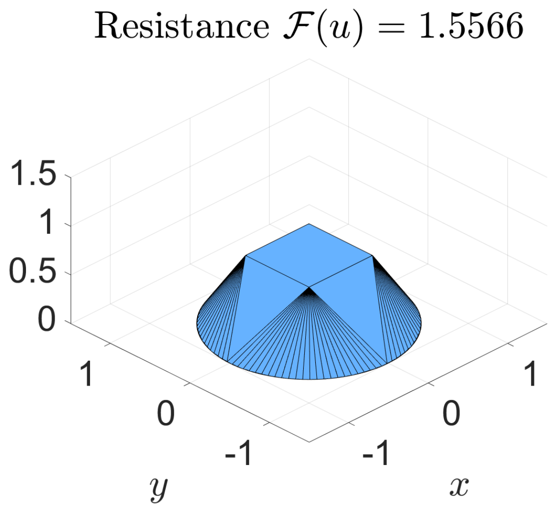
It remains to minimize the resistance $\mathcal{F}(\rho)$ with respect to the function ρ. After analytical and geometric developments, it occurs that the minimizer ρ defines a centred regular polygon as set N_0. We analyze the solutions obtained by the authors in the next section.

2.2.3 Solutions and numerical results

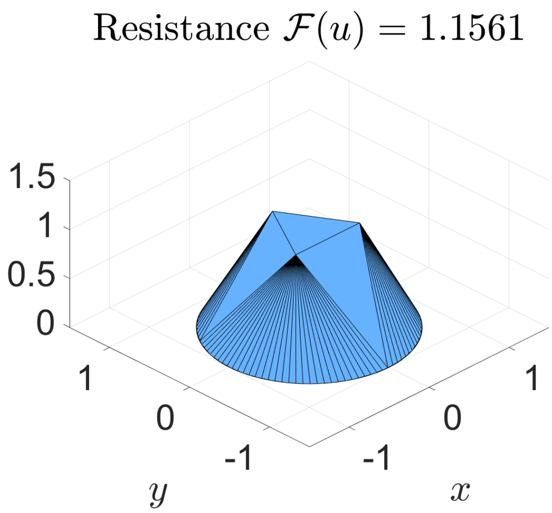
Figs. 2.8 are sketches of the optimal developable bodies and their resistances for the heights $M = 0.4$, 0.7, 1.0 and 1.5.



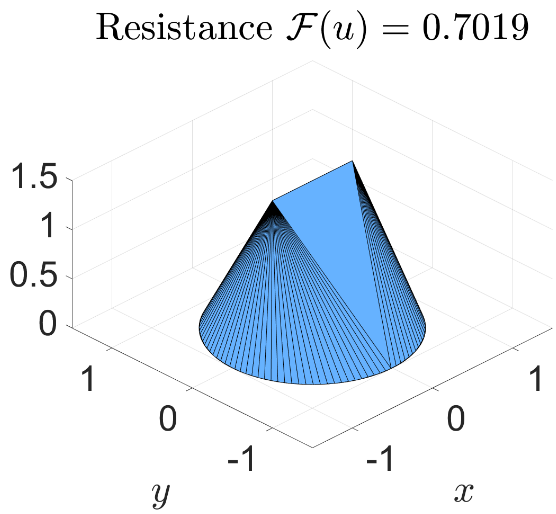
(a) $M = 0.4$.



(b) $M = 0.7$.



(c) $M = 1.0$.



(d) $M = 1.5$.

Fig. 2.8 Optimal developable bodies for different heights M.

The resistances $\mathcal{F}$ reported by the authors are in Table 2.1.

Table 2.1 Resistances of the developable bodies.

Height M	Resistance $\mathcal{F}$
1.5	0.701 9
1.0	1.156 1
0.7	1.556 6
0.4	2.103 4

It occurs that the number of sides of the optimal polygon N_0 found analytically by the authors depends on the height M. More precisely, there exists a decreasing sequence of numbers $M_2, M_3, \ldots$ such that if $M_{m+1} < M < M_m$ then the set N_0 of the optimal body U_d has m sides. In other words, if M_m is the lowest upper bound of M, then N_0 has m sides. The sequence M_m reported by the authors is in Table 2.2.

Table 2.2 Sequence M_m for the developable bodies.

M_2	∞
M_3	1.180
M_4	0.754
M_5	0.561
M_6	0.448
M_7	0.372
M_8	0.318
M_9	0.278

We must observe that the bodies obtained share the same symmetry as a polygon as suggested by Conjecture 1. Moreover, among all the bodies $U_d \in \mathcal{C}_d$, the bodies with the symmetry naturally appear to be the optimal ones.

In Fig. 2.9, we show the function $m(M)$ indicating the symmetry parameter m of the optimal developable body of height M.

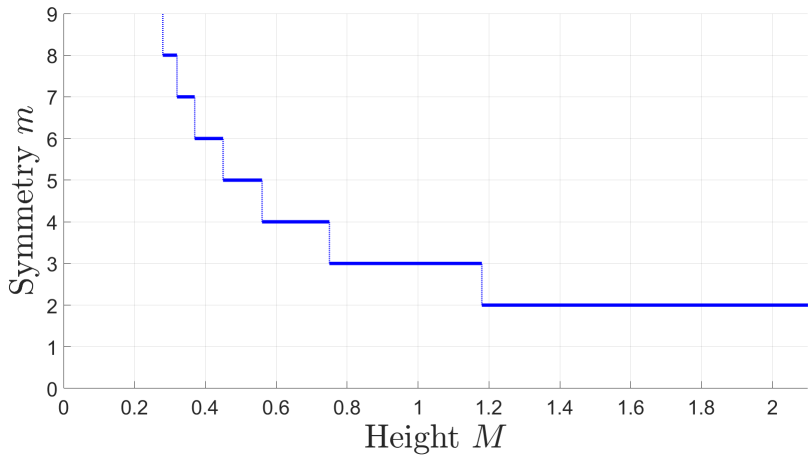


Fig. 2.9 Symmetry parameter $m \in \mathbb{N}$ as a function of the height $M \in \mathbb{R}$.

We observe that the symmetry parameter m, i.e., the number of sides of the polygon, decreases when M increases as stated after Conjecture 1. Moreover, we can start to observe the two limits

$$\lim_{M\to 0} m = \infty \tag{2.12}$$

and

$$\lim_{M\to\infty} m = 2. \tag{2.13}$$

Finally, note that Theorem 3 states that the optimal solution u^* is locally developable. The bodies U_d are globally developable. However, we will see that there exists bodies u not globally developable with better resistance than the resistances of the developable bodies in Table 2.1.

2.3 Polytope body

In 2005 [9], T. Lachand-Robert and E. Oudet approximated the body U by a polytope. They presented their technique to solve general problem involving the convex body constraint and applied it to Newton's problem.

2.3.1 Idea

They replaced the convex body U by a polytope P defined as an intersection of half-spaces. They discretized the functional $\mathcal{F}$ and optimized it with respect to the half-spaces in order to find the optimal polytope P. The optimization is made thanks to the gradient method. Note that Theorem 3 motivates this choice of polytope P since an intersection of half-spaces is everywhere non-strictly concave.

It is well known that a convex set U can be described as the intersection of half-spaces, potentially infinitely many half-spaces. The method consists in working only with a finite number of half-spaces.

2.3.2 Modelization and resolution

The convex body U is replaced by the polytope P defined by the intersection of n half-spaces. The formal definition of P is

$$P = \left\{\mathbf{x} \in \mathbb{R}^3 \mid \nu_i \cdot \mathbf{x} \leq \phi_i,\ \forall i = 1, \ldots, n\right\} = \left\{\mathbf{x} \in \mathbb{R}^3 \mid N\mathbf{x} \leq \Phi\right\} \tag{2.14}$$

where $\nu_i = \begin{pmatrix}\nu_{i,1} & \nu_{i,2} & \nu_{i,3}\end{pmatrix} \in \mathbb{R}^3$, $\phi_i \in \mathbb{R}$ characterize the half-space i, $N = \begin{pmatrix}\nu_1 \\ \vdots \\ \nu_n\end{pmatrix} \in \mathbb{R}^{n\times 3}$, $\Phi = \begin{pmatrix}\phi_1 \\ \vdots \\ \phi_n\end{pmatrix} \in \mathbb{R}^n$ and $\mathbf{x} = \begin{pmatrix}x \\ y \\ z\end{pmatrix} \in \mathbb{R}^3$.

We use the notation $P = P(N, \Phi)$ to highlight the fact that P is entirely characterized by N and Φ. Again, we equivalently use the polytope P or the function u_p. Fig. 2.10 is an example of polytope P.

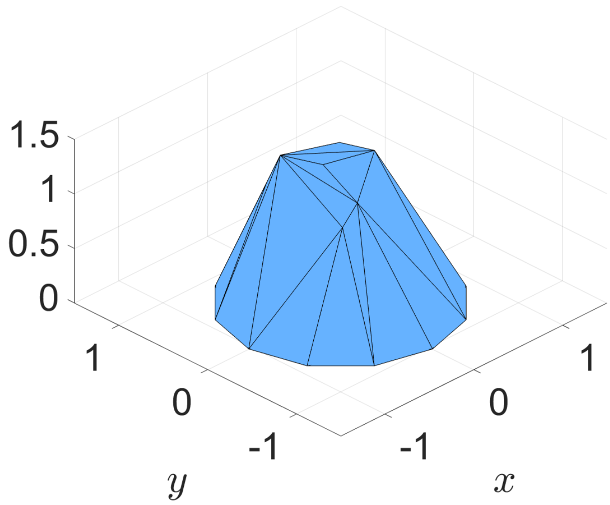


Fig. 2.10 Example of polytope P.

Management of the constraints A polytope is a convex body as it is an intersection of half-spaces which are convex sets.

The constraints $0 \leq u_p \leq M$ are taken into account just by adding the half-spaces $z \geq 0$ and $z \leq M$ in N and Φ. In practice, it yields

- $u_p \geq 0$: half-space $z \geq 0$, i.e.,

$$0x + 0y - 1z \leq 0 \Leftrightarrow \nu_{1,1} = \nu_{1,2} = 0,\ \nu_{1,3} = -1,\ \phi_1 = 0. \tag{2.15}$$

- $u_p \leq M$: half-space $z \leq M$, i.e.,

$$0x + 0y + 1z \leq M \Leftrightarrow \nu_{2,1} = \nu_{2,2} = 0,\ \nu_{2,3} = 1,\ \phi_2 = M. \tag{2.16}$$

Moreover, they imposed the polytope to fit the disk Ω. Of course, we cannot construct a polytope with a circular basis with a finite number of half-spaces, hence, they discretized the disk into a polygon, Ω_ℓ, of ℓ sides. They added ℓ constraints to keep the polytope P inside Ω. The constraints are the following half-spaces

$$\begin{gathered} \{\mathbf{x} \in \mathbb{R}^3 \mid \nu_{3,1}x + \nu_{3,2}y + 0z \leq \phi_3\} \\ \vdots \\ \{\mathbf{x} \in \mathbb{R}^3 \mid \nu_{2+\ell,1}x + \nu_{2+\ell,2}y + 0z \leq \phi_{2+\ell}\} \end{gathered} \tag{2.17}$$

where $\nu_{i,1}, \nu_{i,2}$ and ϕ_i are determined geometrically for $i = 3, \ldots, 2+\ell$.

Fig. 2.11 shows these ℓ half-spaces and constraints.

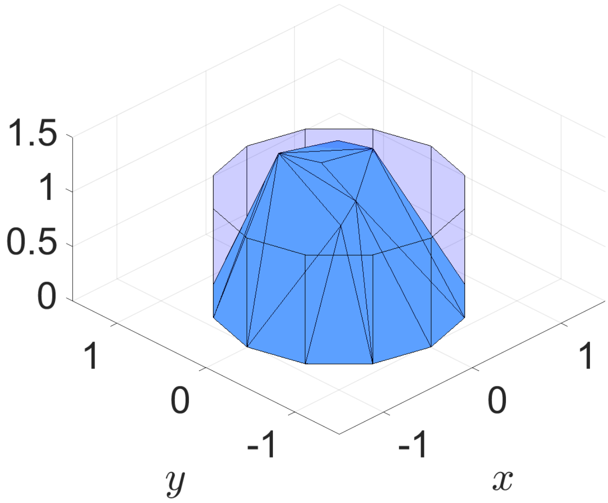


Fig. 2.11 Polytope P with the ℓ constraints.

They chose the ℓ half-spaces such that the disk Ω is an incircle of the polygon Ω_ℓ. The domain used to compute the resistance is Ω_ℓ instead of Ω. Since the resistance $\mathcal{F}$ integrates the function $f(t) = \frac{1}{1+t^2} > 0$, the value of the resistance $\mathcal{F}$ obtained is inherently overestimated. In practice, they used $\ell = 300$.

Evaluation of the resistance In this approach, the gradient of u_p is constant by parts, i.e., ∇u_p is constant on each face of the polytope P. Therefore, the gradient is easily computed on each face. However, computing the faces and knowing on each point (x, y) which half-space is dominant is a tedious task.

A half-space is dominant on a point (x, y) if it actually defines a frontier of the polytope P on this point. Formally, the half-space i is dominant on the point (x, y) if the point $\mathbf{x} = (x, y, z)$, where $z = \frac{\phi_i - \nu_{i,1}x - \nu_{i,2}y}{\nu_{i,3}}$, is on all the other half-spaces, i.e.,

$$\nu_j \cdot \mathbf{x} \leq \phi_j \qquad \forall j = 1, \ldots, n. \tag{2.18}$$

As a consequence, when i is dominant on (x, y)

$$u_p(x, y) = z = \frac{\phi_i - \nu_{i,1}x - \nu_{i,2}y}{\nu_{i,3}}. \tag{2.19}$$

We note $F_i(\nu_i, \phi_i) = F_i$ the face due to the half-space i, in other terms, it is the set of points $(x, y) \in \Omega$ where the half-space i is dominant, i.e.,

$$F_i = \left\{ (x, y) \in \Omega \;\middle|\; \nu_i \cdot \begin{pmatrix} x \\ y \\ u_p(x, y) \end{pmatrix} = \phi_i \right\}. \tag{2.20}$$

Therefore, the gradient of u_p, and its norm, on the face F_i where the half-space i dominates are

$$\nabla u_p(x, y) = \frac{1}{\nu_{i,3}} \begin{pmatrix} \nu_{i,1} \\ \nu_{i,2} \end{pmatrix} \qquad \forall (x, y) \in F_i \tag{2.21}$$

$$||\nabla u_p(x, y)|| = \sqrt{\frac{\nu_{i,1}^2 + \nu_{i,2}^2}{\nu_{i,3}^2}} \qquad \forall (x, y) \in F_i. \tag{2.22}$$

The resistance $\mathcal{F}$ of the polytope is

$$\mathcal{F}(u_p) = \int_\Omega f\big(||\nabla u_p(x,y)||\big)\mathrm{d}x\mathrm{d}y \tag{2.23}$$

$$= \sum_{i=1}^n \int_{F_i} f\big(||\nabla u_p(x,y)||\big)\mathrm{d}x\mathrm{d}y \tag{2.24}$$

$$= \sum_{i=1}^n f\left(\sqrt{\frac{\nu_{i,1}^2+\nu_{i,2}^2}{\nu_{i,3}^2}}\right)\int_{F_i}\mathrm{d}x\mathrm{d}y \tag{2.25}$$

$$= \sum_{i=1}^n \left(1+\frac{\nu_{i,1}^2+\nu_{i,2}^2}{\nu_{i,3}^2}\right)^{-1} Area(F_i). \tag{2.26}$$

Note that if the half-space i is not dominant anywhere, then, $Area(F_i) = 0$.

They can now formulate the problem as a more classical optimization problem with a finite number of variables

$$\min_{\Phi\in\mathbb{R}^{n\times 3}} \quad \sum_{i=1}^n \left(1+\frac{\nu_{i,1}^2+\nu_{i,2}^2}{\nu_{i,3}^2}\right)^{-1} Area\big(F_i(\nu_i,\phi_i)\big) \tag{2.27}$$

where ν_i and ϕ_i are fixed by the half-spaces (2.15), (2.16) and (2.17) for $i = 1,\dots,2+\ell$ and variables for $i = 2+\ell+1,\dots,n$.

Note that the minimization is made with respect to Φ only and not N. Indeed, optimizing N does not improve the method according to some numerical tests of the authors.

The problem (2.27) is unconstrained and can be solved by the gradient method. Performing the optimization via the gradient method requires to compute the derivatives of the objective function that is the resistance $\mathcal{F}$.

Gradient of the resistance The task is not straightforward. The gradient must take into account the topological modifications, i.e., a non-dominant face that becomes dominant and conversely.

After a geometrical analysis, they obtained an expression of the derivatives of $\mathcal{F}$ with respect to the variables ϕ_i.

$$\frac{\partial\mathcal{F}}{\partial\phi_i}(N,\Phi) = \sum_{\substack{j=1\\ j\neq i}}^n Length(F_i\cap F_j)\frac{\nu_{j,3} f\left(\sqrt{\frac{\nu_{j,1}^2+\nu_{j,2}^2}{\nu_{j,3}^2}}\right) - \cos(\theta_{ij})\nu_{i,3} f\left(\sqrt{\frac{\nu_{i,1}^2+\nu_{i,2}^2}{\nu_{i,3}^2}}\right)}{\sin(\theta_{ij})} \tag{2.28}$$

where θ_{ij} is such that $\cos(\theta_{ij}) = |\nu_i\cdot\nu_j|$ and $\sin(\theta_{ij})(\nu_i\cdot\nu_j) \geq 0$.

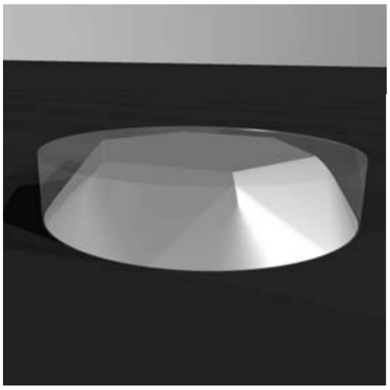
Observe that the expression (2.28) depends only on N, Φ, F_i the faces of the polytope and the geometrical parameters θ_{ij} that also depends on N and Φ. Therefore, knowing the polytope $P(N,\Phi)$, the derivatives can be entirely computed. The authors do not provide details on how to compute the geometry, faces and vertex, of a given polytope $P(N, \Phi)$.

Actually, they used a slightly modified gradient method. At each step of the gradient method, they can add or remove half-spaces and modify the value of N since the optimization is only made with respect to Φ. Moreover, they used a genetic algorithm to find the global minimum and avoid the local ones of Observation 1.

2.3.3 Solutions and numerical results

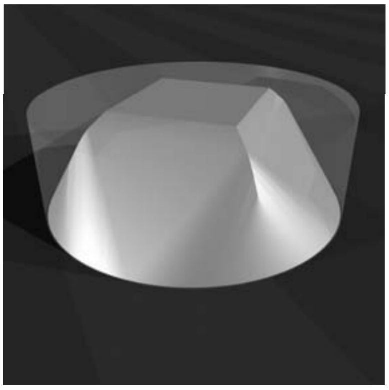
Figs. 2.12 are the optimal polytopes obtained in [9] for the heights $M = 0.4$, 0.7, 1.0 and 1.5.

Resistance $\mathcal{F} = 2.1006$



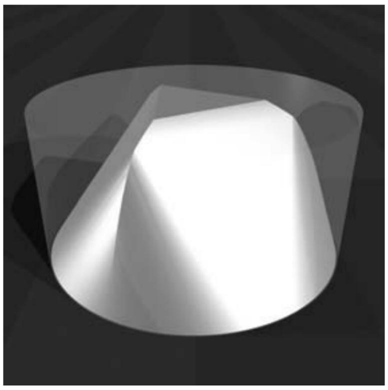
(a) $M = 0.4$.

Resistance $\mathcal{F} = 1.5457$



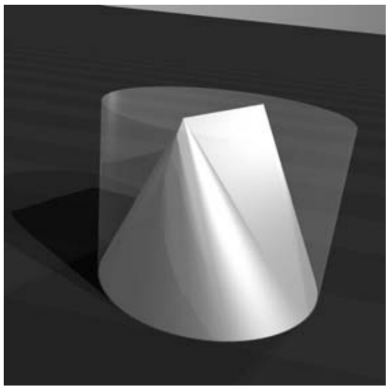
(b) $M = 0.7$.

Resistance $\mathcal{F} = 1.1379$



(c) $M = 1.0$.

Resistance $\mathcal{F} = 0.7012$



(d) $M = 1.5$.

Fig. 2.12 Optimal polytopes for different heights M (from [9]).

The resistances $\mathcal{F}$ reported by the authors are in Table 2.3.

Table 2.3 Resistances of the polytopes.

Height M	Resistance $\mathcal{F}$
1.5	0.701 2
1.0	1.137 9
0.7	1.545 7
0.4	2.100 6

The results obtained by the polytopes are better than the results of the developable bodies for all heights. It can be shown that the solutions provided by the polytope method are not globally developable [9].

The optimal bodies in Fig. 2.12 also seem to exhibit the property of symmetry m. However, it would require a more rigorous analysis to prove that these bodies are indeed symmetric, even if on the figure, they strongly appear to be symmetric. Again, it was not imposed by the choice of modelization (2.14) that the body P must have this symmetry but it naturally becomes symmetric during the optimization.

The sequence M_m of the polytope method is not provided by the authors. We can only deduce the relations in Table 2.4 from the resistances and figures reported. Recall that the sequence M_m indicates the symmetry parameter m of a body smaller than M_m.

Table 2.4 Sequence M_m for the polytopes.

	M_2	$= \infty$
$1.0 <$	M_3	< 1.5
$0.7 <$	M_4	< 1.0
$0.4 <$	M_5	< 0.7
$0.4 <$	M_6	< 0.7
	M_7	< 0.4

The intervals of the different values of M_m are consistent with the ones of the developable bodies in Table 2.2.

2.4 Minimum of linear functions

In 2014 [15], G. Wachsmuth proposed two methods to solve numerically the Newton's problem. The first one is to express the function u as a minimum of linear functions. Again, note that Theorem 3 motivates this choice of function u since a minimum of linear functions is everywhere non-strictly concave.

2.4.1 Idea

He expressed the function $u(x, y)$ as a minimum of linear functions, added the constraint of height M and then formulated the Newton's problem (1.10) as a more classical optimization problem with a finite number of variables. Conceptually, it yields equivalent bodies that the polytope approach but with another description. Yet, it occurs that the minimum of linear functions is more convenient than the polytope approach.

2.4.2 Modelization and resolution

The function $u(x, y)$ is chosen as a minimum of n linear functions

$$u_\ell(x, y) = \min_{i=1,\ldots,n} a_i x + b_i y + c_i \tag{2.29}$$

where a_i, b_i, c_i are real parameters for $i = 1, \ldots, n$. The choice of the these parameters entirely describes the body u_ℓ.

Fig. 2.13 is an example of function u_ℓ.

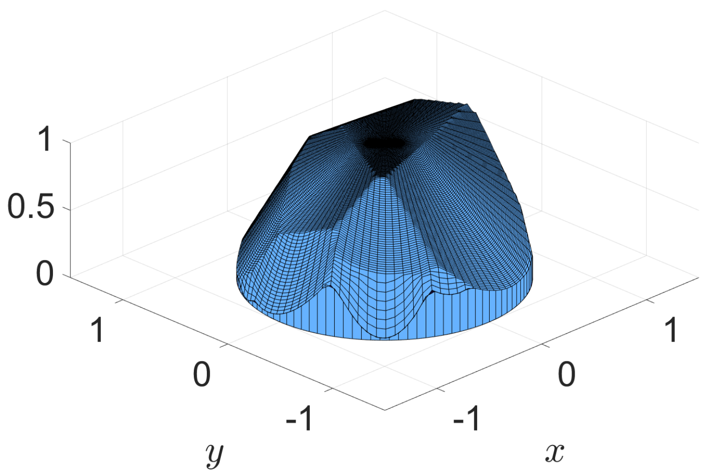


Fig. 2.13 Example of minimum of linear functions u_ℓ.

Management of the constraints The constraint of convexity of the body is directly satisfied since a minimum over linear functions is concave. We can observe it on a 2d example in Fig. 2.14.

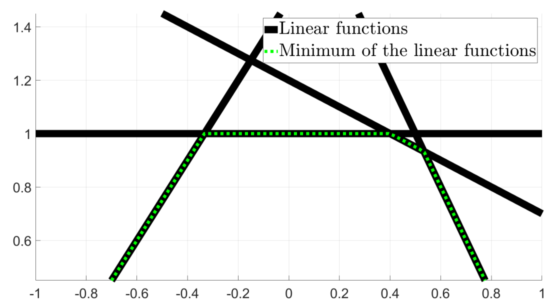


Fig. 2.14 A Minimum of linear functions is concave.

In order to guarantee $u_\ell \leq M$, the author adds the linear function $0x + 0y + M$ by setting $a_1 = b_1 = 0$ and $c_1 = M$. Hence, the minimum u_ℓ could never go upper than M.

In order to guarantee $u_\ell \geq 0$, the author looked at the lowest point of each linear function $a_i x + b_i y + c_i$ on the disk which is the point

$$(x, y) = -\frac{(a_i, b_i)}{\sqrt{a_i^2 + b_i^2}} \tag{2.30}$$

that takes the value

$$a_i x + b_i y + c_i = -\sqrt{a_i^2 + b_i^2} + c_i. \tag{2.31}$$

Hence, the author just imposed

$$c_i \geq \sqrt{a_i^2 + b_i^2} \qquad \forall i = 1, \ldots, n. \tag{2.32}$$

Evaluation of the resistance Similarly to the polytope method, the gradient of a linear function is easily computed on a point (x, y). However, we must know which linear function is dominant on (x, y). Intuitively, the interpretation of a dominant linear function is identical to the one of a half-space.

Formally in this context, a linear function i is dominant on (x, y) if it dominates all the other linear functions j, i.e.,

$$a_i x + b_i y + c_i \leq a_j x + b_j y + c_j \qquad \forall j = 1, \ldots, n. \tag{2.33}$$

The set of points where the linear function i is dominant is

$$F_i = \big\{(x, y) \in \Omega \mid a_i x + b_i y + c_i \leq a_j x + b_j y + c_j, \ \ \forall j = 1, \ldots, n\big\}. \tag{2.34}$$

The gradient of u_ℓ, and its norm, on the points where the linear function i dominates are

$$\nabla u_\ell(x, y) = \begin{pmatrix} a_i \\ b_i \end{pmatrix} \qquad \forall (x, y) \in F_i \tag{2.35}$$

$$||\nabla u_\ell(x, y)|| = \sqrt{a_i^2 + b_i^2} \qquad \forall (x, y) \in F_i. \tag{2.36}$$

The resistance $\mathcal{F}$ is computed exactly as for the polytope method, i.e.,

$$\begin{aligned}
\mathcal{F}(u_\ell) &= \int_\Omega f\big(||\nabla u_\ell(x, y)||\big) \mathrm{d}x\mathrm{d}y && (2.37)\\
&= \sum_{i=1}^{n} \int_{F_i} f\big(||\nabla u_\ell(x, y)||\big) \mathrm{d}x\mathrm{d}y && (2.38)\\
&= \sum_{i=1}^{n} f\left(\sqrt{a_i^2 + b_i^2}\right) \int_{F_i} \mathrm{d}x\mathrm{d}y && (2.39)\\
&= \sum_{i=1}^{n} \big(1 + a_i^2 + b_i^2\big)^{-1} Area(F_i). && (2.40)
\end{aligned}$$

Finally, the author obtained the following constrained optimization problem

$$
\begin{aligned}
\min_{a,b,c\in\mathbb{R}^n} \quad & \sum_{i=1}^{n} \left(1 + a_i^2 + b_i^2\right)^{-1} Area\big(F_i(a_i, b_i, c_i)\big) \\
\text{s.t.} \quad & c_i \geq \sqrt{a_i^2 + b_i^2} \qquad \forall i = 1, \ldots, n.
\end{aligned} \tag{2.41}
$$

where $a_1 = b_1 = 0$ and $c_1 = M$ are fixed.

He solved the problem using `fmincon` [19] of Matlab starting from a random admissible solution a_0, b_0, c_0. It is possible to use `fmincon` with or without the gradient of the objective function. The author provided the gradient of the resistance. He did not report the algorithm used, the available ones in `fmincon` are *interior-point*, *trust-region-reflective*, *sqp*, *sqp-legacy* and *active-set*.

Gradient of the resistance Again similarly to the polytope method, the difficulty is on the topological modifications. After geometric and analytic developments, the gradient of the objective function is

$$
\frac{\partial \mathcal{F}(u)}{\partial \begin{pmatrix} a_i \\ b_i \\ c_i \end{pmatrix}} = -2Area(F_i)\big(1+a_i^2+b_i^2\big)^{-2} \begin{pmatrix} a_i \\ b_i \\ 0 \end{pmatrix} + \sum_{j=1}^{n} \frac{\partial Area(F_j)}{\partial \begin{pmatrix} a_i \\ b_i \\ c_i \end{pmatrix}} \big(1+a_j^2+b_j^2\big)^{-1}. \tag{2.42}
$$

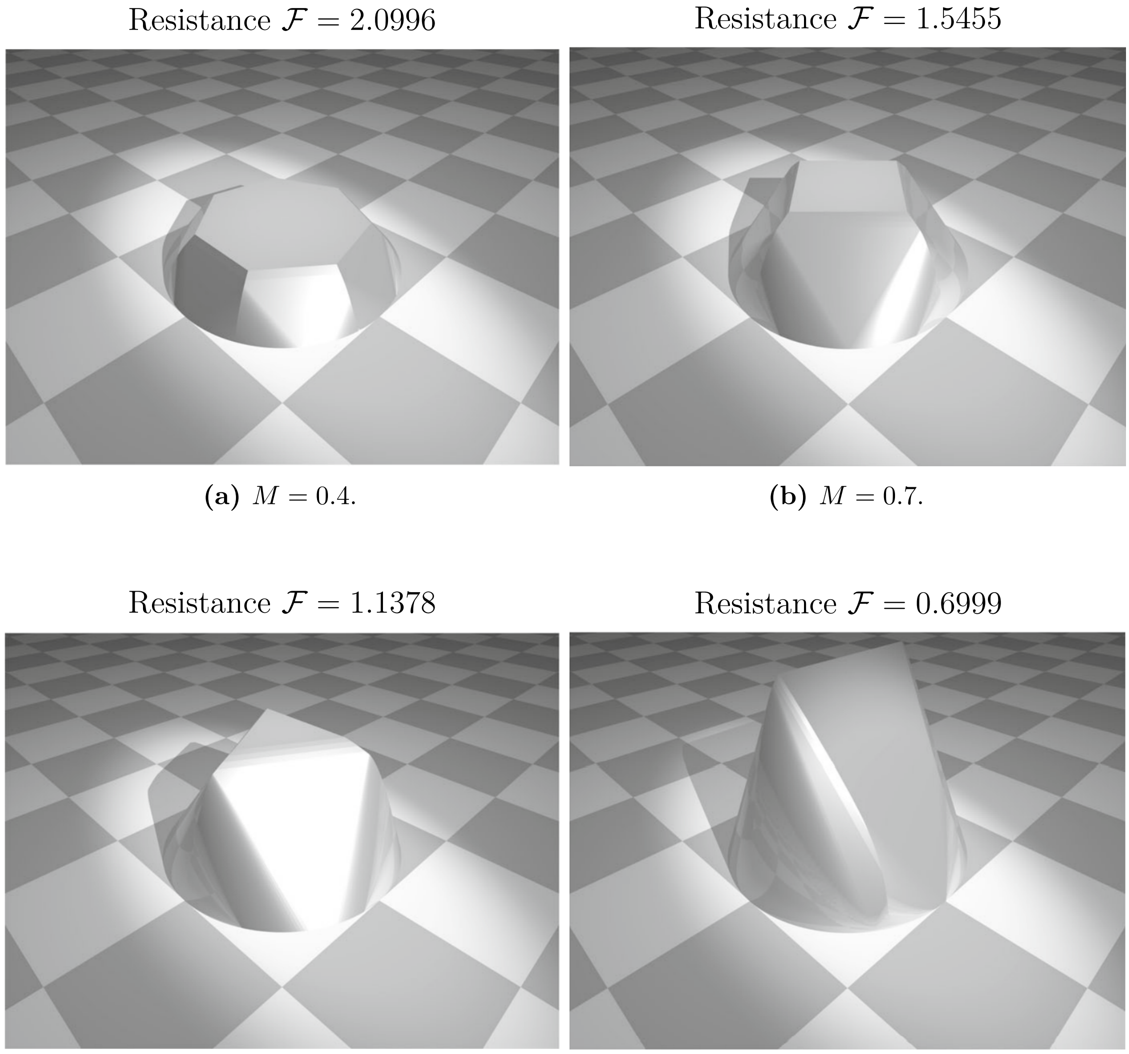
It is possible to reduce the problem of computing the regions of dominance of the linear functions to the problem of computing the power diagram of some weighted points (x_i, y_i, w_i). A power diagram problem is a generalization of the Voronoi diagram problem.

In addition to the optimization performed by `fmincon`, the author can also manually add or remove linear functions when needed.

Moreover, observing the symmetry phenomenon, the author imposed the symmetry to its function u_ℓ. With a body u_ℓ symmetric, the disk Ω can be divided in $2m$ sectors where the function u_ℓ is exactly the same on each sector. Therefore, the resistance can be computed only on one sector of Ω and then multiplied by $2m$. It reduces the execution time of the optimization.

2.4.3 Solutions and numerical results

Figs. 2.15 are the optimal minimum of linear functions obtained in [15] for the heights $M = 0.4$, 0.7, 1.0 and 1.5.



(a) $M = 0.4$. **(b)** $M = 0.7$. **(c)** $M = 1.0$. **(d)** $M = 1.5$.

Fig. 2.15 Optimal minimum of linear functions for different heights M (from [15]).

The resistances $\mathcal{F}$ reported by the author are in Table 2.5.

Table 2.5 Resistances of the minimum of linear functions.

Height M	Resistance $\mathcal{F}$
1.5	0.699 9
1.0	1.137 8
0.7	1.545 5
0.4	2.099 6

The numerical results of this approach, reported in Table 2.5, are better than the results of the polytope approach available in Table 2.3.

Before imposing the symmetry on the body u_ℓ to save execution time, the solutions obtained already seemed to be symmetric as predicted by Conjecture 1. Then, the author imposed the symmetry to reach even better results.

The sequence M_m prescribing the symmetry parameter m is not reported by the author. From the resistances and figures reported, we can only deduce the exact same bounds as the polytope approach reported in Table 2.4.

2.5 Advanced convex hull

The second approach of G. Wachsmuth [15] is to express the body U as a specific convex hull. Given his observations on the solutions obtained by his first approach, he proposed a suitable and convenient way to express the body U.

2.5.1 Idea

G. Wachsmuth observed that all the optimal solutions proposed by the other approaches, including his first one, benefits from the same symmetry as a polygon. Hence, he imposed this symmetry on his new class of body. The bodies are entirely determined by a univariate function g and a convex hull. The resolution phase consists in finding the optimal function g.
Moreover, the definition of its bodies explicitly depends on the symmetry parameter m.

2.5.2 Modelization and resolution

First, the author defined m rays on the disk Ω, one each $\phi = \frac{2\pi}{2m}$. For a fixed $m \geq 2$, the rays are the following sets of points (x, y)

$$\text{ray}_k = \left\{ \begin{pmatrix} r\cos(2k\phi) \\ r\sin(2k\phi) \end{pmatrix} \in \mathbb{R}^2 \;\middle|\; r \in [0,1] \right\} \qquad \forall k = 0, \dots, m-1. \tag{2.43}$$

Fig. 2.16 shows the m rays in the case $m = 4$ and $\phi = \frac{2\pi}{2m} = \frac{\pi}{4}$. Each ray is a segment from $r = 0$ to $r = 1$.

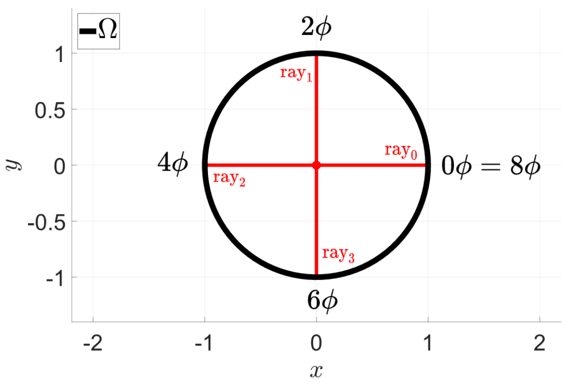


Fig. 2.16 Rays for $m = 4$.

We note the body $U_{g,m}$ and $u_{g,m}$. On the m rays ray_k, the shape of the body $U_{g,m}$ is explicitly defined by the continuous concave decreasing function

$$g : [0,1] \to [0, M] : r \to g(r) \tag{2.44}$$

and implicitly defined by the convex hull between the rays.

More precisely, on the rays, the body $U_{g,m}$ takes the value of $g(r)$, i.e.,

$$(x,y) \in \text{ray}_k \Rightarrow \begin{cases} \big(x,y,z\big) \in U_g & \forall z \in [0, g(r)] \\ u_g(x,y) = g(r) \end{cases} \tag{2.45}$$

where $r = \sqrt{x^2+y^2}$.

Therefore, noting $g_k \subset \mathbb{R}^3$ the graph of the function $g(r)$ on the line ray_k, i.e.,

$$g_k = \Big\{ \big(x,y,g(r)\big) \in \mathbb{R}^3 \;\big|\; (x,y) \in \text{ray}_k,\; r \in [0,1] \Big\}, \tag{2.46}$$

the whole body $U_{g,m}$ is defined by the following convex hull

$$U_{g,m} = conv\Bigg(\Big(\Omega \times \{0\} \Big) \cup \Big(\bigcup_{k=0}^{m-1} g_k \Big) \Bigg). \tag{2.47}$$

Fig. 2.17 is an example of body $U_{g,m}$ for some arbitrary function g and $m=3$.

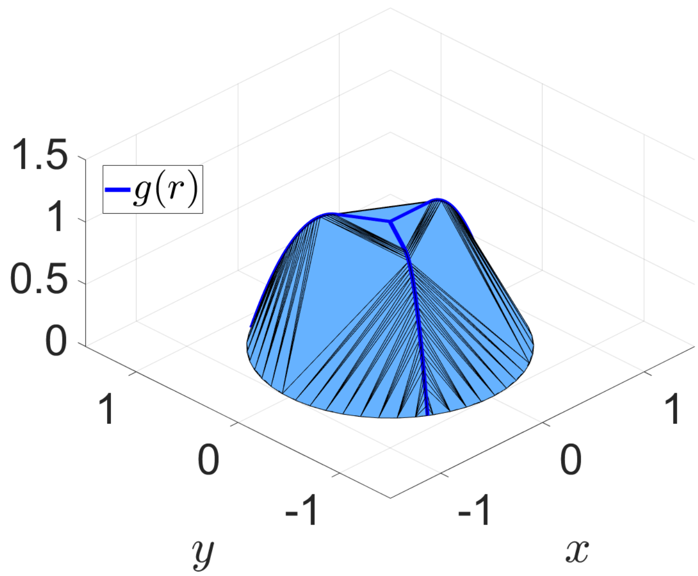


Fig. 2.17 Example of advanced convex hull $U_{g,m}$.

Note that by definition (2.47) of the body $U_{g,m}$, such bodies always have the symmetry m as desired.

Management of the constraints The convexity of $U_{g,m}$ is guaranteed by its definition (2.47) since a convex hull is a convex set. And, in order to guarantee $0 \le u_{g,m} \le M$, the function g must satisfy $g(0) = M$ and $g(1) = 0$.

Due to the symmetry imposed by the definition (2.47) of $U_{g,m}$, the behavior of the function $u_{g,m}$ is the same on each of the $2m$ sectors

$$\theta \in [j\phi,\; (j+1)\phi] \qquad j = 0, \dots, 2m-1 \tag{2.48}$$

of the disk Ω. Recall that $\phi = \frac{2\pi}{2m}$.

Moreover, the author divided each sector in five regions, as shown in Fig. 2.18.

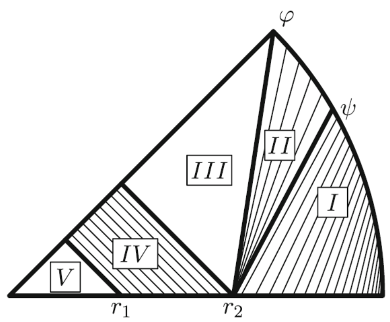


Fig. 2.18 Sector of Ω divided in 5 regions (from [15]).

He introduced additional conditions on g in order to facilitate the computation of $\mathcal{F}(u_{g,m})$ on the sector. In particular, there exist parameters r_1 and r_2 with $0 < r_1 < r_2 < 1$ characterizing the 5 regions and they must respect the following additional conditions

- $g(r) = M \quad \forall r \in [0, r_1]$
- g is twice continuously differentiable on the intervals $[r_1, r_2]$ and $[r_2, 1]$.

Evaluation of the resistance After tedious developments, the author expressed the resistance on each of the 5 regions in Fig. 2.18.

Region I

$$J_1 = \int_{r_2}^{1} \frac{\frac{\mathrm{d}\sin(\alpha(r))}{\mathrm{d}r}\big(\cos(\alpha(r)) - r\big) - \sin(\alpha(r))\Big(\frac{\mathrm{d}\cos(\alpha(r))}{\mathrm{d}r} + 1\Big)}{1 + \big(g(r) - rg'(r)\big)^2} \mathrm{d}r \tag{2.49}$$

where $g'(r)$ is the derivative of g with respect to r and $\alpha(r)$ is such that $\cos(\alpha(r)) = \frac{-g'(r)}{g(r)-rg'(r)}$.

Region II

$$J_2 = \frac{1}{2} \int_{\psi}^{\phi} \frac{1 - r_2\cos(\alpha)}{1 + \Big(\frac{g(r_2)}{1-r_2\cos(\alpha)}\Big)^2} \mathrm{d}\alpha \tag{2.50}$$

Region III

$$J_3 = \frac{1}{2} \frac{r_2\sin(\phi)\big(1 - r_2\cos(\phi)\big)}{1 + \Big(\frac{g(r_2)}{1-r_2\cos(\phi)}\Big)^2} \tag{2.51}$$

Region IV

$$J_4 = \frac{1}{2}\sin(2\phi)\cos^2(\phi)\int_{r_1}^{r_2} \frac{r}{\cos^2(\phi) + g'(r)^2}\mathrm{d}r \tag{2.52}$$

Region V

$$J_5 = \frac{1}{2}r_1^2\cos(\phi)\sin(\phi) \tag{2.53}$$

Finally, the total resistance $\mathcal{F}$ is the sum of each five regions on each of the $2m$ sectors, i.e.,

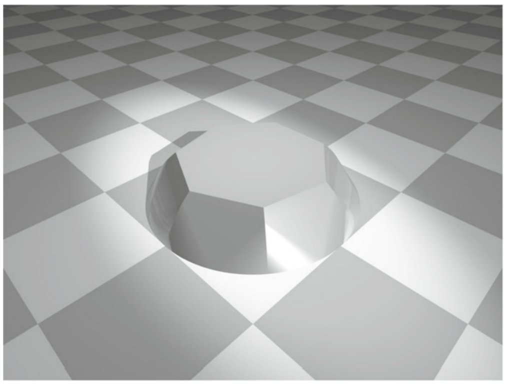
$$\mathcal{F}(u_{g,m}) = 2m(J_1 + J_2 + J_3 + J_4 + J_5). \tag{2.54}$$

The author proposed a strategy to find the optimal function g. Mainly, it consists in finding two parameters, $g'(1)$ and r_2, such that the obtained function g is optimal. Computing J_1 and J_4 can be done by solving numerically the associated Euler-Lagrange equations which are second order differential equations.

2.5.3 Solutions and numerical results

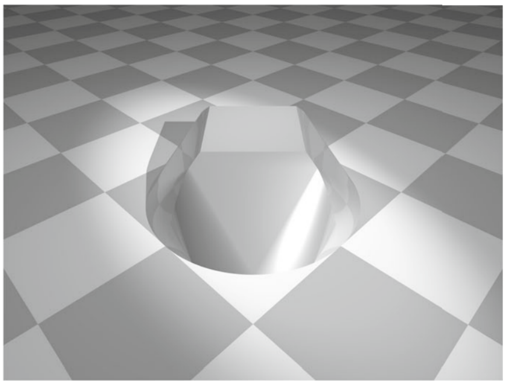
Figs. 2.19 are the optimal advanced convex hulls obtained in [15] for the heights $M = 0.4$, 0.7 and 1.0. The body $u_{g,m}$ for $M = 1.5$ was not illustrated.

Resistance $\mathcal{F} = 2.0996$



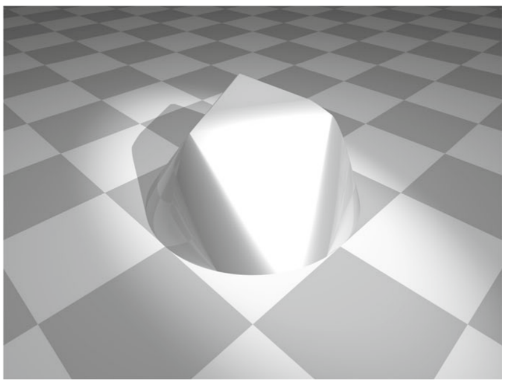
(a) $M = 0.4$.

Resistance $\mathcal{F} = 1.5455$



(b) $M = 0.7$.

Resistance $\mathcal{F} = 1.1378$



(c) $M = 1.0$.

Fig. 2.19 Optimal advanced convex hull for different heights M (from [15]).

The numerical results reported by the author are in Table 2.6.

Table 2.6 Resistances of the advanced convex hull.

Height M	Resistance $\mathcal{F}$
1.5	0.699 9
1.0	1.137 8
0.7	1.545 5
0.4	2.099 6

The author provides bounds on the values of the elements of the sequence M_m for this approach.

Table 2.7 Sequence M_m for the advanced convex hull.

	M_2	$= \infty$
$1.4 <$	M_3	< 1.5
$0.9 <$	M_4	< 1.0
$0.6 <$	M_5	< 0.7
$0.5 <$	M_6	< 0.6
$0.4 <$	M_7	< 0.5

This approach provides the best known results for $M < 1.4$. For $M > 1.5$, it is the minimum of linear functions which is the best. It is not clear for which M the transition occurs. The transition coincides with the change of symmetry parameter m from $m = 3$ to $m = 2$.

2.6 Summary and comments

We start this section by summarizing the different approaches reviewed. The following list contains the different approaches and the way they managed the convexity constraint.

- Radially symmetric body (1686 I. Newton): the radial profile g is concave.
- Developable body (2000 T. Lachand-Robert and M. A. Peletier): the body is a convex hull.
- Polytope (2005 T. Lachand-Robert and E. Oudet): the body is the intersection of half-spaces.
- Minimum of linear functions (2014 G. Wachsmuth): a minimum of linear functions is concave.
- Advanced convex hull (2014 G. Wachsmuth): the body is a convex hull.

Table 2.8 contains the resistances obtained by the different approaches. The current best resistance for each height M are in bold.

Table 2.8 Summary of the resistances $\mathcal{F}$ obtained by the approaches of the literature.

Height	$M = 0.4$	$M = 0.7$	$M = 1.0$	$M = 1.5$
Radially symm.	2.107 4	1.568 5	1.177 5	0.752 6
Developable	2.103 4	1.556 6	1.156 1	0.701 9
Polytope	2.100 6	1.545 7	1.137 9	0.701 2
Min. of linear fun.	2.099 645	1.545 487	1.137 781	**0.699 923**
Adv. convex hull	**2.099 606**	**1.545 461**	**1.137 751**	0.699 949

Now, we propose a list of comments and observations on the different approaches and their results.

Non-optimality of the radially symmetric Newton's solution The Newton's solution is radially symmetric whereas the developable bodies are m-symmetric. And the developable bodies already have a better resistance that the Newton's solution.

It appears that the Newton's solution does not exploit the correct type of symmetry whereas the developable body does, i.e., the symmetry suggested by Conjecture 1.

Non-optimality of developable bodies and non-global developability It is shown by Theorem 3 that the optimal body u^* of the Newton's problem is locally developable. However, assuming the global developability is too restrictive and excludes the optimal solution u^*. Therefore, the developable bodies are not optimal. Indeed, there exist other methods that obtained better results. For instance, polytopes do not make this assumption and get better results with non-globally developable bodies.

Non-optimality of polytopes The polytope approach slightly suffers of impracticabilities. For instance, the optimization is only made with respect to Φ and not N. Even if the authors ensure that they have strategies to solve this issue, I suspect that the difference of results between this approach and the minimum of linear functions comes from here. Indeed, the latter provides very slightly better results.

Difficulty of the polytope and minimum of linear functions The polytope and minimum of linear functions approaches are very similar, we can notice it by comparing the optimization problems (2.27) and (2.41). The complexity of these approaches is due to the geometrical computations. Indeed the gradients ∇u are easily computed but determining the edges, vertexes and dominant regions of the body is some work. Moreover, it is even harder to compute the gradient of the objective function $\nabla\mathcal{F}$ and to anticipate the topological modifications.

Current best solutions The advanced convex hull approach provides the best results for $M < 1.4$ whereas the minimum of linear functions is the best for $M > 1.5$. It is not clear where in the segment $[1.4,\ 1.5]$ the transition happens. But this transition is due to the change of symmetry parameter m from $m = 3$ to $m = 2$.

Natural versus artificial symmetry m The developable body, polytopes and minimum of linear functions all exhibit naturally the phenomenon of symmetry m whereas it has been artificially imposed on the advanced convex hull approach.

Sequence M_m The sequence M_m varies from one approach to another. However, all approaches verify the limits $\lim\limits_{M\to 0} m = \infty$ and $\lim\limits_{M\to\infty} m = 2$.

In conclusion of this chapter, we propose a picture of the evolution in the literature of the best resistance $\mathcal{F}$ for $M = 1.0$ through the years. Fig. 2.21 is a zoom of the green box of Fig. 2.20.

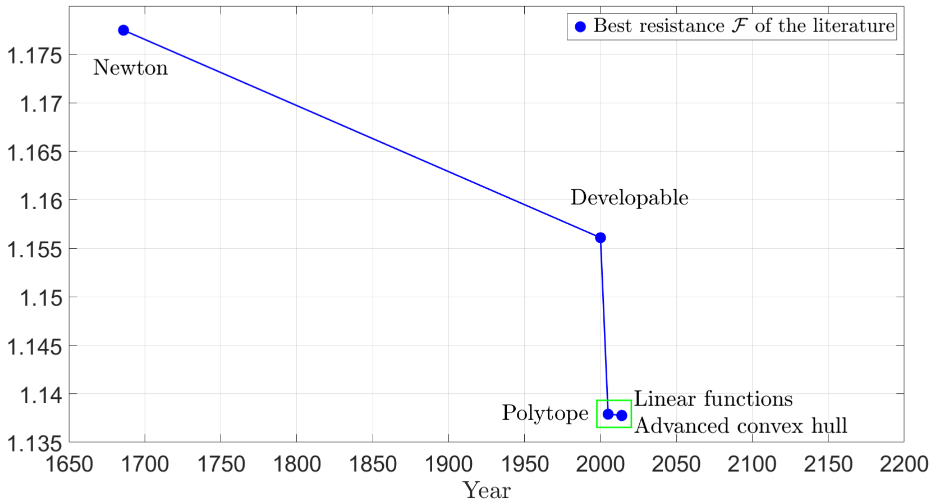


Fig. 2.20 Evolution of the best resistance for $M = 1$ through the years.

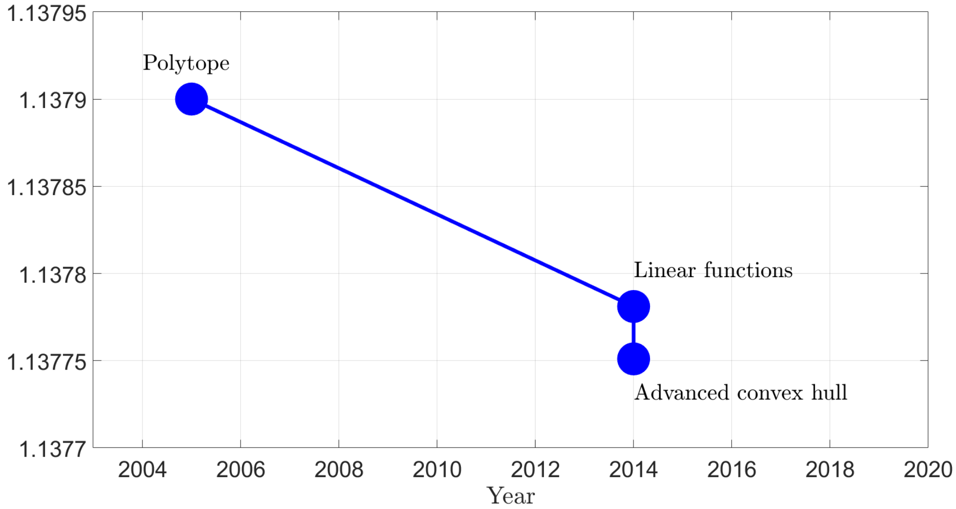


Fig. 2.21 Zoom of Fig. 2.20.

Chapter 3

Smooth minimum method : theoretical framework

In this chapter, we present the theoretical framework of the proposed method. In particular, we detail how to compute the resistance, guarantee the maximal height of the body and how to obtain the optimal solution of our class of functions.

3.1 Idea

The minimum of linear functions provided good results. However, the integral in the resistance required to compute the whole geometry of the current body and the gradient of the resistance is even harder to compute. We propose a new technique to avoid the geometric computations but keep the simple idea of a minimum over linear functions.

We propose to take the smooth minimum of linear functions instead of the exact minimum. We do it thanks to the «LogSumExp» function

$$LSE_\lambda(z_1, z_2, \ldots, z_n) = \frac{1}{\lambda} \log \left(\sum_{i=1}^{n} \exp(\lambda z_i) \right) \tag{3.1}$$

where $\lambda < 0$ is a parameter. The LogSumExp function tends to the true minimum when λ tends to $-\infty$, i.e.,

$$\lim_{\lambda \to -\infty} LSE_\lambda(z_1, \ldots, z_n) = \min\{z_1, \ldots, z_n\}. \tag{3.2}$$

Intuitively, the exact minimum changes suddenly and non-smoothly when a new variable z_i becomes minimal, whereas, the smooth minimum LSE_λ changes smoothly from one variable to another.
Crucially, the exact minimum is not a smooth function and the smooth minimum LSE_λ is a smooth function.

In order to define our body u_λ, we just take the smooth minimum of n linear functions

$$u_\lambda(x, y) = \frac{1}{\lambda} \log \left(\sum_{i=1}^{n} \exp\left(\lambda(a_i x + b_i y + c_i)\right) \right). \tag{3.3}$$

Fig. 3.1 is the comparison, in 2 dimensions, of the exact minimum (in black) and the LSE_λ function for $\lambda = -1$ (in blue), $\lambda = -5$ (in green) and $\lambda = -20$ (in red).

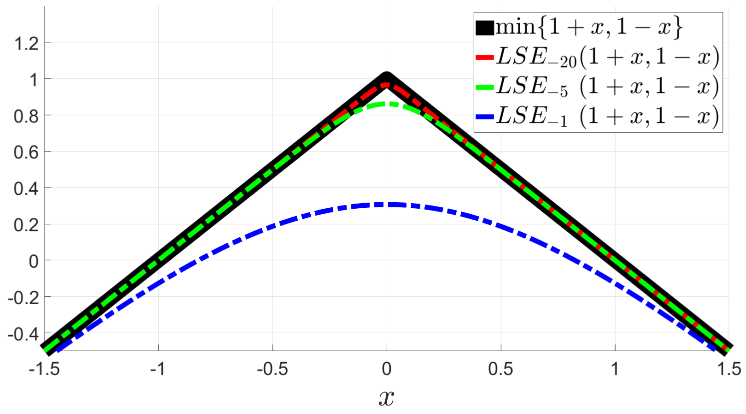


Fig. 3.1 Illustration in 2 dimensions of the LogSumExp.

Similarly, Fig. 3.2 is the comparison, in 3 dimensions, of the exact minimum (in black) and the LogSumExp function (in red) for $\lambda = -10$.

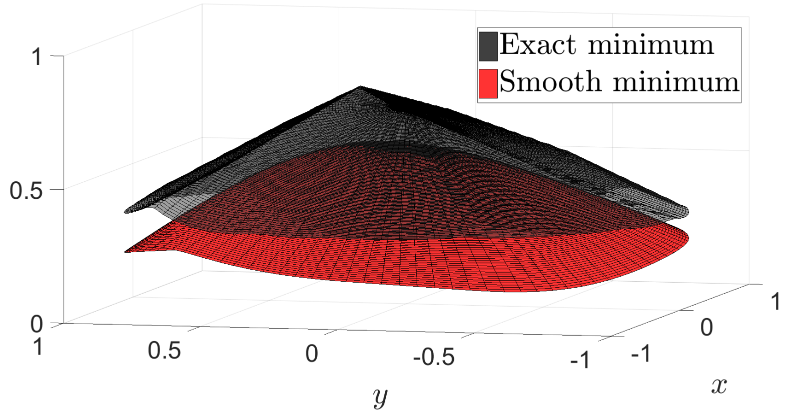


Fig. 3.2 Illustration in 3 dimensions of the LogSumExp.

Fig. 3.1 and Fig. 3.2 exhibit the two important properties of the LogSumExp function, namely, its smoothness and convergence to the exact minimum when λ decreases.

The motivations to represent the body u_λ with the LogSumExp function are multiple.

- The function u_λ is smooth and can be integrated classically without geometric computations of any dominant region. Actually, there is only one «dominant» region, the whole disk Ω, where all linear functions contribute. Recall that the minimum of linear functions approach required to compute the whole geometry of the body at each iteration of the optimization.

- Asymptotically, the solutions obtained with u_λ should be as good as the solutions obtained with u_ℓ the minimum of linear functions method when λ tends to $-\infty$. Recall that the minimum of linear functions is the current best approach for $M > 1.5$ and the second best one for $M < 1.4$.

- The last motivation comes from an observation of a picture of G. Wachsmuth in [15]. The picture is Fig. 3.3. It is a solution obtained by its minimum of linear functions method with $n = 48$ linear functions.

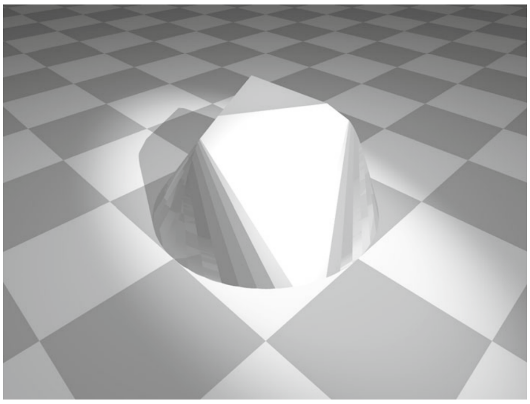


Fig. 3.3 Minimum of $n = 48$ linear functions (from [15]).

On the lateral face of the body, it looks like the linear functions try to approximate a smooth nonlinear function. The hope is that using the smooth minimum will naturally coincide with this unknown smooth nonlinear function. Fig. 3.4 is the illustration of the suspected phenomenon with the exact minimum (in black) and the smooth minimum (in red) of $n = 4$ linear functions with $\lambda = -5$.

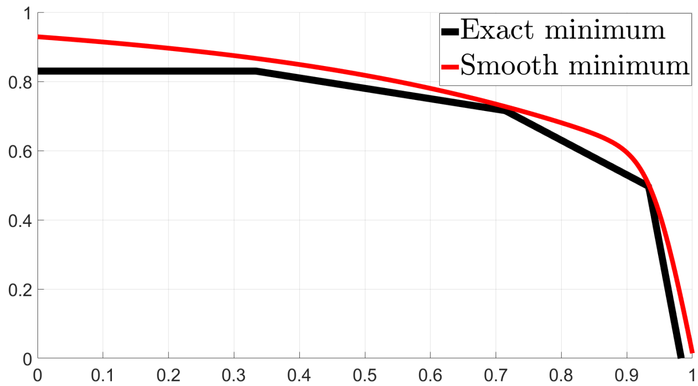


Fig. 3.4 The exact minimum (in black) approximates the smooth minimum (in red).

It is interesting to see that the Cheeger's sets, Fig. 1.7, obtained with the intersection of half-spaces in [9] also seem to be a linear discretization of some smooth function.

We must mention an apparent drawback of the method. The smooth minimum is strictly concave except when all the n variables are equal. It is in contradiction with Theorem 3. However, using a large λ in absolute value will make the smooth minimum tends to the exact minimum and the former will be less and less «strictly concave». Moreover, excluding the optimal solution u^* of our class of functions u_λ does not necessarily mean that we could not reach good results and in particular better results that the current literature.

Finally, we just mention that we will write the problem of finding the optimal function u_λ as a constrained optimization problem. Solving the problem will be done either by `fmincon` of Matlab or by a projected gradient method.

3.2 Computation of the resistance

We present how to compute the resistance of a function u_λ. It is computed with a single numerical integral.

More precisely, recall that the expression of the resistance of a body u is

$$\mathcal{F}(u) = \int_\Omega \frac{1}{1 + ||\nabla u(x,y)||^2} \mathrm{d}x\mathrm{d}y. \tag{3.4}$$

The gradient of the LogSumExp u_λ is everywhere

$$\nabla u_\lambda(x,y) = \frac{\begin{pmatrix} \sum_{i=1}^n a_i \exp\big(\lambda(a_i x + b_i y + c_i)\big) \\ \sum_{i=1}^n b_i \exp\big(\lambda(a_i x + b_i y + c_i)\big) \end{pmatrix}}{\sum_{i=1}^n \exp\big(\lambda(a_i x + b_i y + c_i)\big)} \quad \forall (x,y) \in \Omega \tag{3.5}$$

and its squared norm is

$$||\nabla u_\lambda(x,y)||^2 = \left(\frac{\sum_{i=1}^n a_i \exp\big(\lambda(a_i x + b_i y + c_i)\big)}{\sum_{i=1}^n \exp\big(\lambda(a_i x + b_i y + c_i)\big)} \right)^2 + \left(\frac{\sum_{i=1}^n b_i \exp\big(\lambda(a_i x + b_i y + c_i)\big)}{\sum_{i=1}^n \exp\big(\lambda(a_i x + b_i y + c_i)\big)} \right)^2 \quad \forall (x,y) \in \Omega. \tag{3.6}$$

One of the key point of the smooth minimum method is that the gradient (3.5) is described as a single expression everywhere on the disk Ω. There is no need to compute dominant regions contrary to the expression (2.35) of the minimum of linear functions where the expression depends on the function i.

Computing the resistance (3.4) of the LogSumExp function u_λ yields

$$\mathcal{F}(u_\lambda) = \int_\Omega \frac{1}{1 + \left(\frac{\sum_{i=1}^n a_i \exp\left(\lambda(a_i x + b_i y + c_i)\right)}{\sum_{i=1}^n \exp\left(\lambda(a_i x + b_i y + c_i)\right)} \right)^2 + \left(\frac{\sum_{i=1}^n b_i \exp\left(\lambda(a_i x + b_i y + c_i)\right)}{\sum_{i=1}^n \exp\left(\lambda(a_i x + b_i y + c_i)\right)} \right)^2} \mathrm{d}x\mathrm{d}y \tag{3.7}$$

$$= \int_\Omega \frac{\left(\sum_{i=1}^n \exp\big(\lambda(a_i x + b_i y + c_i)\big) \right)^2}{\left(\sum_{i=1}^n \exp\big(\lambda(a_i x + b_i y + c_i)\big) \right)^2 + \left(\sum_{i=1}^n a_i \exp\big(\lambda(a_i x + b_i y + c_i)\big) \right)^2 + \left(\sum_{i=1}^n b_i \exp\big(\lambda(a_i x + b_i y + c_i)\big) \right)^2} \mathrm{d}x\mathrm{d}y \tag{3.8}$$

$$= \int_\Omega \frac{\sum_{i=1}^n \sum_{j=1}^n \exp\big(\lambda((a_i + a_j)x + (b_i + b_j)y + (c_i + c_j))\big)}{\sum_{k=1}^n \sum_{l=1}^n (1 + a_k a_l + b_k b_l) \exp\big(\lambda((a_k + a_l)x + (b_k + b_l)y + (c_k + c_l))\big)} \mathrm{d}x\mathrm{d}y \tag{3.9}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \int\limits_{\Omega} \frac{\exp\Big(\lambda((a_i + a_j)x + (b_i + b_j)y + (c_i + c_j))\Big)}{\sum\limits_{k=1}^{n}\sum\limits_{l=1}^{n}(1 + a_k a_l + b_k b_l)\exp\Big(\lambda((a_k + a_l)x + (b_k + b_l)y + (c_k + c_l))\Big)} \mathrm{d}x\mathrm{d}y \tag{3.10}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \int\limits_{\Omega} \left(\sum_{k=1}^{n} \sum_{l=1}^{n} \alpha_{kl} \exp\Big(\lambda g_{ijkl}(x, y)\Big) \right)^{-1} \mathrm{d}x\mathrm{d}y. \tag{3.11}$$

The expression (3.11) is a compact way to write the integral where

$$\alpha_{kl} = 1 + a_k a_l + b_k b_l \tag{3.12}$$

and

$$g_{ijkl}(x, y) = (a_k + a_l - a_i - a_j)x + (b_k + b_l - b_i - b_j)y + (c_k + c_l - c_i - c_j) \tag{3.13}$$

The expression (3.8) is the one actually used in the code.

It occurs that there is no way that I am aware of to compute this integral analytically. Hence, we appeal to numerical integration. The numerical integration can either be done using the function `integral2` of MATLAB or using an integration rule adapted to the problem. We will discuss this choice in Sect. 4.3.

3.3 Management of the constraints

3.3.1 Convexity of the body

The convexity constraint of the function u_λ is guaranteed by definition of the LogSumExp. Indeed, the LogSumExp function is a concave function.

Proposition 2. *The LogSumExp function is a concave function for $\lambda < 0$.*

Proof. We show that the LogSumExp fulfills the definition of a concave function, i.e.,

$$f\Big(tx + (1 - t)y\Big) \geq tf(x) + (1 - t)f(y) \quad \forall t \in [0, 1] \text{ and } \forall x, y \in \mathbb{R}^n. \tag{3.14}$$

The proposed proof involves Hölder's inequality

$$\sum_{i=1}^{n} |a_i b_i| \leq \left(\sum_{i=1}^{n} |a_i|^p \right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} |b_i|^q \right)^{\frac{1}{q}} \quad \forall a, b \in \mathbb{R}^n \text{ and } \frac{1}{p} + \frac{1}{q} = 1. \tag{3.15}$$

We will omit the absolute values in Hölder's inequality (3.15) since we will only use strictly positive a_i and b_i. We note the LogSumExp function

$$f(z) = f(z_1, \ldots, z_n) = \frac{1}{\lambda} \log\left(\sum_{i=1}^{n} \exp\Big(\lambda z_i\Big) \right). \tag{3.16}$$

Developing the concavity definition (3.14) with the LogSumExp (3.16) yields

$$f\big(tx+(1-t)y\big) = \frac{1}{\lambda}\log\left(\sum_{i=1}^{n}\exp\big(\lambda(tx_i+(1-t)y_i)\big)\right) \tag{3.17}$$

$$\text{Set } \begin{cases} a_i = \exp(\lambda t x_i) \\ b_i = \exp(\lambda(1-t)y_i) \end{cases} \Bigg\downarrow$$

$$= \frac{1}{\lambda}\log\left(\sum_{i=1}^{n} a_i b_i\right) \tag{3.18}$$

$$\text{Hölder's inequality (3.15) and } \lambda < 0 \Bigg\downarrow$$

$$\geq \frac{1}{\lambda}\log\left(\left(\sum_{i=1}^{n} a_i^p\right)^{\frac{1}{p}}\left(\sum_{i=1}^{n} b_i^q\right)^{\frac{1}{q}}\right) \tag{3.19}$$

$$\text{Set } \begin{cases} t & = \frac{1}{p} \\ 1-t & = \frac{1}{q} \end{cases} \Bigg\downarrow$$

$$= \frac{1}{\lambda}\log\left(\left(\sum_{i=1}^{n} a_i^{\frac{1}{t}}\right)^{t}\left(\sum_{i=1}^{n} b_i^{\frac{1}{1-t}}\right)^{1-t}\right) \tag{3.20}$$

$$= \frac{t}{\lambda}\log\left(\sum_{i=1}^{n} a_i^{\frac{1}{t}}\right) + \frac{1-t}{\lambda}\log\left(\sum_{i=1}^{n} b_i^{\frac{1}{1-t}}\right) \tag{3.21}$$

$$\text{Replace } \begin{cases} a_i^{\frac{1}{t}} & = \exp(\lambda x_i) \\ b_i^{\frac{1}{1-t}} & = \exp(\lambda y_i) \end{cases} \Bigg\downarrow$$

$$= \frac{t}{\lambda}\log\left(\sum_{i=1}^{n}\exp(\lambda x_i)\right) + \frac{1-t}{\lambda}\log\left(\sum_{i=1}^{n}\exp(\lambda y_i)\right) \tag{3.22}$$

$$= tf(x) + (1-t)f(y). \tag{3.23}$$

□

Finally, defining the affine map

$$A(x,y) : \mathbb{R}^2 \to \mathbb{R}^n : (x,y) \to A(x,y) = \begin{pmatrix} a_1x + b_1y + c_1 \\ \vdots \\ a_nx + b_ny + c_n \end{pmatrix}, \tag{3.24}$$

the function

$$u_\lambda(x,y) = f(A(x,y)) = \frac{1}{\lambda}\log\left(\sum_{i=1}^{n}\exp\big(\lambda(a_ix + b_iy + c_i)\big)\right) \tag{3.25}$$

is concave since it is the composition of a concave function, $f(z)$, by an affine map, $A(x,y)$.

3.3.2 Maximal height M

Constraining the body u_λ to have a maximal height M and to attain it is quite hard. Actually, it is the main difficulty and drawback of the smooth minimum method.

Developing the constraints

First, we propose to develop the constraints $u_\lambda \geq 0$ and $u_\lambda \leq M$ in order to emphasize their complexity. Recall that both constraints must hold everywhere on the disk Ω.

Equivalences to $u_\lambda \geq 0$:

$$u_\lambda(x,y) \geq 0 \quad \forall(x,y)\in\Omega \iff \frac{1}{\lambda}\log\left(\sum_{i=1}^{n}\exp\big(\lambda(a_ix+b_iy+c_i)\big)\right) \geq 0 \quad \forall(x,y)\in\Omega$$

$$\lambda < 0 \Big\downarrow$$

$$\iff \log\left(\sum_{i=1}^{n}\exp\big(\lambda(a_ix+b_iy+c_i)\big)\right) \leq 0 \quad \forall(x,y)\in\Omega \tag{3.26}$$

$$\iff \sum_{i=1}^{n}\exp\big(\lambda(a_ix+b_iy+c_i)\big) \leq 1 \quad \forall(x,y)\in\Omega \tag{3.27}$$

$$\iff \begin{cases} \max\limits_{(x,y)\in\mathbb{R}^2} & \sum\limits_{i=1}^{n}\exp\big(\lambda(a_ix+b_iy+c_i)\big) \\ & x^2+y^2 \leq 1 \end{cases} \leq 1 \tag{3.28}$$

$$\text{Minimal value on the frontier} \Big\downarrow$$

$$\iff \begin{cases} \max\limits_{(x,y)\in\mathbb{R}^2} & \sum\limits_{i=1}^{n}\exp\big(\lambda(a_ix+b_iy+c_i)\big) \\ & x^2+y^2 = 1 \end{cases} \leq 1 \tag{3.29}$$

$$\text{Polar coordinates} \Big\downarrow$$

$$\iff \max_{\theta\in[0,2\pi]}\sum_{i=1}^{n}\exp\big(\lambda(a_i\theta+b_i\theta+c_i)\big) \leq 1. \tag{3.30}$$

The condition (3.30) is the simplest characterization of $u_\lambda \geq 0$ everywhere on the disk Ω. Unfortunately, the optimization problem cannot be solved analytically. Indeed, it is an unconstrained problem with the following gradient of objective function to cancel

$$-\lambda\sum_{i=1}^{n}(a_i+b_i)\exp\big(\lambda(a_i\theta+b_i\theta+c_i)\big)=0 \tag{3.31}$$

which can be resumed to solving the following equation in variable θ, noting $\alpha_i = (a_i+b_i)\exp(\lambda c_i)$ and $\beta_i = \lambda(a_i+b_i)$,

$$\sum_{i=1}^{n}\alpha_i\exp(\beta_i\theta)=0. \tag{3.32}$$

There is no way that I am aware of to solve analytically the problem (3.32).

Moreover, note that the problem (3.30) is a nonconvex optimization problem.

Equivalences to $u_\lambda \leq M$:

$$u_\lambda(x,y)\leq M \quad \forall(x,y)\in\Omega \iff \frac{1}{\lambda}\log\left(\sum_{i=1}^{n}\exp\big(\lambda(a_ix+b_iy+c_i)\big)\right) \leq M \quad \forall(x,y)\in\Omega$$

$$\Big\downarrow\ \lambda<0$$

$$\iff \log\left(\sum_{i=1}^{n}\exp\big(\lambda(a_ix+b_iy+c_i)\big)\right) \geq \lambda M \quad \forall(x,y)\in\Omega \tag{3.33}$$

$$\iff \sum_{i=1}^{n}\exp\big(\lambda(a_ix+b_iy+c_i)\big) \geq \exp(\lambda M) \quad \forall(x,y)\in\Omega \tag{3.34}$$

$$\iff \begin{cases}\displaystyle\min_{(x,y)\in\mathbb{R}^2} & \displaystyle\sum_{i=1}^{n}\exp\big(\lambda(a_ix+b_iy+c_i)\big) \\ & x^2+y^2\leq 1\end{cases} \geq \exp(\lambda M) \tag{3.35}$$

The condition (3.35) is the simplest characterization of $u_\lambda \leq M$ everywhere on the disk Ω. Again, it cannot be handled analytically. However, it is a convex optimization problem.

There are several remarks about these equivalences to $u_\lambda \geq 0$ and $u_\lambda \leq M$:

- Both conditions (3.27) and (3.34), respectively for $u_\lambda \geq 0$ and $u_\lambda \leq M$, are unpractical since they must hold for infinitely many points $(x,y)\in\Omega$.
- However, a practical way to use these conditions is to discretize the disk Ω and use only a finite number of the constraints (3.27) and (3.34). Note that for the constraint (3.27), we only need to discretize $\partial\Omega$.

- Besides, note that the optimization problems in (3.30) and (3.35) respectively compute the minimal and maximal values of the body u_λ on the disk Ω, i.e.,

$$\begin{cases} \min\limits_{(x,y)\in\Omega} & u_\lambda(x,y) \\ \\ \max\limits_{(x,y)\in\Omega} & u_\lambda(x,y). \end{cases} \tag{3.36}$$

  Computing these minimum and maximum at each iteration of the optimization of the resistance and constraining them to be respectively higher than zero and lower than M do not work well in practice. Yet, it is important to be able to compute them and we will discuss it later.

Guarantee from the smooth minimum

We saw that applying directly the constraints $0 \leq u_\lambda \leq M$ is not straightforward, therefore, we will look at another way to handle this constraint of maximal height. Indeed, there is a practical and cheap way to guarantee that the body u_λ approximately has a height M. It uses a guarantee on the LogSumExp function.

The LogSumExp function benefits from a property that links it to the exact minimum function. More precisely, the LogSumExp function is always contained between the exact minimum minus a known constant δ and the exact minimum.

Property 1.

$$\min\{z_1,\dots,z_n\} - \delta \leq LSE_\lambda(z_1,\dots,z_n) \leq \min\{z_1,\dots,z_n\} \qquad \forall z \in \mathbb{R}^n \tag{3.37}$$

where $\delta = \frac{\log(1/n)}{\lambda} > 0$.

The positive constant $\delta = \frac{\log(1/n)}{\lambda}$ decreases when λ increases in absolute value. Therefore, in Property 1, the LogSumExp gets closer to the exact minimum. Indeed, we already saw that increasing λ, in absolute value, makes the LogSumExp closer to the exact minimum.

In our context, with the notation u_λ for the smooth minimum of linear functions and u_ℓ for the exact minimum of linear functions, it yields

$$u_\ell(x,y) - \delta \leq u_\lambda(x,y) \leq u_\ell(x,y) \qquad \forall (x,y) \in \Omega. \tag{3.38}$$

Note that we already partially observed Property (3.38) on Fig. 3.1 and Fig. 3.2, the exact minimum (in black) was above the LogSumExp (in red). Fig. 3.5 illustrates completely the property. It shows the exact minimum u_ℓ (in black), the LogSumExp u_λ (in red) and the exact minimum minus the constant $u_\ell - \delta$ (in blue).

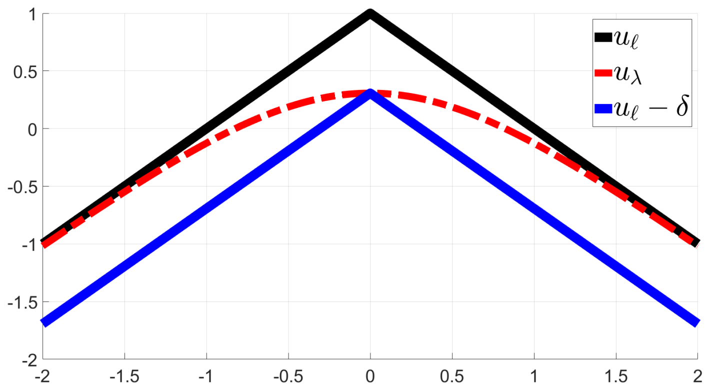


Fig. 3.5 Illustration of Property 1.

Property 1 suggests a way to obtain necessary or sufficient conditions to $0 \leq u_\lambda \leq M$. Indeed, from the right inequality of (3.38) required to be contained between zero and M, i.e.,

$$0 \leq u_\lambda \leq u_\ell \leq M \tag{3.39}$$

we can deduce

$$0 \leq u_\lambda \implies 0 \leq u_\ell \tag{3.40}$$
$$u_\ell \leq M \implies u_\lambda \leq M \tag{3.41}$$

and from the left inequality of (3.38) required to be contained between zero and M, i.e.,

$$0 \leq u_\ell - \delta \leq u_\lambda \leq M \tag{3.42}$$

we can deduce

$$0 \leq u_\ell - \delta \implies 0 \leq u_\lambda \tag{3.43}$$
$$u_\lambda \leq M \implies u_\ell - \delta \leq M \tag{3.44}$$

Now, recall that in the exact minimum case, there were simple conditions to guarantee $0 \leq u_\ell \leq M$, i.e.,

$$u_\ell(x,y) \geq 0 \quad \forall (x,y) \in \Omega \iff c_i \geq \sqrt{a_i^2 + b_i^2} \quad \forall i = 1, \dots, n \tag{3.45}$$
$$u_\ell(x,y) \leq M \quad \forall (x,y) \in \Omega \iff a_1 = b_1 = 0 \text{ and } c_1 = M. \tag{3.46}$$

Remark that $a_1 = b_1 = 0$, $c_1 = M$ is only a sufficient condition to $u_\ell \leq M$. However, we assume that such a plane $z = k$ for some real k is always part of the optimal solution u^*. Therefore, we can consider the relation (3.46) as an equivalence.

Combining the implications (3.40), (3.41), (3.43), (3.44) and the conditions (3.45), (3.46) yields the necessary or sufficient conditions for the smooth case, i.e., $0 \leq u_\lambda$ or $u_\lambda \leq M$.

- Necessary condition to $u_\lambda \geq 0$:

$$c_i \geq \sqrt{a_i^2 + b_i^2} \quad \forall i = 1, \dots, n \iff u_\ell \geq 0 \impliedby u_\lambda \geq 0. \tag{3.47}$$

It exactly corresponds to impose that each linear function i is such that

$$a_i x + b_i y + c_i \geq 0 \tag{3.48}$$

$$\Leftrightarrow \exp\Big(\lambda(a_i x + b_i y + c_i)\Big) \leq 1. \tag{3.49}$$

Clearly, it is guaranteed by (3.27) which is an equivalence to $u_\lambda \geq 0$.

- Sufficient condition to $u_\lambda \geq 0$:

$$c_i \geq \sqrt{a_i^2 + b_i^2} + \delta \quad \forall i = 1, \ldots, n \iff u_\ell - \delta \geq 0 \implies u_\lambda \geq 0. \tag{3.50}$$

It exactly corresponds to impose that each linear function i is such that

$$a_i x + b_i y + c_i \geq \frac{\log(1/n)}{\lambda} = \delta \tag{3.51}$$

$$\Leftrightarrow \exp\Big(\lambda(a_i x + b_i y + c_i)\Big) \leq \frac{1}{n} \tag{3.52}$$

Clearly, it guarantees (3.27) which is an equivalence to $u_\lambda \geq 0$.

- Necessary condition to $u_\lambda \leq M$:

$$a_1 = b_1 = 0 \text{ and } c_1 = M + \delta \iff u_\ell - \delta \leq M \impliedby u_\lambda \leq M. \tag{3.53}$$

- Sufficient condition to $u_\lambda \leq M$:

$$a_1 = b_1 = 0 \text{ and } c_1 = M \iff u_\ell \leq M \implies u_\lambda \leq M. \tag{3.54}$$

It exactly corresponds to impose the linear function $0x + 0y + M$ in the sum of exponential of the LogSumExp which becomes

$$\sum_{i=1}^{n} \exp\Big(\lambda(a_i x + b_i y + c_i)\Big) = \exp(\lambda M) + \sum_{i=2}^{n} \exp\Big(\lambda(a_i x + b_i y + c_i)\Big) \tag{3.55}$$

$$= \exp(\lambda M) + \sum_{i=2}^{n} \exp\Big(\lambda(a_i x + b_i y + c_i)\Big) \tag{3.56}$$

Clearly, it guarantees (3.34) which is an equivalence to $u_\lambda \leq M$.

In the following, we note $\operatorname{Im} u \subset [0, M]$ for a function u taking values in the segment $[0, M]$ but without necessarily reaching zero and M. Whereas, we note $\operatorname{Im} u = [0, M]$ for a function u taking values in the segment $[0, M]$ and that is guaranteed, by the structure of the function u, to reach zero and M.

The pair of sufficient conditions permit us to have a body strictly lying in the segment $[0, M]$. The issue is that these conditions are generally too restrictive, i.e., $u_\lambda(x, y) = 0$ and $u_\lambda(x, y) = M$ are not reached for any single point $(x, y) \in \Omega$.

Slightly more generally, the sufficient conditions modified

$$c_i \geq \sqrt{a_i^2 + b_i^2} + \delta - k_1 \quad \forall i = 1, \ldots, n \tag{3.57}$$

$$a_1 = b_1 = 0 \text{ and } c_1 = M + k_2 \tag{3.58}$$

guarantee that $\operatorname{Im} u_\lambda \subset [-k_1, M + k_2]$ for some real parameter k_1 and k_2. Moreover, the necessary conditions inform us that it is useless to take k_1 and k_2 greater than δ.

Finally, we have a pair of conditions in which we have to select the parameters $k_1 \leq \delta$ and $k_2 \leq \delta$. Using these conditions guarantees us that the body lies in $[-k_1, M + k_2]$. However, the limits $-k_1$ and $M + k_2$ are not necessarily reached.

The practical choice of k_1 and k_2 is discussed later.

Computing the minimal and maximal values

It is important to be able to compute the minimal and maximal values of the body $u_\lambda(x, y)$ on the disk Ω in order to certify whether the body is admissible or not. We propose practical ways to compute them.

- For the minimal value of u_λ on the disk Ω, we already saw that the problem to solve is the unconstrained nonconvex optimization problem of one variable (3.30). We solve it thanks to MATLAB by evaluating 10 000 points θ, taking the discrete minimizer and improving locally it with `fmincon`.
- For the maximal value of u_λ on the disk Ω, it requires to solve the problem (3.35). It occurs that it is a convex optimization problem. Therefore, we compute it thanks to `fmincon`.

  Note that, we cannot predict anything on the location of the maximizer (x, y). In particular, even if at the optimal solution u^*, the maximal value of u^* will be on the point $(0, 0)$, it is not true for the bodies obtained during the optimization. For instance, imposing the single constraint $u_\lambda(0, 0) \leq M$ does not guarantee $u_\lambda(x, y) \leq M \quad \forall(x, y) \in \Omega$. We can observe this phenomenon on the following figure. Fig. 3.6 is the resulting body of the optimization of the resistance $\mathcal{F}$ while using the single constraint $u_\lambda(0, 0) \leq M = 1$.

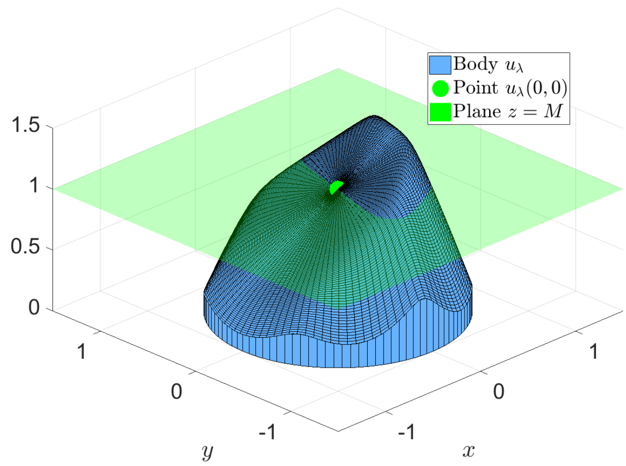


Fig. 3.6 Body obtained using the single constraint $u_\lambda(0, 0) \leq M = 1$.

  We observe that even if $u_\lambda(0, 0) \leq M = 1$, the body gets around the point $(x, y, z) = (0, 0, M = 1)$ to end up higher than $M = 1$.

  However, it will occur that this issue will be naturally solved. In Sect. 4.1, we will impose the symmetry m to the body u_λ. With such a symmetry, the situation in Fig. 3.6 cannot happen anymore.

Cutting and rescaling the body

We saw that there exist constraints which guarantee that the body u_λ strictly lies in $[-k_1, M + k_2]$ for any $k_1 \leq \delta$ and $k_2 \leq \delta$.

The issue is that it does not guarantee that there exists a point (x, y) such that $u_\lambda(x, y) = -k_1$ nor $u_\lambda(x, y) = M + k_2$. Therefore, choosing $k_1 = 0$ and $k_2 = 0$ and even $k_1 > 0$ and $k_2 > 0$ close to zero will generally provide a body with height strictly smaller than M. And having a body lower than M makes us loose some resistance.

Conversely, choosing k_1 and k_2 close or equal to δ will generally provide a body with height strictly higher than M. It is not acceptable. The body must have a total height M since even a slight exceeding may already improve the resistance.

Table 3.1 illustrates both phenomenons. It compares the resistances of the optimal bodies of the literature of height $M = 1$, with a body u_λ smaller than $M = 1$ and a body u_λ higher than $M = 1$. The parenthesis contain the relative differences (r.d.) with the height of reference $M = 1$ or with the resistance of reference $\mathcal{F} = 1.138$.

Table 3.1 Comparison of the resistances of bodies of height exactly M, smaller than M and higher than M.

	Height M	r.d.	Resistance $\mathcal{F}$	r.d.
Optimal	1		1.138	
Smaller	0.94	(−6%)	1.228	(+8%)
Higher	1.02	(+2%)	1.126	(−1%)

We can observe that a small difference in height yields a significant difference in resistance.

It is crucial to have a procedure such that from a body u_λ with $\operatorname{Im} u_\lambda \subset [u_{\min}, u_{\max}]$, we obtain a body $\tilde{u}_\lambda$ with $\operatorname{Im} \tilde{u}_\lambda = [0, M]$. Without altering too much the resistance when $u_{\min} < 0$ or $u_{\max} > M$ and even improving the resistance when $u_{\min} > 0$ or $u_{\max} < M$.

We propose two similar ways to proceed. The first one is the rescaling and the second one is the cutting and rescaling.

Rescaling (R): The idea is to apply the linear map

$$r(z) = Az + B \tag{3.59}$$

to the function $z = u_\lambda(x, y)$ such that

$$\tilde{u}(x, y) = r(u(x, y)) = Au(x, y) + B \in [0, M]. \tag{3.60}$$

The values of A and B are

$$A = \frac{M}{u_{\max} - u_{\min}} \tag{3.61}$$

$$B = \frac{-Mu_{\min}}{u_{\max} - u_{\min}}. \tag{3.62}$$

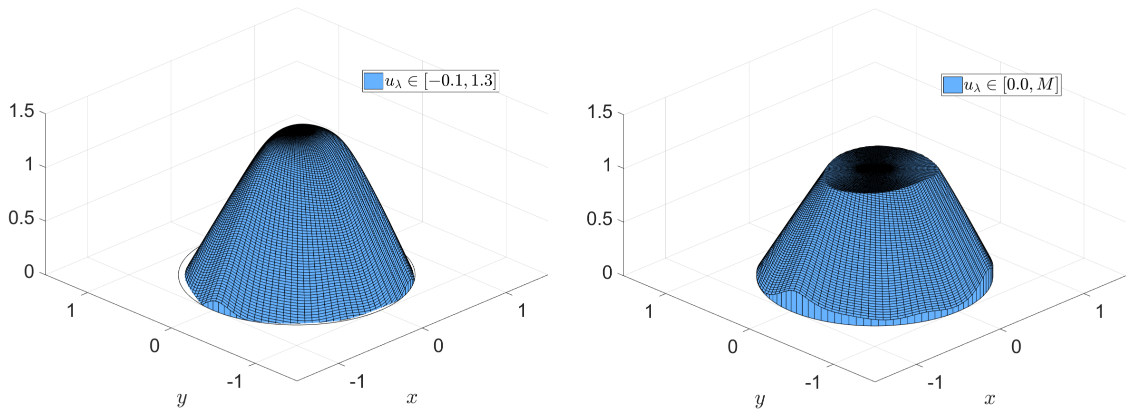
Note that this rescaling does not introduce any additional difficulty in the numerical computation of the resistance $\mathcal{F}$.

Cutting and rescaling (CR): The second way is to cut the body at $z = M$ to have $u_\lambda \leq M$, and then rescale it to have $u_\lambda \geq 0$. Formally, the function $\tilde{u}_\lambda$ after the cutting is

$$\tilde{u}_\lambda(x, y) = \min\{u_\lambda(x, y), M\}. \tag{3.63}$$

Indeed, it corresponds to mix a smooth minimum with an exact minimum. Observe that $\tilde{u}_\lambda$ is still a concave function since it is the exact minimum of two concave functions. After the cutting, we do a rescaling like in (R).

Figs. 3.7 illustrate the effect of the cutting and rescaling applied on a body (left) with height larger than $M = 1$. We can observe that the result of the (CR) (right) lies in $[0, M]$.



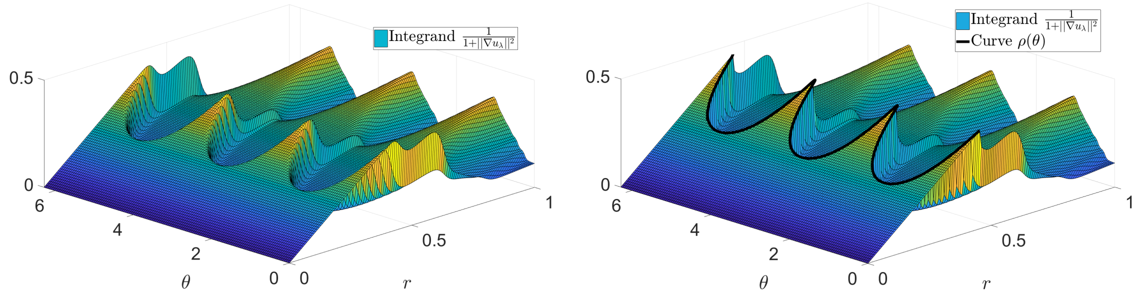
(a) Body u_λ with height $M > 1$. **(b)** Body $\tilde{u}_\lambda$ after the cutting and rescaling (CR).

Fig. 3.7 Effet of cutting and rescaling (CR).

In practice, it occurs that the (CR) is better than (R) for the bodies whose symmetry parameter is $m \geq 3$. For $m = 2$, (R) is better and we will discuss this phenomenon later. As desired, (R) and (CR) improves the resistance of a body smaller than M.

We must notice that the cutting makes the function $\tilde{u}_\lambda$ non-smooth, thus, the function to integrate, $\frac{1}{1+||\nabla \tilde{u}||^2}$, becomes non-continuous. The computation of the resistance cannot be performed in once anymore. The solution that we propose is to separate the area of the domain Ω where $\tilde{u}_\lambda = M$, due to the cutting, and the one where $\tilde{u}_\lambda < M$ thanks to a curve $\rho(\theta)$ that separates these two areas.

Fig. 3.8 is the integrand of the resistance $\mathcal{F}$ of a body before (left) and after (right) the cutting at M. The domain of integration Ω is in polar coordinates (r, θ). In Fig. 3.8b, we distinguish two areas, one from $r = 0$ to $r = \rho(\theta)$, with a constant slope, and one from $r = \rho(\theta)$ to $r = 1$. Formally, the black curve is the set of points $(r, \theta, z) = \left(\rho(\theta), \theta, \frac{1}{1+||\nabla \tilde{u}_\lambda(\rho(\theta),\theta)||^2}\right)$.



(a) Function $\frac{1}{1+||\nabla u_\lambda||^2}$ before cutting.

(b) Function $\frac{1}{1+||\nabla \tilde{u}_\lambda||^2}$ after cutting.

Fig. 3.8 Effect of cutting and rescaling (CR) on the integrand $\frac{1}{1+||\nabla u_\lambda||^2}$ of the resistance. The domain Ω is represented with the polar coordinates.

In practice, we numerically compute the curve $\rho(\theta)$ that separates the two areas. More precisely, we discretize θ into θ_k, and for each k, we search by bisection the ρ^* such that the following equivalence to $u_\lambda(\rho^*, \theta_k) = M$ holds

$$\sum_{i=1}^{n} \exp\big(\lambda(a_i\rho^* \cos(\theta_k) + b_i\rho^* \sin(\theta_k) + c_i)\big) = \exp\big(\lambda M\big) \tag{3.64}$$

knowing that $\rho = 0$ and $\rho = 1$ are lower and upper bounds of the solution ρ^*.

Then, the resistance is

$$\mathcal{F}(\tilde{u}_\lambda) = \int_\Omega \frac{1}{1+||\nabla \tilde{u}_\lambda(x,y)||^2} \mathrm{d}x\mathrm{d}y \tag{3.65}$$

$$= \int_0^{2\pi}\int_0^1 \frac{r}{1+||\nabla \tilde{u}_\lambda(r,\theta)||^2} \mathrm{d}r\mathrm{d}\theta \tag{3.66}$$

$$= \int_0^{2\pi}\int_0^{\rho(\theta)} \frac{r}{1+||0||^2}\mathrm{d}r\mathrm{d}\theta + \int_0^{2\pi}\int_{\rho(\theta)}^1 \frac{r}{1+||\nabla u_\lambda(r,\theta)||^2}\mathrm{d}r\mathrm{d}\theta \tag{3.67}$$

$$= \underbrace{\int_0^{2\pi} \rho^2(\theta)\mathrm{d}\theta}_{\textit{flat part}} + \underbrace{\int_0^{2\pi}\int_{\rho(\theta)}^1 \frac{r}{1+||\nabla u_\lambda(r,\theta)||^2}\mathrm{d}r\mathrm{d}\theta}_{\textit{lateral face}}\,. \tag{3.68}$$

Discussion

In our smoothed case, it is much harder to guarantee $0 \leq u_\lambda \leq M$ whereas in the exact minimum case it was quite simple. The reason is that in the exact minimum case, there was always only one linear function that contributes to u_ℓ on a point (x, y), hence, the conditions are applied individually to each linear function. In the smoothed case, on a point (x, y), all the functions contribute to u_λ. It makes the conditions very hard to establish and impossible to separate.

3.4 Gradient of the resistance

In order to apply a gradient method, we need to be able to compute the gradient of the objective function. As already discussed, we do not have access to an analytic expression of the resistance due to the integral that must be done numerically. Therefore, we cannot directly compute the partial derivatives of the resistance (3.8) with respect to the variables a, b, c.

Yet, it is possible to integrate the derivatives of the integrand instead of differentiating the integral.

$$\nabla_{(a,b,c)}\Big(\mathcal{F}(u_\lambda)\Big) = \nabla_{(a,b,c)}\Big(\int_\Omega \frac{1}{1+||\nabla u_\lambda||^2}\mathrm{d}x\mathrm{d}y\Big) \tag{3.69}$$

$$= \int_\Omega \nabla_{(a,b,c)}\Big(\frac{1}{1+||\nabla u_\lambda||^2}\Big)\mathrm{d}x\mathrm{d}y \tag{3.70}$$

where $\nabla_{(a,b,c)}$ denotes the gradient with respect to the variables a, b, c whereas ∇ denotes the gradient with respect to the geometrical variables x and y.

We note $I(a,b,c) = \frac{1}{1+||\nabla u_\lambda||^2}$ the integrand of the resistance $\mathcal{F}(u_\lambda)$ and

$$\Sigma = \sum_{i=1}^{n} \exp\big(\lambda(a_i x + b_i y + c_i)\big) \tag{3.71}$$

$$\Sigma_a = \sum_{i=1}^{n} a_i \exp\big(\lambda(a_i x + b_i y + c_i)\big) \tag{3.72}$$

$$\Sigma_b = \sum_{i=1}^{n} b_i \exp\big(\lambda(a_i x + b_i y + c_i)\big) \tag{3.73}$$

$$G = \Sigma^2 + \Sigma_a^2 + \Sigma_b^2. \tag{3.74}$$

The derivatives of $I(a,b,c)$ are

$$\frac{\partial I(a,b,c)}{\partial \begin{pmatrix} a_t \\ b_t \\ c_t \end{pmatrix}} = \frac{2\lambda}{G^2}\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \exp\big(\lambda(a_t x + b_t y + c_t)\big)\Big(\Sigma G - \Sigma^2\Big(\Sigma + a_t\Sigma_a + b_t\Sigma_b\Big)\Big) \tag{3.75}$$

Again, there is no way to integrate the expression (3.75) analytically. Hence, we perform a numerical integration. Note that the there are $3n$ integrals to compute.

Observe that the expression (3.75) has a lot of constant terms with respect to the derivative variable a_t, b_t or c_t. We emphasized in red the only term that the depends on the variable of derivation.

3.5 Projection on the admissible set

The only practical constraints available are the constraints (3.57) for $u_\lambda \geq 0$ and (3.58) for $u_\lambda \leq M$. The latter can be directly satisfied by considering the variables a_1, b_1 and c_1 as fixed. For the former, we keep it on the constraints of the following optimization problem

$$\begin{aligned} \min_{(a,b,c)\in\mathbb{R}^{3n}} \quad & \mathcal{F}(u_\lambda) \\ \text{s.t.} \quad & c_i \geq \sqrt{a_i^2 + b_i^2} + \delta - k_1 \qquad \forall i = 1, \ldots, m \end{aligned} \tag{3.76}$$

where $\mathcal{F}(u_\lambda)$ is the resistance given by the expression (3.8) depending on a, b and c. The parameters δ and k_1 are known and fixed.

We want to apply a projected gradient method to the problem (3.76). It requires to have a way to project a given solution (a, b, c) on the set of admissible solutions, i.e.,

$$\mathcal{X} = \left\{ (a, b, c) \in \mathbb{R}^{3n} \mid c_i \geq \sqrt{a_i^2 + b_i^2} + \delta - k_1 \quad \forall i = 1, \dots, m \right\}. \tag{3.77}$$

Fig. 3.9 is a 3d representation of the admissible set $\mathcal{X}$ for $\delta - k_1 = 0$. The admissible points (a, b, c) are «inside» the green surface, i.e., they belong to its epigraph.

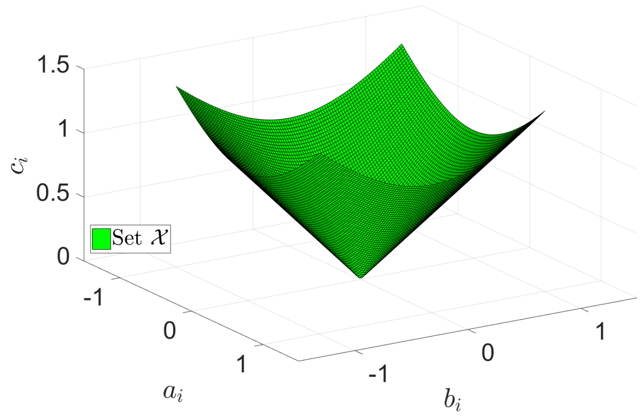


Fig. 3.9 Set $\mathcal{X}$ of admissible points (a, b, c).

Given the point (a, b, c), the problem that provides the projected point $(\tilde{a}, \tilde{b}, \tilde{c})$ is

$$\begin{aligned} \min_{(\tilde{a},\tilde{b},\tilde{c}) \in \mathbb{R}^{3n}} \quad & \sum_{i=1}^{n} (\tilde{a}_i - a_i)^2 + (\tilde{b}_i - b_i)^2 + (\tilde{c}_i - c_i)^2 \\ \text{s.t.} \quad & \tilde{c}_i \geq \sqrt{\tilde{a}_i^2 + \tilde{b}_i^2} + \delta - k_1 \qquad \forall i = 1, \dots, m. \end{aligned} \tag{3.78}$$

Clearly, the problem (3.78) is separable with respect to each function i. Indeed, the projection corresponds to take each function $a_i x + b_i y + c_i$ and to find the closest function $\tilde{a}_i x + \tilde{b}_i y + \tilde{c}_i$ such that the latter is higher than zero everywhere on the disk Ω.

Therefore, we can solve the problem for each function i

$$\begin{aligned} \min_{(\tilde{a}_i,\tilde{b}_i,\tilde{c}_i) \in \mathbb{R}^{3}} \quad & (\tilde{a}_i - a_i)^2 + (\tilde{b}_i - b_i)^2 + (\tilde{c}_i - c_i)^2 \\ \text{s.t.} \quad & \tilde{c}_i \geq \sqrt{\tilde{a}_i^2 + \tilde{b}_i^2} + \delta - k_1. \end{aligned} \tag{3.79}$$

Solving the problem (3.79) analytically yields, $\forall i$ such that $c_i < \sqrt{a_i^2 + b_i^2} + \delta - k_1$,

$$\tilde{a}_i = \frac{a_i}{2} \left(1 + \frac{c_i}{\sqrt{a_i^2 + b_i^2}} \right) \tag{3.80}$$

$$\tilde{b}_i = \frac{b_i}{2} \left(1 + \frac{c_i}{\sqrt{a_i^2 + b_i^2}} \right) \tag{3.81}$$

$$\tilde{c}_i = \sqrt{\tilde{a}_i^2 + \tilde{b}_i^2} + \delta - k_1. \tag{3.82}$$

and we do not modify the a_i, b_i and c_i that do respect the condition $c_i \geq \sqrt{a_i^2 + b_i^2} + \delta - k_1$. Note that the projection of the point $(a, b, c) \in \mathbb{R}^{3n}$ has a computational cost of $\mathcal{O}(n)$.

Chapter 4

Smooth minimum method : practical aspects

In this chapter, we analyze the practical details of the smooth minimum method. In particular, we present how it is possible to use Conjecture 1 of symmetry to save computational efforts, we also discuss the choice of the parameter λ and the numerical precision and stability of the computation of the resistance and its gradient.

Finally, we summarize the whole numerical procedure followed to obtain the best possible results. Mainly, we either optimize the resistance thanks to `fmincon` of MATLAB or by a projected gradient method and we make the body admissible by means of (CR) or (R). While using `fmincon`, the algorithms *sqp* and *active-set* provided the best results regarding the execution time. Therefore, each time we refer to `fmincon`, we used it with one of these two algorithms.

4.1 Taking advantage of the symmetry

It appears that all the bodies proposed in the literature benefits from the symmetry of parameter m defined in Conjecture 1. G. Wachsmuth. took advantage of this symmetry and imposed it on its minimum of linear functions to save computational cost. The latter being precious for our approach, we also use this assumption.

Nevertheless, we started by performing experiments without the assumption of symmetry m. It occurs that the symmetry naturally appears by itself. Therefore, we considered to use the assumption. We present how it can be done practically.

Given n linear functions, the body u_λ expressed in polar coordinates is

$$u_\lambda(r,\theta) = \frac{1}{\lambda}\log\left(\sum_{i=1}^{n}\exp\left(\lambda f_i(r,\theta)\right)\right) \tag{4.1}$$

where $f_i(r,\theta) = a_i r\cos(\theta) + b_i r\sin(\theta) + c_i$ with a_i, b_i, c_i known.

In order to apply the symmetry m, we copy each function f_i on each of the $2m$ sectors of the disk Ω. Therefore, the body becomes

$$u_\lambda(r,\theta) = \frac{1}{\lambda}\log\left(\sum_{i=1}^{n}\sum_{j=0}^{m-1}\exp\left(\lambda g_{i,j}(r,\theta)\right) + \exp\left(\lambda h_{i,j}(r,\theta)\right)\right) \tag{4.2}$$

where $g_{i,j}(r,\theta) = a^g_{i,j} r\cos(\theta) + b^g_{i,j} r\sin(\theta) + c^g_{i,j}$ and $h_{i,j}(r,\theta) = a^h_{i,j} r\cos(\theta) + b^h_{i,j} r\sin(\theta) + c^h_{i,j}$ are linear functions with parameters to determine geometrically.

More precisely, noting the angles $\phi_j = \frac{2\pi}{2m} j$, the definition of the functions g and h are

$$g_{i,j}(r,\theta) = f_i(r, \theta - \phi_{2j}) \qquad \forall i = 1,\dots,n,\ j = 0,\dots,m-1 \tag{4.3}$$
$$h_{i,j}(r,\theta) = f_i(r, -\theta - \phi_{2j+2}) \qquad \forall i = 1,\dots,n,\ j = 0,\dots,m-1. \tag{4.4}$$

In other words, $g_{i,j}$ is the function f_i but on the sector $\theta \in [\phi_{2j}, \phi_{2j+1}]$ whereas $h_{i,j}$ is the function f_i but reversed and on the sector $\theta \in [\phi_{2j+1}, \phi_{2j+2}]$.

Fig. 4.1 is a diagram of the situation for $m = 3$. It indicates in which sector the functions $g_{i,j}$ or $h_{i,j}$ copy the function f_i. The arrows indicate if the function f_i is reversed (green arrow) or not (blue arrow). The angle ϕ_j are represented.

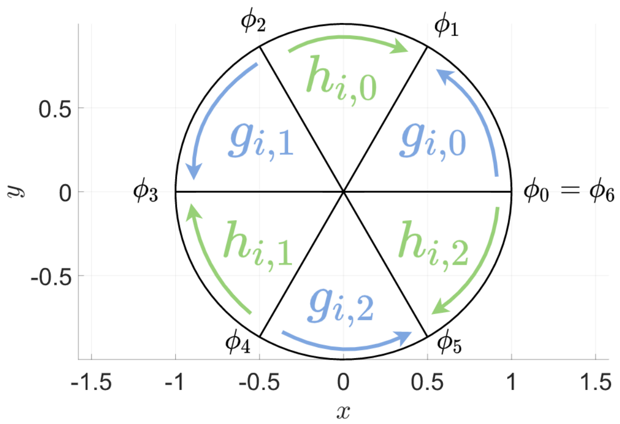


Fig. 4.1 Sectors in which the functions $g_{i,j}$ and $h_{i,j}$ copy the function f_i.

We can determine explicitly the parameters $a^g_{i,j}, b^g_{i,j}, c^g_{i,j}, a^h_{i,j}, b^h_{i,j}, c^h_{i,j}$ in terms of a_i, b_i, c_i. Indeed, by definition of the function $g_{i,j}(r,\theta) = f_i(r, \theta - \phi_{2j})$, it must satisfy

$$\begin{cases} g_{i,j}(0,0) & = & f_i(0,0) \\ g_{i,j}(1,\phi_{2j}) & = & f_i(1,0) \\ g_{i,j}(1,\phi_{2j+1}) & = & f_i(1,\phi_1). \end{cases} \tag{4.5}$$

In addition to $c^g_{i,j} = c_i$, it yields the following linear system

$$\begin{pmatrix} \cos(\phi_{2j}) & \sin(\phi_{2j}) \\ \cos(\phi_{2j+1}) & \sin(\phi_{2j+1}) \end{pmatrix} \begin{pmatrix} a^g_{i,j} \\ b^g_{i,j} \end{pmatrix} = \begin{pmatrix} a_i \\ a_i\cos(\phi_1) + b_i \sin(\phi_1) \end{pmatrix} \tag{4.6}$$

and the solution is

$$\begin{pmatrix} a^g_{i,j} \\ b^g_{i,j} \end{pmatrix} = \frac{\begin{pmatrix} \sin(\phi_{2j+1}) a_i - \sin(\phi_{2j})\big(a_i \cos(\phi_1) + b_i \sin(\phi_1)\big) \\ -\cos(\phi_{2j+1}) a_i + \cos(\phi_{2j})\big(a_i\cos(\phi_1) + b_i\sin(\phi_1)\big) \end{pmatrix}}{\cos(\phi_{2j})\sin(\phi_{2j+1}) - \sin(\phi_{2j})\cos(\phi_{2j+1})}. \tag{4.7}$$

Similarly, by definition of the function $h_{i,j}(r,\theta) = f_i(r, -\theta + \phi_{2j+2)})$, it must satisfy

$$\begin{cases} h_{i,j}(0,0) & = & f_i(0,0) \\ h_{i,j}(1,\phi_{2j+2}) & = & f_i(1,0) \\ h_{i,j}(1,\phi_{2j+1}) & = & f_i(1,\phi_1). \end{cases} \tag{4.8}$$

In addition to $c_{i,j}^h = c_i$, it yields the following linear system

$$\begin{pmatrix} \cos(\phi_{2j+2}) & \sin(\phi_{2j+2}) \\ \cos(\phi_{2j+1}) & \sin(\phi_{2j+1}) \end{pmatrix} \begin{pmatrix} a_{i,j}^h \\ b_{i,j}^h \end{pmatrix} = \begin{pmatrix} a_i \\ a_i \cos(\phi_1) + b_i \sin(\phi_1) \end{pmatrix} \tag{4.9}$$

and the solution is

$$\begin{pmatrix} a_{i,j}^h \\ b_{i,j}^h \end{pmatrix} = \frac{\begin{pmatrix} \sin(\phi_{2j+1})a_i - \sin(\phi_{2j+2})\big(a_i \cos(\phi_1) + b_i \sin(\phi_1)\big) \\ -\cos(\phi_{2j+1})a_i + \cos(\theta_{2j+2})\big(a_i \cos(\phi_1) + b_i \sin(\phi_1)\big) \end{pmatrix}}{\cos(\phi_{2j+2})\sin(\phi_{2j+1}) - \sin(\phi_{2j+2})\cos(\phi_{2j+1})}. \tag{4.10}$$

We define the function

$$sym : \mathbb{R}^{3n} \to \mathbb{R}^{2m3n} : (a,b,c) \to sym(a,b,c) = (a^g, a^h, b^g, b^h, c^g, c^h) \tag{4.11}$$

where a^g, b^g and a^h, b^h are respectively given by (4.7) and (4.10) and $c^g = c^h = c$.

In order to clarify, recall that $a, b, c \in \mathbb{R}^n$ and $a^g, a^h, b^g, b^h, c^g, c^h \in \mathbb{R}^{nm}$. In the following, we will note $(a_{aug}, b_{aug}, c_{aug}) = (a^g, a^h, b^g, b^h, c^g, c^h)$.

Knowing the parameters (a, b, c), the resistance $\mathcal{F}$ depends on the the whole parameters $(a_{aug}, b_{aug}, c_{aug}) = sym(a, b, c)$. In particular, the resistance is

$$\mathcal{F}(sym(a,b,c)). \tag{4.12}$$

Moreover, at each iteration of the projected gradient method, we update the parameters (a, b, c). Therefore, we need to compute the derivatives of the function $\mathcal{F}(sym(a, b, c))$ with respect to the parameters (a, b, c). Thanks to the chain rule, we have

$$\nabla_a \mathcal{F}(sym(a,b,c)) = \nabla_{(a_{aug},b_{aug},c_{aug})} \mathcal{F}(a_{aug}, b_{aug}, c_{aug})\Big|_{(a_{aug},b_{aug},c_{aug})=sym(a,b,c)} \cdot J_a \tag{4.13}$$

$$\nabla_b \mathcal{F}(sym(a,b,c)) = \nabla_{(a_{aug},b_{aug},c_{aug})} \mathcal{F}(a_{aug}, b_{aug}, c_{aug})\Big|_{(a_{aug},b_{aug},c_{aug})=sym(a,b,c)} \cdot J_b \tag{4.14}$$

$$\nabla_c \mathcal{F}(sym(a,b,c)) = \nabla_{(a_{aug},b_{aug},c_{aug})} \mathcal{F}(a_{aug}, b_{aug}, c_{aug})\Big|_{(a_{aug},b_{aug},c_{aug})=sym(a,b,c)} \cdot J_c \tag{4.15}$$

where J_a, J_b, J_c are some part of the Jacobian of the function sym. More precisely, the Jacobian J of sym is $(2m3n) \times 3n$ and J_a is the n first columns, J_b the next n and J_c the last n columns. Observe that, since sym is linear, J does not depend on (a, b, c).

Assuming the symmetry m provides an appreciable gain of computational effort. Indeed, since the body u_λ is m-symmetric, we can compute the integral only on one sector of the disk Ω and then we multiply the result by $2m$. As the numerical integration is the most expensive step of our method, the improvement is significant. Note that the larger m is, the larger the gain is.

Note that we can observe in Fig. 3.8 that the function $\frac{1}{1+||\nabla u_\lambda||^2}$ to integrate has indeed a symmetry $m = 3$.

By the way, as we mentioned in Sect. 3.3.2, with such a symmetry m, imposing $u_\lambda(0,0) \leq M$ will guarantee $u_\lambda(x,y) \leq M \quad \forall (x,y) \in \Omega$. Indeed, with this symmetry the body cannot outflank the point $(0,0,M)$ by one side only, as in Fig. 3.6, since by symmetry it should do the same in the other sides.

Finally, note that the effective number of linear functions is now given by $2mn$.

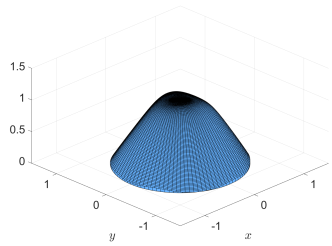
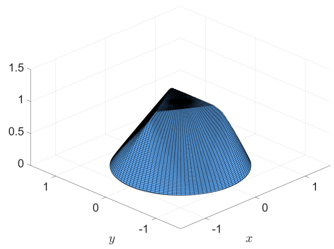
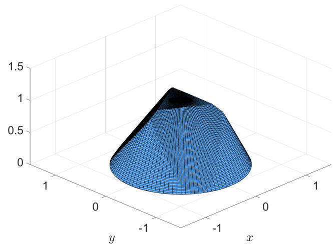
4.2 Choice of λ

A body u_λ is characterized by the decision variables, a, b, c, and by the parameter λ. Indeed, for the moment, we considered λ as a fixed parameter.

However, the choice of λ is very important and we must choose it wisely. At first sight, it is not clear which value of λ we should choose or even how we can determine such a value. The only clear information is that taking $\lambda \to -\infty$ should provide bodies as good as the minimum of linear functions approach. Yet, it is not excluded that there are better possible values for λ. Moreover, note that a large λ, in absolute value, introduces numerical instabilities since we are dealing with expressions of the form $\exp\big(\lambda(\ldots)\big)$. Furthermore, when λ tends to $-\infty$, the body u_λ becomes less smooth and the integrand of the resistance, $\frac{1}{1+||\nabla u_\lambda||^2}$, becomes less continuous. Therefore, the integral to compute the resistance is harder to perform with the same accuracy and takes more time.

First of all, we compare bodies of different λ and their resistances $\mathcal{F}$. More precisely, the bodies were optimized with 20 iterations of `fmincon`. Table 4.1 contains the resistance $\mathcal{F}$, the execution time of the 20 iterations and an illustration of the bodies u_λ for $\lambda = -10$, $\lambda = -100$ and $\lambda = -1000$. The bodies have $2mn = 180$ linear functions and a height $M = 1$.

Table 4.1 Bodies optimized with 20 iterations of `fmincon` for different values of λ.

	$\lambda = -10$	$\lambda = -100$	$\lambda = -1000$
$\mathcal{F}$	1.231 7	1.149 4	1.146 5
Time	14s	97s	827s
			

The execution time is considerably larger when λ is increased in absolute value. For $\lambda = -1000$, the optimization lasts almost 15 minutes. It is due to the integral in the resistance that takes much more time to reach the same level of precision when increasing λ in absolute value. Moreover, the resistance decreases when λ increases in absolute value.

Secondly, given a good body (a, b, c) optimized for some λ, we analyze the evolution of the resistance while varying the parameter λ only.

Fig. 4.2 is the evolution of the resistance $\mathcal{F}$ while varying the parameter λ of a body optimized for $\lambda = -100$. We applied the cutting and rescaling before reporting the resistance.

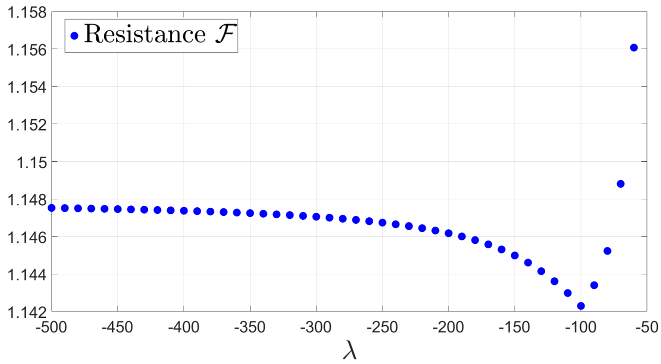


Fig. 4.2 Evolution of the resistance while changing λ.

It appears that the body optimized for $\lambda = -100$ is optimal with respect to λ in $\lambda = -100$. Recall that we apply (CR) such that all the bodies have a height of M exactly.

Thirdly, note that it is possible to apply a block coordinate descent. Namely, we optimize with respect to (a, b, c) with fixed λ and, then, we optimize with respect to λ with fixed (a, b, c) and so on. This strategy will be used to find the best results.

Finally, we cannot clearly conclude on how to choose the parameter λ. It occurs that increasing λ increases the execution time of the optimization. Therefore, in practice, we tested several λ and found empirically the best one. We will provide the best results and their associated λ in Sect. 5.2.

4.3 Precision and performance of the numerical integration

Unfortunately, the resistance $\mathcal{F}(u_\lambda)$ in (3.8) and the gradient of the resistance $\nabla\mathcal{F}(u_\lambda)$ in (3.69) cannot be integrated analytically. Therefore, we must perform these integrals numerically. The shape of the integrand of $\mathcal{F}(u_\lambda)$ to integrate was depicted in Fig. 3.8.

We propose two ways to compute these integrals. The first one is to use the function `integral2` [20] of MATLAB and the second one is to use the 2d Simpson's rule [21]. The first one is potentially very well designed whereas the second can be particularized to our specific problems.

`integral2` is a MATLAB function that allows to compute a two-dimensional integral of the form

$$I = \int_a^b \int_{c(x)}^{d(x)} f(x, y)\mathrm{d}x\mathrm{d}y. \tag{4.16}$$

It is an adaptive strategy that evaluates the function f on different points, and locally increases the number of evaluated points on the region of the domain where f appears to be less regular. We can either specify the maximal number of evaluated points by `integral2` or a desired precision ϵ. Indeed, the function tries to provide an estimation $I_{\texttt{integral2}}$ of the integral I such that

$$|I - I_{\texttt{integral2}}| \leq \epsilon. \tag{4.17}$$

If we do not specify a desired precision ϵ, then the function will return an estimation of the error $\epsilon = |I - I_{\texttt{integral2}}|$.

In the following, we attempt to compare the precision and the execution times of `integral2` and the 2d Simpson's rule for the computation of the resistance and gradient of the resistance. For the resistance, compare the case without cutting and the case CR.

Integrand of the resistance

Before analyzing the computation of the integral of the resistance $\mathcal{F}$, we study the integrand function of this integral over the domain Ω.

In the code, the integrand of the resistance (3.8) is implemented by the function `integrand`. The function has a computational cost of $\mathcal{O}(n)$, n being the effective number of linear functions of the body. The computation of the integral depends on the parameter λ whereas the evaluation of the integrand does not depend it.

Fig. 4.3 is the evolution of the execution time of the function `integrand` to evaluate 10^6 points (r, θ) for different number n of linear functions. It shows that the practical complexity is indeed $\mathcal{O}(n)$.

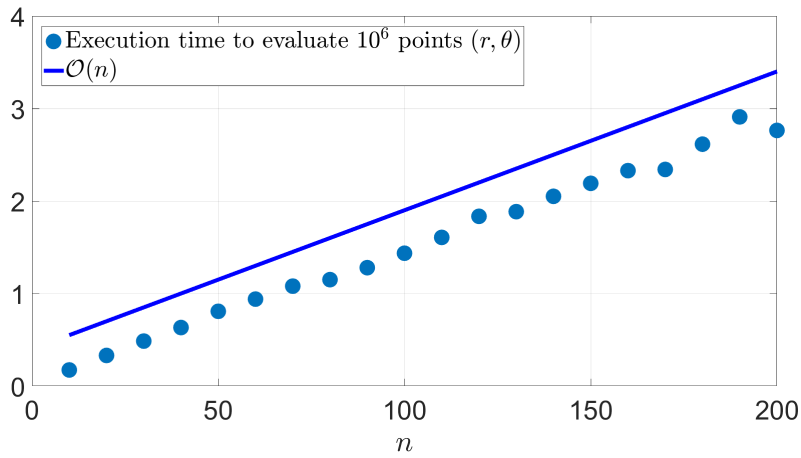


Fig. 4.3 Execution time of the function `integrand` to evaluate 10^6 points (r, θ).

Resistance without cutting

We want to compute the single integral in (3.8). In Table 4.2, we report the resistance $\mathcal{F}$ computed by `integral2` and the 2d Simpson's rule, their execution time and the guarantee of maximal error ϵ of `integral2` for different number of points used by `integral2` and the 2d Simpson's rule. The body has $2mn = 180$ linear functions and $\lambda = -100$. We bold the digits common with the most accurate resistance available.

Table 4.2 Comparison of the performances of `integral2` and the 2d Simpson's rule to compute the resistance $\mathcal{F}$ with $2mn = 180$ linear functions and $\lambda = -100$.

	`integral2`			2d Simpson's rule	
Points	Resistance $\mathcal{F}$	Time	ϵ	Resistance $\mathcal{F}$	Time
10	**1.146** 909 466 761 723	0.006s	3.4 10^{-4}	**1.1**91 218 802 791 675	0.001s
10^2	**1.146 679** 397 305 740	0.035s	1.4 10^{-7}	**1.14**3 365 478 047 454	0.001s
10^3	**1.146 679 410 266 07**2	0.296s	2.3 10^{-13}	**1.146** 945 342 253 993	0.003s
10^4	**1.146 679 410 266 080**	0.431s	2.5 10^{-14}	**1.146 6**29 953 802 800	0.022s
10^5	**1.146 679 410 266 080**	0.420s	2.5 10^{-14}	**1.146 679** 397 940 379	0.263s

We observe that with 10^4 points evaluated, `integral2` already reaches a precision close to the epsilon-machine $\epsilon_m \approx 10^{-16}$. Moreover, increasing the number of points consistently increases the number of supposed correct digits. It strongly makes us confident that `integral2` is computing the correct value of the resistance with a sufficient precision. Therefore, in the following, all the resistances reported were calculated with `integral2` with a sufficient precision.

The resistances obtained by the 2d Simpson's rule are also consistent but less accurate.

We note $\mathcal{F}^*$ the resistance obtained by `integral2` with a desired precision $\epsilon = 10^{-16}$. Fig. 4.4 contains the error $|\mathcal{F}^* - \mathcal{F}|$ (blue dots) for different values of desired precision ϵ. It also contains the desired precision ϵ (blue line). Actually, the blue line is $2m\epsilon$ since we compute the integral on one sector and multiply the result by $2m$.

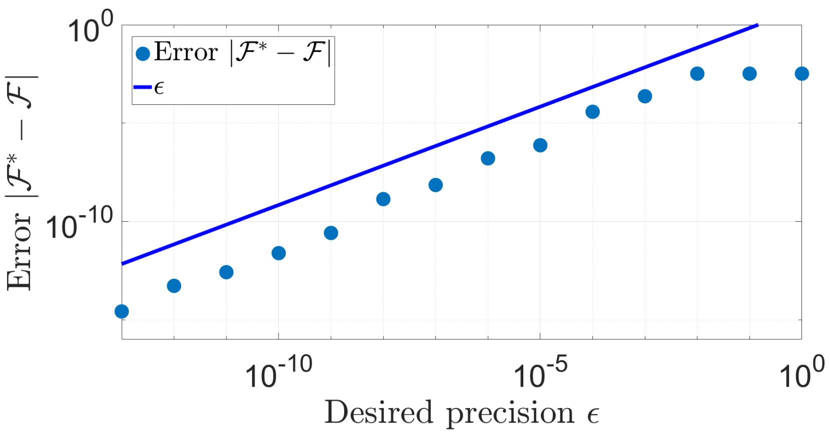


Fig. 4.4 Error $|\mathcal{F}^* - \mathcal{F}|$ of `integral2` for different values of desired precision ϵ.

We observe that the guarantee ϵ is always valid.

Recall that the LogSumExp function tends to the exact minimum when λ decreases, see Property 1 and equation (3.2). It has a consequence on the performance of the integration. For instance, $u_{\lambda=-1}$ is smoother, thus, easier to integrate, than $u_{\lambda=-100}$.

We illustrate this phenomenon. Table 4.3 contains the execution time of `integral2` to reach a precision $\epsilon = 10^{-14}$ to compute the resistance $\mathcal{F}$ of a body with $2mn = 180$ functions for different λ.

Table 4.3 Evolution of the execution time of `integral2` for different values of λ.

λ	Time
-1	0.0s
-10	0.1s
-10^2	0.7s
-10^3	8.8s
-10^4	361.5s

We indeed observe that reaching the same level of precision requires much more time for a large absolute value of λ than for a smaller one. Similarly, Fig. 4.5 is the evolution of the time to compute the resistance $\mathcal{F}$ for different values of λ. The horizontal axis is $|\lambda|$. We can observe that the execution time (green dots) follows a $\mathcal{O}(|\lambda|)$ curve (green line).

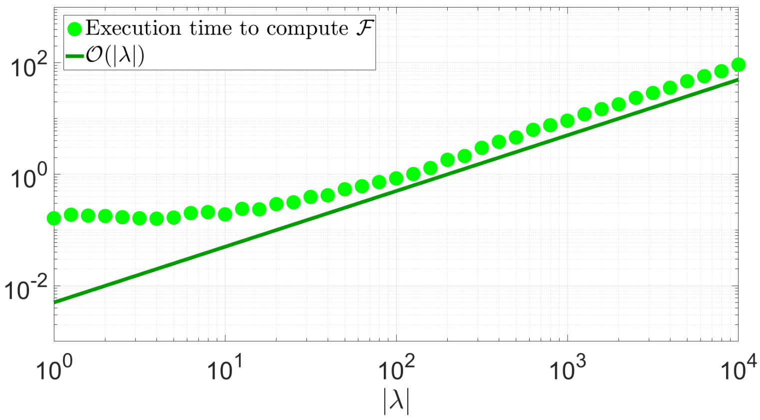


Fig. 4.5 Evolution of the execution time to compute $\mathcal{F}$ for different values of λ.

Resistance with cutting

We now run the same experiments in the case (CR) where we cut the body at M. Indeed, the integrand becomes non-continuous, hence, the integral must be evaluated in two steps thanks to the curve $\rho(\theta)$.

In Table 4.4, we report the resistance $\mathcal{F}$, the execution time and the guarantee of maximal error for a body with $2mn = 180$ linear functions, $\lambda = -100$ and (CR) for different number of evaluated points by `integral2`.

Table 4.4 Performances of `integral2` to compute the resistance $\mathcal{F}$ with $2mn = 90$ linear functions and (CR).

Points	Resistance $\mathcal{F}$	Time	ϵ
10	**1.143** 572 819 037 340	0.598s	6.2 10^{-4}
10^2	**1.143 494** 581 028 606	0.713s	1.5 10^{-7}
10^3	**1.143 494 729 324** 450	1.708s	5.5 10^{-14}
10^4	**1.143 494 729 324 471**	2.368s	1.8 10^{-14}
10^5	**1.143 494 729 324 471**	2.342s	1.8 10^{-14}

Similarly to the case without (CR), Table 4.2, the epsilon-machine is reached with 10^4 points. And, the number of correct digits increases consistently. Fig. 4.6 contains the error $|\mathcal{F}^* - \mathcal{F}|$ (blue dots) for different values of desired precision ϵ. Again, the guarantee holds for all values of ϵ.

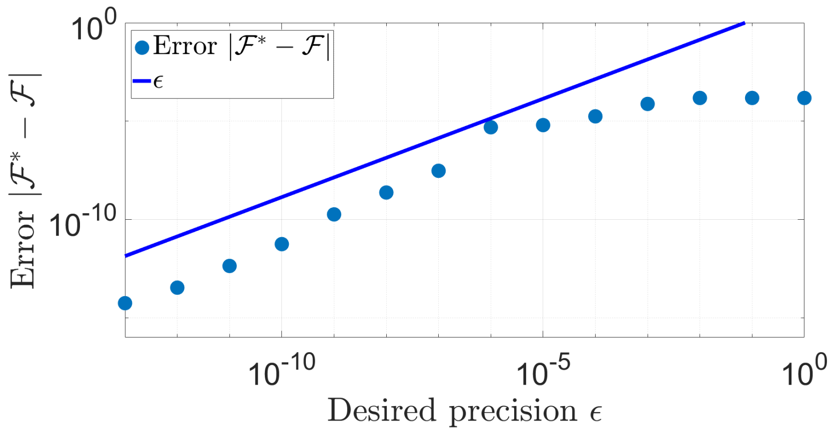


Fig. 4.6 Error $|\mathcal{F}^* - \mathcal{F}|$ of `integral2` for different values of desired precision ϵ in the (CR) case.

The computation of the integral also suffers from the same problem of lower smoothness while increasing the absolute value of λ. Table 4.5 contains the execution time of `integral2` to reach a precision $\epsilon = 10^{-14}$ to compute the resistance $\mathcal{F}$ of a body with $2mn = 180$ linear functions and CR for different values of λ.

Table 4.5 Evolution of the execution time of `integral2` for different values of λ in the CR case.

λ	Time
-1	0.18s
-10	0.57s
-10^2	2.25s
-10^3	23.89s
-10^4	448.41s

Similarly to the case without cutting, Table 4.3, increasing λ in absolute value increases the execution time.

Integration of the gradient of the resistance $\nabla\mathcal{F}$

Now, we want to compute the gradient of the resistance, i.e., we integrate the function (3.75).

Recall that there are $3n$ integrals to compute and that there are a lot of constant terms. These terms are constant with respect to the variable of derivations a_t, b_t and c_t but not to the variable of integration x and y. However, `integral2` forbids to integrate a vector. Therefore, we are forced to compute all the terms, even the constant ones, for each integral.

On the contrary, we do not face this issue while using the 2d Simpson's rule implemented by hand. We can compute and store the constant terms and use them when needed. The reason why it works here is the fact that the Simpson's rule always uses the same points (x, y) to evaluate the functions whereas `integral` uses an adaptive strategy and evaluates different points (x, y) for each function.

Table 4.6 contains the norm of the gradient of the resistance obtained by `integral2` and the 2d Simpson's rule and their execution time in different cases. The body has $2mn = 180$ linear functions and $\lambda = -100$. For `integral2`, we prescribed the desired precision ϵ whereas for the 2d Simpson's rule, we prescribed the number of points evaluated. We chose the desired precision and the number of points such that the results are illustrative and comparable.

Table 4.6 Performances of `integral2` and the 2d Simpson's rule to compute the gradient of the resistance $\nabla\mathcal{F}$ with $2mn = 180$ linear functions and $\lambda = -100$.

`integral2`			2d Simpson's rule		
ϵ	$\|\|\nabla\mathcal{F}\|\|$	Time	Points	$\|\|\nabla\mathcal{F}\|\|$	Time
10^{-3}	**0.553** 245 129 099	1.5s	3.4 10^{-4}	**0.5**61 576 970 145	0.3s
10^{-6}	**0.55**5 826 766 621	6.0s	1.4 10^{-7}	**0.556** 212 589 565	1.6s
10^{-9}	**0.556 063** 810 190	15.7s	2.3 10^{-13}	**0.556 06**4 894 067	5.4s
10^{-12}	**0.556 063 9**76 527	36.2s	2.5 10^{-14}	**0.556 063 94**7 716	13.9s
10^{-15}	**0.556 063 94**2 756	110.0s	2.5 10^{-14}	**0.556 063 94**3 248	25.3s

We observe that the 2d Simpson's rule is faster. The reason is that there are $2mn$ integrals to compute to obtain the vector $\nabla\mathcal{F}$. However, `integral2` does not allow to integrate a $2mn-$dimensional function, thus, we must call `integral2` $2mn$ times. As the $2mn$ integrals have a lot of common terms in (3.69), we miss a great opportunity.

We do not miss this opportunity with the 2d Simpson's rule. Indeed, we compute the integrals thanks to the same points (x, y), therefore, the common terms can be computed once for all and not $2mn$ times. It explains the huge difference of performance between `integral2` and the 2d Simpson's rule in Table 4.6. Note that increasing the number of linear functions increases the difference.

4.4 Numerical stability

We propose an improvement of the numerical evaluation of the LogSumExp and related expressions. It is a known numerical phenomenon that the LogSumExp can suffer of imprecision due to the sum of potentially large exponential. The improvement applies at different part of our problem.

We use the following identity

$$\sum_{i=1}^{n} \exp\big(\lambda(a_i x+b_i y+c_i)\big) = \sum_{i=1}^{n} \exp\big(\lambda(a_i x+b_i y+c_i-\sigma)\big)\exp(\lambda\sigma) = \exp\big(\lambda\sigma\big)\sum_{i=1}^{n} \exp\big(\lambda(a_i x+b_i y+c_i-\sigma\big) \tag{4.18}$$

where $\sigma = \min\limits_{i=1,\ldots,n}(a_i x + b_i y + c_i)$. Recall that $\lambda < 0$.

Evaluation of the function u_λ Thanks to the identity (4.18), instead of evaluating the function u_λ with

$$u_\lambda(x,y) = \frac{1}{\lambda}\log\left(\sum_{i=1}^{n} \exp\big(\lambda(a_i x + b_i y + c_i)\big)\right) \tag{4.19}$$

it is equivalent and more accurate to use

$$u_\lambda(x,y) = \sigma + \frac{1}{\lambda}\log\left(\sum_{i=1}^{n} \exp\big(\lambda(a_i x + b_i y + c_i - \sigma)\big)\right). \tag{4.20}$$

Without the usage of σ, MATLAB can evaluate u_λ up to $\lambda = -10^4$, above it returns $u_\lambda(x,y) = \infty$.

Note that in practice, evaluating u_λ is only useful to represent graphically the body as the values of u_λ are not used to compute or optimize the resistance, only the values of the gradient ∇u_λ are.

Integrand of the resistance $\mathcal{F}(u_\lambda)$ The expression (3.8) of the integrand of the resistance $\mathcal{F}(u_\lambda)$ is

$$\frac{\left(\sum\limits_{i=1}^{n} \exp\big(\lambda(a_i x + b_i y + c_i)\big)\right)^2}{\left(\sum\limits_{i=1}^{n} \exp\big(\lambda(a_i x + b_i y + c_i)\big)\right)^2 + \left(\sum\limits_{i=1}^{n} a_i\exp\big(\lambda(a_i x + b_i y + c_i)\big)\right)^2 + \left(\sum\limits_{i=1}^{n} b_i\exp\big(\lambda(a_i x + b_i y + c_i)\big)\right)^2}. \tag{4.21}$$

Using the identity (4.18) in this expression (4.21), the $\exp^2(\lambda\sigma)$ simplify in the numerator and denominator and yields

$$\frac{\left(\sum\limits_{i=1}^{n} \exp\big(\lambda(a_i x + b_i y + c_i - \sigma)\big)\right)^2}{\left(\sum\limits_{i=1}^{n} \exp\big(\lambda(a_i x + b_i y + c_i - \sigma)\big)\right)^2 + \left(\sum\limits_{i=1}^{n} a_i\exp\big(\lambda(a_i x + b_i y + c_i - \sigma)\big)\right)^2 + \left(\sum\limits_{i=1}^{n} b_i\exp\big(\lambda(a_i x + b_i y + c_i - \sigma)\big)\right)^2}. \tag{4.22}$$

Again, without σ we cannot use λ lower than -10^4, whereas with σ, we can use any value of λ.

Note that unlike the evaluation of u, improving the evaluation of the integrand of the resistance $\mathcal{F}$ is very important. Without σ, it is impossible to compute $\mathcal{F}$ for value $\lambda < -10^4$ since we could even not evaluate the integrand. Now, the limiting factor is the non-smoothness of the integrand rather than the numerical instability.

Integrand of the gradient of the resistance $\nabla\mathcal{F}$ The idea scales very well on the gradient of the integrand of the resistance $\nabla\mathcal{F}$. We simply replace all the $a_ix + b_iy + c_i$ of all the exponentials of the expression (3.75) by $a_ix + b_iy + c_i - \sigma$ and all the $\exp(\lambda\sigma)$ simplify.

4.5 Procedure

In this section, we provide details on the actual procedure used to find the optimal smooth minimum u_λ. We can distinguish two relatively independent approaches to perform the optimization.

Optimization by `fmincon`: We give the problem to the MATLAB function `fmincon`. It has the advantage to manage internally all the practical details: strategy to avoid local minimum, stopping criterion, step length and so on. Moreover, `fmincon` allows to solve a problem knowing its objective function only and do not necessarily need the gradient of the objective function. Finally, it can treat any constraint of the form $f_i(\mathbf{x}) \leq 0$ where $f_i(\mathbf{x})$ can be any non-linear or non-continuous function. However, if f_i is not sufficiently regular, it leads to bad results.

Optimization by projected gradient method: We apply a hand-made projected gradient method. Unlike `fmincon`, we have to manage the practical details. Also, at each step, we need to project the new solution on the set of admissible solutions. If this projection on the set $\{\mathbf{x} \mid f_i(\mathbf{x}) \leq 0 \quad \forall i\}$ cannot be done explicitly, it requires to solve the projection problem with `fmincon`. We will only use explicit projection since the reason to use a project gradient method is to avoid the usage of `fmincon`. Finally, by applying projected gradient iterations, we have a good control and understanding on what is happening.

Ultimately, we will combine iterations of one approach with iterations of the other. Now, we present how each approach is done in practice.

4.5.1 Solving the problem with `fmincon`

Solving an optimization problem with `fmincon` requires three inputs.

- An initial solution (a_0, b_0, c_0).
- The objective function $\mathcal{F}$ as a MATLAB function.
- The constraints as a MATLAB function.

We provide brief explanations on these three inputs.

Initial solution For the initial solution, we just randomly generate three vectors $a_0, b_0, c_0 \in \mathbb{R}^n$ and correct them to have a feasible body. An example of random body with $\lambda = -10$ is given in Fig. 4.7.

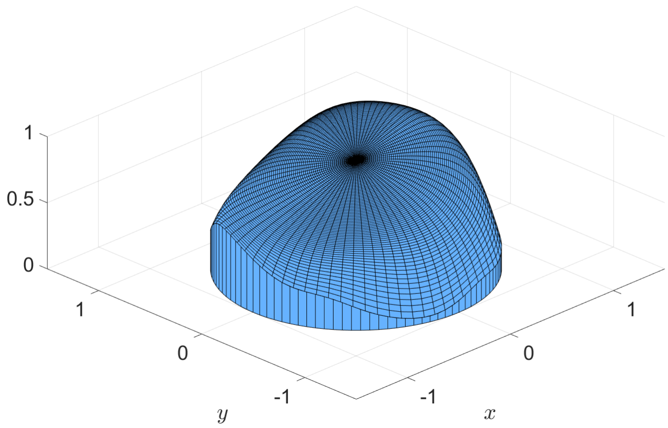


Fig. 4.7 Initial random body $u_{\lambda=-10}$

Objective function In practice, we implemented the function `integrand(a,b,c,r,theta)` that evaluates the integrand of the resistance $\mathcal{F}$ in the expression (3.8). We already presented this function in Sect. 4.3. Then, we integrate this function `integrand` thanks to `integral2` and we give it to `fmincon`.

Constraints As we briefly discussed, `fmincon` can receive any constraints of the form $f_i(x)$. Therefore, we have different possibilities coming from Sect. 3.3.2:

- Discretizing the constraints $0 \leq u_\lambda \leq M$ on the disk Ω. In particular, we use the constraints (3.27) and (3.34). For $u_\lambda \leq M$, we can even restrict ourself to just impose it on the point $(x, y) = (0, 0)$, namely, $u_\lambda(0, 0) \leq M$. It works but required to carefully choose the points of discretization where the constraints will be applied.

- Another possibility is to compute the maximal and minimal values of a given u_λ and to impose them to be, respectively, lower than M and greater than zero. However, it does not behave very well. We suspect that such constraints are not regular enough since it consists in solving optimization problems.

- The last possibility is to use the conditions extracted from the exact minimum case. These conditions were the following

$$c_i \geq \sqrt{a_i^2 + b_i^2} + \delta - k_1 \quad \forall i = 1, \ldots, n \tag{4.23}$$

$$a_1 = b_1 = 0 \text{ and } c_1 = M + k_2 \tag{4.24}$$

  where k_1 and k_2 are parameters to carefully select. Recall that these constraints ensured that the body obtained lies in $[-k_1, M + k_2]$.

4.5.2 Projected gradient

The second approach is to apply a projected gradient. Such method mainly requires five ingredients.

- An initial solution (a_0, b_0, c_0).

- The gradient of the objective function $\nabla\mathcal{F}$.
- A projector on the feasible set.
- A procedure to compute the length of the steps.
- A strategy to avoid local minimum.

We already discussed the three first points. Again, we provide details on these ingredients.

Initial solution The generation of the initial solution is identical to the initial solution described for the `fmincon` approach in Sect. 4.5.1.

Gradient of the objective function The gradient of the objective function $\nabla\mathcal{F}$ is the integral of the expression (3.75). We already discussed in Sect. 4.3 how this integral can be performed thanks to the 2d Simpson's rule.

Projection on the feasible set Since we only want to perform explicit projection, the only possibility is the constraint from the exact minimum case. Therefore, at each iteration, we project the parameter a_i, b_i and c_i thanks to the analytical solution of the problem (3.79). Recall, that the projected body is only guaranteed to lie into $[k_1, M + k_2]$.

Length of the step We use Armijo backtracking using the gradient mapping as criterion. We note the objective function $\mathcal{F}$ and the decision variables $\mathbf{x} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$.
More precisely, at each iteration of the projected gradient method, starting with a sufficiently large step size α, we decrease α until

$$\mathcal{F}(\mathbf{x}) - \mathcal{F}\big(\mathbf{x} - \alpha\nabla\mathcal{F}(\mathbf{x})\big) \geq \alpha t \tag{4.25}$$

where t is some fraction of the square of the norm of the gradient mapping $||G||^2$ defined as

$$G = \frac{\mathbf{x} - P\big(\mathbf{x} - \alpha\nabla\mathcal{F}(\mathbf{x})\big)}{\alpha}. \tag{4.26}$$

Avoiding local minimum In Chapter 2, we presented Observation 1 stating that the Newton's problem has local minimums. As we will see, we can indeed observe it on our approach. Therefore, we need a strategy to escape from a local minimum.

We simply propose to add a random perturbation on the variables a, b, c each 5 iterations of the projected gradient method.

Fig. 4.8 is the evolution of the resistance $\mathcal{F}$ during the projected gradient iterations with a perturbation each 5 iterations.

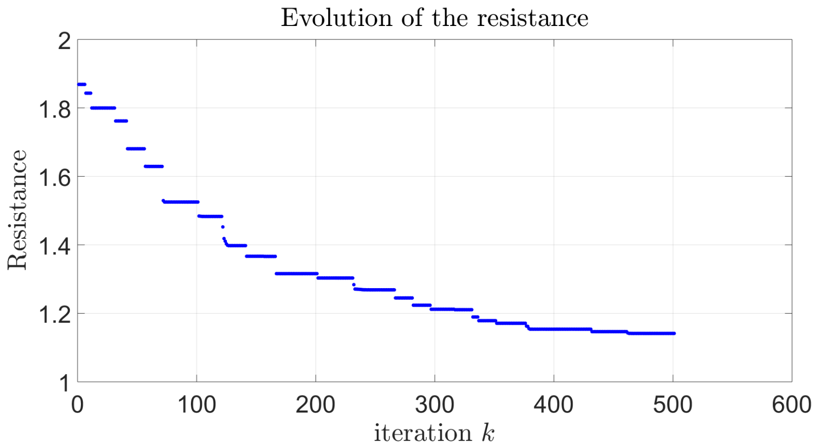


Fig. 4.8 Evolution of the resistance $\mathcal{F}$ during the projected gradient iterations with perturbations.

The figure exhibits that the resistance $\mathcal{F}$ quickly reaches a plateau and then the perturbations allow to reach a lower plateau and so on.

This phenomenon is consistent with the observation done in the literature. The polytope approach also used a gradient method and required the application of a genetic algorithm in order to avoid the local minimums. Whereas the exact minimum of linear functions used `fmincon` and did not mention any usage of randomness since this work is done by `fmincon`.

4.5.3 Cutting and rescaling

Recall that the only explicit constraints only allow to have a body lying in $[-k_1, M + k_2)]$ and that we need to apply the (CR) or (R) in order to have a body lying in $[0, M]$.

It appears that (CR) is better for bodies with symmetry parameter $m \geq 3$ whereas (R) is better for $m = 2$. More precisely, a body higher than M with symmetry $m \geq 3$ (resp. $m = 2$) will lose less (resp. more) optimality with (CR) than with (R). Indeed, the bodies with symmetry $m \geq 3$ have a top flat part. Therefore, the «cutting» of (CR) increases the optimality. Conversely, the bodies with $m = 2$ do not have the flat part and then the cutting introduces a deteriorating flat part.

The (CR) and (R) scalings can either be applied at the end of the optimization or at each iteration of the optimization. However, as the CR involves an exact minimum, if we apply the CR at each iteration, then we are minimizing a function $\tilde{\mathcal{F}}$ involving the strict minimum. Therefore, we minimize a function $\tilde{\mathcal{F}}$ while using the gradient of another function $\mathcal{F}$.

4.5.4 Choice of the parameters

There are several parameters to set before the optimization. We distinguish two types of parameters The technical parameters as the number of iterations, `OptimalityTolerance` and so on of `fmincon`, the maximal number of evaluated points of `integral2` and the 2d Simpson's rule, the precision of `integral2`, the frequency and amplitude of the perturbations to avoid the local minimums during the projected gradient method and many others. We do not comment the choice of these parameters.

The second type of parameters are those directly linked to the modelization of the body. More precisely, there is the following list of parameters.

- The LogSumExp parameter λ: as already discussed, it is not totally clear how to choose it. A good choice seems to be a large value but sufficiently small to perform rapid optimization.
- The number of linear functions n: A larger number of linear function certainly cover a strictly larger set of possible body u_λ. However, increasing n also increases the execution time of the simulation. As an indication, G. Wachsmuth used $2mn \approx 200$ linear functions.
- The symmetry parameter m: Conjecture 1 suggests that the parameter m depends on the height M. We can base our choice on the sequence M_m obtained for the best approach of the literature. The sequence was reported in Table 2.7.
- The bounds k_1 and k_2: These parameters indicate that the relaxed constraints will provide a body lying into $[-k_1, M + k_2]$. But, the choice $k_1 = k_2 = 0$ yield a body strictly lower that M. The latter situation is not necessarily an issue as (CR) and (R) increases the optimality of a body lower than M. The problem is that arbitrary values of k_1 and k_2 are possible. Therefore, we empirically tested different values.

4.5.5 Refinement strategy

Until now, we statically set the number of linear functions n. However, it may be interesting to dynamically remove or add specific linear functions during the optimization. This idea was done in the polytope and exact minimum of linear functions approaches [9, 15].

In our case, it occurs that the optimization naturally excludes some linear function i by giving large values for a_i, b_i and particularly $c_i \geq 0$. Indeed, in the LogSumExp function (4.27) with $\lambda < 0$

$$u_\lambda(x, y) = \frac{1}{\lambda} \log \left(\sum_{i=1}^{n} \exp\Big(\lambda(a_i x + b_i y + c_i)\Big) \right), \tag{4.27}$$

if c_i is large for some function i, then $\exp\Big(\lambda(a_i x + b_i y + c_i)\Big) \approx 0 \ \ \forall (x, y) \in \Omega$ and we can remove this function i. More precisely, in practice, we remove all the functions i such that their lowest point on the disk Ω is lower than M, i.e., the functions i such that

$$c_i - \sqrt{a_i^2 + b_i^2} < M. \tag{4.28}$$

Concerning the additional functions, it is less straightforward. A working strategy is to select an existing function i and add a new function with the same parameters a_i, b_i and c_i but slightly perturbed. Moreover, we tried to add linear functions to obtain bodies that look like the best current bodies of the literature.

In order to obtain the best resistances, we both removed and added linear functions as described here.

Chapter 5

Numerical results

In this chapter, we present the numerical results obtained with the smooth minimum method. First, we recall and comment the different features used to obtain the best results, then we analyze the best resistances and bodies and finally we provide the potential limitations of the method.

5.1 Overview of the features

We propose to list and comment the different features used to perform the optimization.

Evaluation of the resistance $\mathcal{F}$ We saw in Sect. 4.3 that `integral2` is faster than the 2d Simpson's rule to evaluate the resistance $\mathcal{F}$. Therefore, we always use `integral2` to evaluate $\mathcal{F}$.

Evaluation of the gradient of the resistance $\nabla\mathcal{F}$ Whereas, we also saw in Sect. 4.3 that the 2d Simpson's rule is faster than `integral2` to evaluate the $2mn$ integrals of the gradient of the resistance $\nabla\mathcal{F}$. Therefore, we use the 2d Simpson's rule to evaluate $\nabla\mathcal{F}$.

Optimization The optimization alternates between iterations of `fmincon`, described in Sect. 4.5.1, and projected gradient iterations, described in Sect. 4.5.2. The former appears better to avoid the local minimums and find a good candidate whereas the later was useful to refine this good candidate.

Constraints For the constraints, we mainly use the explicit ones as all the others did not provide satisfactory results. These constraints were derived in Sect. 3.3.2 and the way to perform the projection in Sect. 3.5.

Cutting and rescaling Since these constraints yield bodies higher or lower than M, we are forced to use the «cutting and rescaling» (R) or (CR) described in Sect. 3.3.2.

Symmetry In order to reduce the execution time, we apply the symmetry m as explained in Sect. 4.1. Therefore, the resistance $\mathcal{F}$ and its gradient $\nabla\mathcal{F}$ are computed on one sector only.

Refinement Finally, we also apply the refinement, i.e., adding and removing linear functions, as explained in Sect. 4.5.5.

5.2 Resistances

We present the bodies with best resistances $\mathcal{F}$ obtained by the smooth minimum method for the different heights. Table 5.1 contains the best resistances obtained in addition to the parameters m, n, λ and (R) or (CR) of the associated bodies. They are not better than the best resistances of the literature, yet, they come very close. The parenthesis contain a «percentage of closeness» between the best resistance of the literature $\mathcal{F}_{lit}$ and the best resistance of our approach $\mathcal{F}_\lambda$ for a given height M, i.e.,

$$percentage\ of\ closeness = \left(1 - \frac{|\mathcal{F}_{lit} - \mathcal{F}_\lambda|}{\mathcal{F}_{lit}}\right) \times 100\%. \tag{5.1}$$

The best resistance $\mathcal{F}_\lambda$ were obtained by alterning iterations of `fmincon` and projected gradient method.

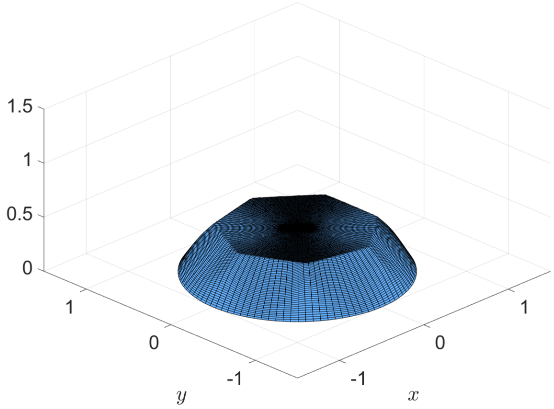
Table 5.1 Best resistances of the smooth minimum of linear functions.

Height M	$\mathcal{F}_\lambda$ (*closeness*)	Symmetry m	n	λ	R/CR	$\mathcal{F}_{lit}$
1.5	0.707 5 (98.9%)	2	15	-50	R	0.699 9
1.0	1.142 3 (99.6%)	3	7	-100	CR	1.137 8
0.7	1.550 0 (99.7%)	4	13	-110	CR	1.545 5
0.4	2.101 7 (99.9%)	7	7	-110	CR	2.099 6

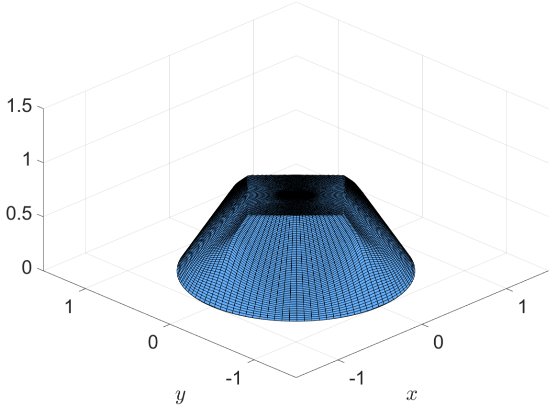
The results $\mathcal{F}_\lambda$ come close to the best results of the literature $\mathcal{F}_{lit}$ but do not surpass them. In particular, we observe that the «percentage of closeness» increases when the height M decreases. We suspect that is due to the computational time saved by the symmetry. Indeed, as the resistance $\mathcal{F}$ is computed on only one of the $2m$ sectors of Ω, increasing m decreases the execution time.

Moreover, note that in each cases, the optimization started with $n = 30$ linear functions but some of them became useless and got removed. Therefore, we end up with a very low number of effective linear functions $2mn$. For instances, the exact minimum of linear functions of G. Wachsmuth had around $2mn \approx 200$ where we have $2mn \approx 50$. We did not manage to suitably add new linear functions.

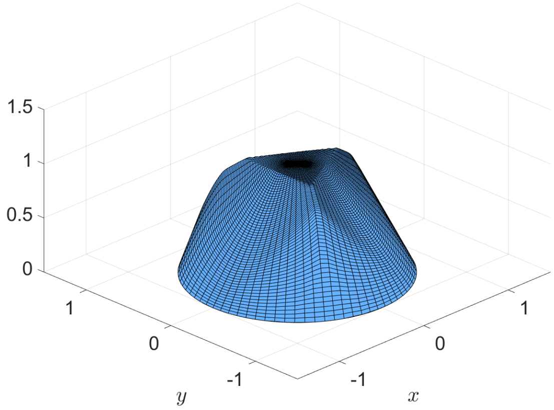
Figs. 5.1 are the illustrations of the bodies u_λ reaching the resistances in Table 5.1.



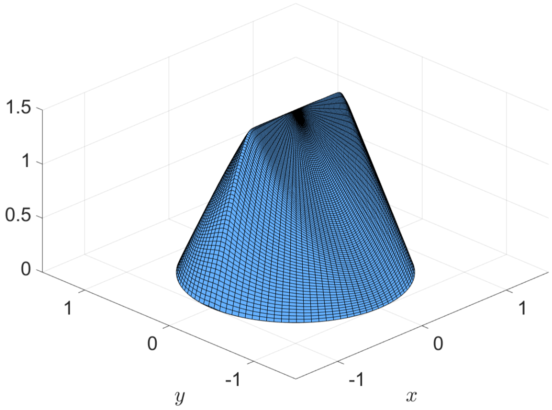
(a) $M = 0.4$.



(b) $M = 0.7$.



(c) $M = 1.0$.



(d) $M = 1.5$.

Fig. 5.1 Best smooth minimum bodies for different heights M.

One of the motivation of the usage of the smooth minimum method is that the exact minimum seemed to approximate a smooth function. In particular, there were two areas where we observe this «approximation». Firstly, on the transition between a lateral face and the top face and, secondly, on the transition between two lateral faces.

Fig. 5.2 illustrates the transitions between a lateral face and the top face or another lateral face of the best body u_λ for $M = 1$.

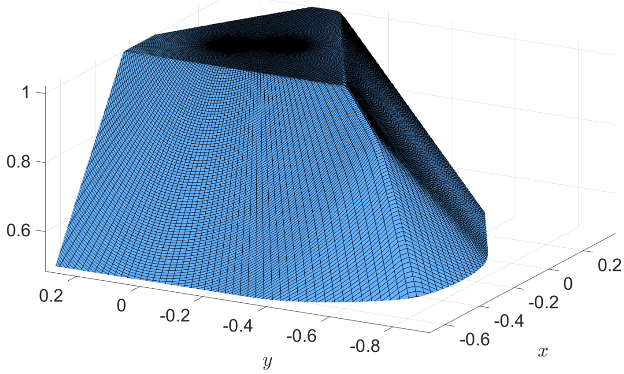


Fig. 5.2 Areas where the exact minimum seemed to approximate a smooth function.

The transition between a lateral face and the top face is abrupt and not smooth. Despite the refinement, we were not able to make the desired behavior appears. Conversely, the transition between two lateral faces is smoothed compared to the exact minimum case in Fig. 3.4.

Finally, we propose to update Fig. 2.20 summarizing the best resistances of the different approaches through the years.

Fig. 5.3 is the evolution of the best resistances (blue dots) and the best resistance of the smooth minimum method (green line) for $M = 1$.

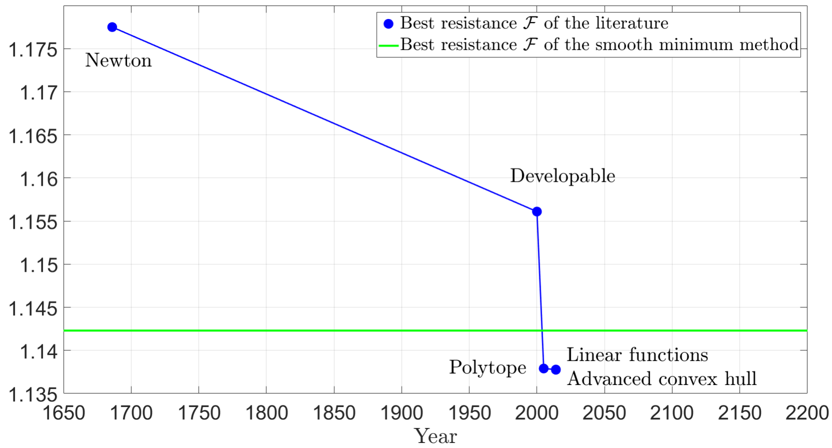


Fig. 5.3 Evolution of the best resistance through the years (blue dots) and the best resistance of the smooth minimum method (green line) for $M = 1$.

5.3 Limitations

The smooth minimum method provided good results but does not surpass the results of the polytopes, minimum of linear functions and advanced convex hull approaches. The reasons are multiple therefore we try to enumerate the possible improvements.

- The execution time of the optimization takes a considerable time. For instances, each solution in Fig. 5.1 took around 10 minutes. Therefore, it is tedious to perform an analysis of the good technical parameters to set. Note that it is the integral to compute the resistance $\mathcal{F}$ and its gradient $\nabla\mathcal{F}$ that take 99% of the execution time.

- Clearly, guaranteeing that the body u_λ has a maximal height M is a complicated task. We proposed several possibilities, however, none of them are perfect. They are either approximate or correspond to constraints not regular enough to be taken exactly into account while performing an optimization.

- The choice of the parameter λ is not clear. We empirically selected the best one.

- The cutting in (CR) is not taken into account in the computation of the gradient.

- Concerning the gradient without cutting, the option chosen was to compute analytically the integrand of the gradient of the resistance $\nabla\mathcal{F}$ and to numerically integrate it. Another option would be to numerically integrate the resistance with the 2d Simpson's rule, therefore, obtaining an explicit expression of the resistance $\mathcal{F}$. And then, analytically differentiate this explicit expression with respect to the variables a_i, b_i and c_i to obtain the gradient of the resistance.

- The refinement can be improved. Indeed, we did not manage to have the expected result in Fig 5.2, as we believe that we should more add effective linear functions between the lateral faces and the top face.

Chapter 6

Conclusion

Newton's problem aims to find the shape of a body minimizing the resistance of a fluid on this body. It is physically realistic in the case of a body moving at low speed through a rare gas. Mathematically, it is formalized as an optimization problem where the decision variable is the concave function u or the convex set U. It appears that there exist several problems formulated as the optimization of some function of a convex set U. Therefore, a method designed to solve the Newton's problem should also be designed to solve these other problems.

After formally presenting the Newton's problem, we looked at some known theoretical results about it. In particular, the optimal solution u^* is locally developable, i.e., locally non-strictly concave, and it is conjectured that the optimal solution has the symmetry m, i.e., the same type of symmetry that a polygon of m sides. Note that the radially symmetric solution proposed by Newton is not locally developable and has $m = \infty$ whereas the conjecture rather expects a finite m for height $M > 0$.

We analyzed four approaches of the literature: developable bodies [18], polytopes [9], minimum of linear functions [15] and advanced convex hull [15]. They are all locally developable and benefit from the symmetry m. It occurs that the advanced convex hull is the best approach for height $M < 1.4$ and that the minimum of linear functions is the best for $M > 1.5$.

The resistance of a body u is given by the integral of a function of its gradient ∇u. Therefore, the minimum of linear functions method requires to compute the gradient of u. It is necessary to compute the geometry of the body u, i.e., its vertex, faces and so on, in order to know which linear function is dominant. It also demands to anticipate the topological change of the body.

In light of this, we proposed to modelize the body u as a smooth minimum of linear functions thanks to the LogSumExp function. It kept the simple idea of minimum of linear functions and avoid the geometric computations as the integral can be performed in once. It occurs that this method actually suffers from practical drawbacks. More precisely, it is complicated to impose the maximal height M and the parameter λ of the LogSumExp is not clear to choose.

Nevertheless, the method provides good results. The resistances obtained by the smooth minimum method are approximately 98% as good as the best resistances of the literature.

There are several possible improvements as discussed in Sect. 5.3. Firstly, the refinement strategy, i.e., the way to add linear functions during the optimization, should be better. Indeed, comparing graphically our solutions to the best solutions of the literature shows that our solutions lack of linear functions in some area. Secondly, we used the gradient without «cutting» to optimize the resistance of a body with «cutting». It would be better to integrate the effect of the cutting directly into the gradient. Finally, it may be possible to obtain better results by investigating even more deeply the choice of the parameters k_1, k_2 and λ.

Nomenclature

Bodies

$\mathcal{S}_d$ Set of developable sets U_d

$\mathcal{C}_d$ Set of developable functions u_d

$\mathcal{S}$ Set of convex sets U

$\mathcal{C}$ Set of concave functions u

$f(t) = \frac{1}{1+t^2}$ Integrand of the resistance of the body

P Body as polytope (set)

U Set that describes the body

u Function that describes the body

u_ℓ Body as exact minimum (function)

u_λ Body as smooth minimum (function)

U_d Body as developable set

u_d Body as developable function

u_p Body as polytope (function)

$U_{g,m}$ Body as advanced convex hull (set)

$u_{g,m}$ Body as advanced convex hull (function)

Constant

$\delta = \frac{\log(1/n)}{\lambda} > 0$ Constant of the LogSumExp

$\sigma = \min_i a_i x + b_i y + c_i$ Constant of numerical stability

Parameters

$\lambda < 0$ Parameter of the LogSumExp function

$k_1 \leq \delta$ Parameter of the constraint to ensure $u_\lambda \geq -k_1$

$k_2 \leq \delta$ Parameter of the constraint to ensure $u_\lambda \leq M + k_2$

Bibliography

[1] Isaac Newton. *Philosophiae naturalis principia mathematica*, volume 2. typis A. et JM Duncan, 1833.

[2] Angelo Miele. On the theory of optimum aerodynamic shapes. Technical report, RICE UNIV HOUSTON TX AERO-ASTRONAUTICS GROUP, 1968.

[3] Giuseppe Buttazzo. A survey on the newton problem of optimal profiles. In *Variational analysis and aerospace engineering*, pages 33–48. Springer, 2009.

[4] Giuseppe Buttazzo, Vincenzo Ferone, and Bernhard Kawohl. Minimum problems over sets of concave functions and related questions. *Mathematische Nachrichten*, 173(1):71–89, 1995.

[5] Myriam Comte and Thomas Lachand-Robert. Newton's problem of the body of minimal resistance under a single-impact assumption. *Calculus of Variations and Partial Differential Equations*, 12(2):173–211, 2001.

[6] Myriam Comte and Thomas Lachand-Robert. Existence of minimizers for newton's problem of the body of minimal resistance under a single impact assumption. *Journal d'Analyse Mathématique*, 83(1):313–335, 2001.

[7] Alexander Plakhov. Newton's problem of minimal resistance under the single-impact assumption. *Nonlinearity*, 29(2):465, 2016.

[8] Herbert Busemann and Clinton M Petty. Problems on convex bodies. *Mathematica Scandinavica*, pages 88–94, 1956.

[9] Thomas Lachand-Robert and Édouard Oudet. Minimizing within convex bodies using a convex hull method. *SIAM Journal on Optimization*, 16(2):368–379, 2005.

[10] Jeff Cheeger. A lower bound for the smallest eigenvalue of the laplacian. In *Problems in analysis*, pages 195–200. Princeton University Press, 2015.

[11] Enea Parini. An introduction to the cheeger problem. *Surv. Math. Appl.*, 6:9–21, 2011.

[12] Wikipedia page «caccioppoli set». https://en.wikipedia.org/wiki/Caccioppoli_set#Caccioppoli_definition, Consulted: 21-04-21.

[13] François Alter and Vicent Caselles. Uniqueness of the cheeger set of a convex body. *Nonlinear Analysis: Theory, Methods & Applications*, 70(1):32–44, 2009.

[14] Lorenzo Brasco, Erik Lindgren, and Enea Parini. The fractional cheeger problem. *Interfaces and Free Boundaries*, 16(3):419–458, 2014.

[15] Gerd Wachsmuth. The numerical solution of newton's problem of least resistance. *Mathematical Programming*, 147(1-2):331–350, 2014.

[16] Lev V Lokutsievskiy and Mikhail I Zelikin. Hessian measures in the aerodynamic newton problem. *Journal of Dynamical and Control Systems*, 24(3):475–495, 2018.

[17] Thomas Lachand-Robert and Mark A Peletier. An example of non-convex minimization and an application to newton's problem of the body of least resistance. In *Annales de l'Institut Henri Poincare (C) Non Linear Analysis*, volume 18, pages 179–198. Elsevier, 2001.

[18] Thomas Lachand-Robert and Mark A Peletier. Newton's problem of the body of minimal resistance in the class of convex developable functions. *Mathematische Nachrichten*, 226(1):153–176, 2001.

[19] Matlab page «fmincon». https://nl.mathworks.com/help/optim/ug/fmincon.html, Consulted: 12-06-21.

[20] Lawrence F Shampine. Matlab program for quadrature in 2d. *Applied Mathematics and Computation*, 202(1):266–274, 2008.

[21] Whayne Padden. 2d simpson's integrator. https://www.mathworks.com/matlabcentral/fileexchange/23204-2d-simpson-s-integrator, 2021, Consulted: 08-06-21.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain
Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/epl**