CIMAT

Centro de Investigación en Matemáticas, A.C.

THE MONGE-AMPERE EQUATION AND THE NEWTON PROBLEM OF MINIMAL RESISTANCE

T H E S I S

As a requirement to obtain the degree of

Master of Science

With speciality in

Applied Mathematics

Presented by

Alejandro Méndez Rojas

Under the advice of

Dr. Miguel Angel Moreles Vázquez

Autorización de la versión final

Guanajuato, Gto., September 2020

Contents

Acknowledgements

I would like to express my sincere gratitute to my advisor Dr. Miguel Angel Moreles Vázquez for the continuous support and patience during this process.

Thanks to CONACYT for the schoolarship that I got during my two years at CIMAT.

Thanks to Dr. Renato Gabriel Iturriaga Acevedo and Dr. Pedro González Casanova Enriquez for accepting being my synodals and for their comments on this work.

Thanks to my classmates who worked with me during my first year at CIMAT.

Thanks for all the people who gived me their support during this work. Special thanks to those who helped me to grow as a person during this last year. In particular thanks to Paola Patricia García Lira for her guide, help and support, which helped me to continue with my dreams and have a better understanding of myself.

Introduction

The Monge-Ampère (M-A) equation is an important fully nonlinear elliptic equation. Its study is motivated from problems in different areas of knowledge. Research on the subject is vast, and we shall not list the extensive literature. We just refer the reader to [3] and references therein.

We study the M-A equation motivated by shape optimization. In particular, the Newton problem of minimal resistance. In 1685 Newton studied the problem of finding the shape of a body which moves in a fluid with minimal resistance to motion. A solution in his words (from Principia Mathematica):

If in a rare medium, consisting of equal particles freely disposed at equal distances from each other, a globe and a cylinder described on equal diameter move with equal velocities in the direction of the axis of the cylinder, (then) the resistance of the globe will be half as great as that of the cylinder. . . . I reckon that this proposition will be not without application in the building of ships.

The problem is still open and of great interest, see the survey in [2]. A more recent work is [6].

A first objective of this work, it to present the modern Variational Analysis version of the Newton problem. The underlying model is based, essentially, on the same assumptions that Newton made. It is shown that the problem corresponds to the minimization of a functional. Remarkably, in subdomains where the solution is smooth, the Monge Ampere equation must be satisfied as a necessary condition of optimality.

The M-A is a fully nonlinear elliptic equation, the well posedness of the Dirichlet Problem is a natural and nontrivial problem to be addressed. Although, the Newton problem leads to smooth solutions of the M-A equation, here we are content with establishing existence and uniqueness of general-

ized, as well as, viscosity solutions. We follow the presentation of [5].

With the Newton problem in perspective, we explore in the smooth case, the numerical solution of the M-A equation. In [1], a meshless approach is developed with polynomials a basis functions. Alternatively, we propose the classic Radial Basis Functions (RBF) interpolation. We develop a scheme using the Gaussian RBF as basis for approximation. Our results are preliminary but satisfactory. The potential will become apparent.

The outline of this multidisciplinary presentation of the M-A equation is as follows.

In Chapter 1, we present the modern analysis of the Newton Problem as in [3] or [5]. We revisit the problem in the jargon of fluid mechanics to derive the mathematical model. We are led to a variational problem, namely, the minimization of the functional of minimal resistance. The latter is evaluated on the half sphere and cylinder, obtaining Newton's conclusion.

Next, we introduce an admissible class of functions to study the minimal resistance functional. The variational analysis is carried out next. We show that the minimization problem is solvable, and smooth solutions on open subsets, necessarily satisfy the Monge-Ampere equation therein.

The well posedness of the Dirichlet problem for the Monge-Ampere equation is the content of Chapter 2.

In Chapter 3 we develop a numerical method to solve the M-A equation with Dirichlet conditions. We review the basics of RBF interpolation to develop a meshless scheme for approximation.

Conclusions and future work are described in the last chapter.

Chapter 1

The Newton problem of minimal resistance

In 1685 Newton studied the problem of finding the shape of a body which moves in a fluid with minimal resistance to motion. A solution in his words (from Principia Mathematica):

If in a rare medium, consisting of equal particles freely disposed at equal distances from each other, a globe and a cylinder described on equal diameter move with equal velocities in the direction of the axis of the cylinder, (then) the resistance of the globe will be half as great as that of the cylinder. . . . I reckon that this proposition will be not without application in the building of ships.

This chapter presents the modern analysis as in [3] or [5]. We revisit the problem in the jargon of fluid mechanics to derive the mathematical model. We are led to a variational problem, namely, the minimization of the functional of minimal resistance. The latter is evaluated on the half sphere and cylinder, obtaining Newton's conclusion.

Next, we introduce an admissible class of functions to study the minimal resistance functional. The variational analysis is carried out next. We show that the minimization problem is solvable, and smooth solutions on open subsets, necessarily satisfy the Monge-Ampere equation therein.

1.1 Mathematical Modeling

Let us start with an introduction to the Newton problem of aerodynamical profiles. Finding the profile of a body with minimal resistance (either aerodynamical or hydrodynamical) to movement is one of the first problems in variational calculus.

In 1685 Sir Isaac Newton studied this problem presenting a model to study the resistance of a profile to the motion in an inviscid and incompressible medium. The assumptions used by Newton to simplify his model are the following:

- the body is imbedded in a particle flow of uniform velocity v_0;
- the resistance comes as result of particles hitting the surface, and the particles hit with a perfect elastic collision;
- other effects like vorticity and turbulence are not considered.

We consider our surface as the graph of a function $u = u(x, y)$ defined in a domain $D \subset \mathbb{R}^2$.

Let us start deducing a functional J based on the previous hypothesis. This funcional should be such that minimizing J is equivalent to finding a surface generated by u that minimizes motion resistance.

In order to find J we start studying the impact of a particle with a surface. Consider a particle of an inviscid fluid hitting the surface with perfect elastic collision.

The momentum before the hit is given by

$$M_- = mv_0, \tag{1.1}$$

where v_0 is the velocity vector. Then, if ν is the normal vector to the surface at the hitting point, we have that

$$M_- = mv_0 = m((v_0 \cdot \nu)\nu + (v_0 \cdot \tau)\tau). \tag{1.2}$$

Here τ is the tangential component at the hitting point.

Perfect elastic collision means that the direction of the velocity changes, while the modulus of the velocity (the kinetic energy) is preserved. The direction of the velocity changes by preserving its tangential component and

reflecting its normal component. Consequently, the momentum after the hit is given by

$$M_+ = m(-(v_0 \cdot \nu)\nu + (v_0 \cdot \tau)\tau). \tag{1.3}$$

Then the difference is

$$\Delta M = -2m(v_0 \cdot \nu)\nu. \tag{1.4}$$

Also, by Newton's second law, the derivative of the momentum is the force

$$F(x,t) = \frac{d}{dt}M(x,t), \tag{1.5}$$

where

$$M(x,t) = \begin{cases} M_- & \text{if } t < t_0, \\ M_+ & \text{if } t > t_0. \end{cases} \tag{1.6}$$

Since the M is discontinuous on t_0, we take the derivative in the generalized sense.

So, if φ is a C_0^∞ function

$$\begin{aligned} \langle \dot{M}, \varphi \rangle &= -\int_{\mathbb{R}} M\varphi' \\ &= -M_- \int_{-\infty}^{t_0} \varphi' - M_+ \int_{t_0}^{\infty} \varphi' \\ &= -M_-\varphi(t_0) + M_+\varphi(t_0) \\ &= (M_+ - M_-)\varphi(t_0) \\ &= \langle (M_+ - M_-)\delta_{t_0}, \varphi \rangle. \end{aligned}$$

Consequently,

$$\dot{M} = (M_+ - M_-)\delta_{t_0} \equiv \Delta M \delta_{t_0}.$$

Let us consider $v_0 = (0, 0, -z_0)$ a velocity vector, $u = u(x,y)$ a surface in $\mathbb{R}^3$, with u defined on $D \subset \mathbb{R}^2$, and ν an orthogonal vector to u, given by

$$\nu = \frac{1}{\sqrt{(\frac{\partial u}{\partial x})^2 + (\frac{\partial u}{\partial y})^2 + 1}} \left(-\frac{\partial u}{\partial x}, -\frac{\partial u}{\partial y}, 1 \right).$$

Then, the force is

$$F = -2mv_0 \cdot \nu \delta_{t_0} \nu = \frac{2mz_0}{\sqrt{|\nabla u|^2 + 1}} \delta_{t_0} \nu. \tag{1.7}$$

The resistance is given by the third coordinate of F, that is

$$F_{res} = \frac{2mz_0}{|\nabla u|^2 + 1}. \tag{1.8}$$

The total resistance can be expressed as

$$2mz_0 \int_D \frac{1}{|\nabla u|^2 + 1} dxdy. \tag{1.9}$$

Hence, the functional to minimize is

$$J(u) = \int_D \frac{1}{|\nabla u|^2 + 1} dxdy. \tag{1.10}$$

Unless otherwise stated, J is always the Newton functional obtained above.

In the book *Principia Mathematica*, Newton claims that, under the mentioned hypothesis, if a globe and a cylinder described on equal diameter move with equal velocities in the direction of the axis of the cylinder, the resistance of the globe will be half as great as that of the cylinder.

Indeed, if D is the disc of radius R with center in the origin, then

$$u(x, y) = \varphi(r) = \sqrt{R^2 - r^2} \tag{1.11}$$

and

$$\frac{\partial u}{\partial x} = \varphi'(r)\frac{\partial r}{\partial x} = \varphi'(r)\frac{x}{r}. \tag{1.12}$$

Then

$$
\begin{aligned}
J(u) &= 2\pi \int_0^R \frac{1}{\left(\varphi'(r)\left(\left(\frac{x^2}{r^2}\right)+\left(\frac{y^2}{r^2}\right)\right)\right)^2+1} r dr \\
&= 2\pi \int_0^R \frac{r}{\varphi'(r)^2+1} \\
&= 2\pi \int_0^R \frac{r}{(\frac{-r}{\varphi(r)})^2+1} \\
&= 2\pi \int_0^R \frac{(R^2-r^2)r}{R^2} dr \\
&= 2\pi \left(\frac{r^2}{2} - \frac{1}{R^2}\frac{r^4}{4}\right)\Bigg|_0^R \\
&= \frac{\pi R^2}{2}.
\end{aligned}
$$

On the other hand, in the cylinder case, for $\tilde{u}(x,y)=1$ we have that

$$
\begin{aligned}
J(\tilde{u}) &= \int_D \frac{1}{1} dA \\
&= \pi R^2.
\end{aligned}
$$

Notheworthy, the result claimed by Newton is obtained.

1.2 The M-A equation as a constraint in the Newton Problem

This section relies heavily on Analysis. We use results from Real and Variational Analysis as in [4] and [3], [7]. Some of these results are listed in the appendix. We only provide some proofs to illustrate the flavour of the theory.

1.2.1 An admissible class of functions

We are led to study the resistance functional. As customary, a class of admissible functions needs to be determined.

In the next examples we show that, the class of admissible functions contains concave and bounded functions. Otherwise, solutions for the Newton problem need not exist.

First let us see that boundedness is a necessary condition.

Example 1 Let us consider the sequence $\{u_n\}$ defined as

$$u_n(x) = nd(x, \partial D),$$

where $d(x, \partial D)$ is the distance function.

Letting $n \to \infty$ we have that $J(u_n) \to 0$. But $J(u) > 0$ for every function u, which means that it is not possible to find an optimal function.

Now let us see that concaveness is a necessary condition.

Example 2 Lets consider a sequence of functions $\{u_n\}$ with $0 \leq u_n \leq M$, defined as

$$u_n(x) = M \sin^2(n|x|^2).$$

Then

$$\nabla u_n(x) = 4Mn \cos(n|x|^2) \sin(n|x|^2) \cdot x.$$

Notice that the set

$$S = \{x : \cos(n|x|^2) = 0 \text{ or } \sin(n|x|^2) = 0 \text{ for some } n \in \mathbb{N}\}$$

has measure zero. Then

$$\begin{aligned} J(u_n) &= \int_D \frac{1}{1 + |\nabla u_n(x)|^2} dx \\ &= \int_{D \setminus S} \frac{1}{1 + |\nabla u_n(x)|^2} dx + \int_S \frac{1}{1 + |\nabla u_n(x)|^2} dx \\ &= \int_{D \setminus S} \frac{1}{1 + |\nabla u_n(x)|^2} dx, \end{aligned}$$

and then $J(u_n) \to 0$ when $n \to \infty$.

Lets choose then a set of admissible functions where we can guarantee the existence of solutions for the Newton problem.

Definition 3 For every $M > 0$, we define C_M the set

$$\{u\colon u \text{ is concave in } D \text{ and } 0 \leq u \leq M\}.$$

Let us consider the general case of cost functionals of the form

$$F(u) = \int_D f(x, u, \nabla u)dx,$$

where D is a given subset of $\mathbb{R}^N$, and the integrand f satisfies the next hypothesis:

H1 the function $f : D \times \mathbb{R} \times \mathbb{R}^N \to \overline{\mathbb{R}}$ is nonnegative and measurable in the σ-algebra $\mathcal{L}_{\mathcal{N}} \otimes \mathcal{B} \otimes \mathcal{B}_{\mathcal{N}}$;

H2 for all $x \in D$ the function $f(x, \cdot, \cdot)$ is lower-semicontinuous in $\mathbb{R} \times \mathbb{R}^N$.

The resistance Newton functional is described by the integrand

$$f(z) = \frac{1}{1 + |z|^2}.$$

The set of admissible functions is C_M, for a given $M > 0$. Then, the minimum problem to consider is

$$\min\{F(u) : u \in C_M\}.$$

1.2.2 The Newton functional and the M-A equation

As the functions on C_M are continuous and bounded, they are locally Lipschitz, and then measurable.

Lets first remember the definition of Sobolev spaces.

Definition 4 The Sobolev space

$$W^{k,p}(U)$$

consists of all summable functions $u : U \to \mathbb{R}$ such that for each multi-index α with $|\alpha| \leq k$, $D^\alpha u$ exist in the weak sense and belongs to $L^P(U)$. We say

that $u \in W^{k,p}_{loc}(U)$ if for every $V \subset\subset U$, $u \in W^{k,p}(V)$.

We need the next lemma to show that our minimization problem is solvable.

Lema 5 For every $M > 0$ and $p < \infty$, the class C_M is compact respect to the strong topology of $W^{1,p}_{loc}(D)$.

The next theorem says that our minimization problem has a solution on C_M.

Theorem 6 Under the hypothesis **H1** and **H2**, for every $M > 0$, the minimum problem

$$\min\{F(u) : u \in C_M\}$$

has at least one solution.

In particular, the Newton problem has a solution for every $M > 0$.

If u is a solution of the previous minimum problem and, in an open set ω we have that:

1. u is of class C^2;
2. u does not attain the maximal value M;
3. u is strictly concave in the sense that its Hessian matrix is negative definite,

then, it can be proven that the second variation is given by,

$$\delta^2 J(u, \phi, \phi) = \int_\omega \frac{2}{(1 + |\nabla u|^2)^3}(4(\nabla u \nabla \phi)^2 - (1 + |\nabla u|^2)|\nabla \phi|^2)dx \geq 0. \quad (1.13)$$

For the following two theorems, we included a brief appendix on variational calculus. In particular, using the notation from the appendix, $U = C_M$ and $V = W^{1,p}_{loc}(D)$.

From the expression (1.13) we get the following theorem.

Theorem 7 Let D be a disc. Then, a solution of the Newton problem

$$\min\left\{\int_D \frac{1}{1+|\nabla u|^2}dx : u \in C_M\right\}$$

cannot be radial.

Proof Let us proceed by contradiction and assume that u is a radial solution. It can be proven that, outside of a circle of radius r_0 where $u \equiv M$, the function u is smooth, strictly concave and does not attain its maximum M. Then, we can use (1.13). We can take a function ϕ such that $\phi(r,\theta) = \eta(r)\sin(k\theta)$ for some $k \in \mathbb{N}$ with $supp(\eta) \subset (r_0, R)$ where R is the radius of D. Then, from (1.13) and using the change of variable theorem, and since u is a radial function

$$\begin{aligned}
0 \leq & \int_{r_0}^{R} \frac{r}{(1+u_r^2)^3}\Bigg[-(1+u_r^2)\Bigg(\eta_r^2(r)\int_0^{2\pi}\sin^2(k\theta)d\theta \\
& + 2\eta_r(r)\frac{\eta(r)}{r}k\int_0^{2\pi}\sin(k\theta)\cos(k\theta)d\theta + \frac{k^2\eta^2(r)}{r^2}\int_0^{2\pi}\cos^2(k\theta)d\theta\Bigg) + 4u_r^2\phi_r^2\Bigg]dr \\
= & \int_{r_0}^{R} \frac{r}{(1+u_r^2)^3}\Bigg[-(1+u_r^2)\Bigg(\eta_r^2(r)\pi + \frac{k^2\eta^2(r)}{r^2}\pi\Bigg) + 4u_r^2\phi_r^2\Bigg]dr.
\end{aligned}$$

Then, for k big enough, the integrand becomes negative, which is a contradiction. ■

Lets notice that the function $u \pm \varepsilon\phi$ is in C_M for a suitable $\varepsilon < \varepsilon_0$. Indeed, lets define $K = supp(\eta)$. Since u is strictly concave on (r_0, R), then the eigenvalues of it's Hessian matrix are both negative. We will show that it is possible to find ε such that the eigenvalues of the Hessian of the function $v(x,y) := u(x,y) \pm \varepsilon\eta(r)\sin(k\theta)$ are negative too. First lets notice that

$$\begin{aligned}
\phi_x &= \eta_r(r)\sin(k\theta)\frac{x}{\sqrt{x^2+y^y}} - \eta(r)\cos(k\theta)k\frac{y}{x^2+y^2}, \\
\phi_y &= \eta_r(r)\sin(k\theta)\frac{y}{\sqrt{x^2+y^y}} + \eta(r)\cos(k\theta)k\frac{x}{x^2+y^2}.
\end{aligned}$$

Then,

$$v_{xx} = u_{xx} \pm \varepsilon\left(\eta_{rr}(r)\sin(k\theta)\frac{x^2}{x^2+y^2} - \eta(r)\sin(k\theta)k^2\frac{y^2}{(x^2+y^2)^2}\right),$$
$$v_{yy} = u_{yy} \pm \varepsilon\left(\eta_{rr}(r)\sin(k\theta)\frac{y^2}{x^2+y^2} + \eta(r)\sin(k\theta)k^2\frac{x^2}{(x^2+y^2)^2}\right).$$

The eigenvalues of the Hessian matrix are both negative if and only if

$$u_{xx} + u_{yy} < 0 \text{ and } u_{xx}u_{yy} > 0.$$

We have that

$$\begin{aligned} v_{xx}v_{yy} = u_{xx}u_{yy} &\pm \varepsilon u_{xx}\left(\eta_{rr}(r)\sin(k\theta)\frac{y^2}{x^2+y^2} + \eta(r)\sin(k\theta)k^2\frac{x^2}{(x^2+y^2)^2}\right) \\ &\pm \varepsilon u_{yy}\left(\eta_{rr}(r)\sin(k\theta)\frac{x^2}{x^2+y^2} - \eta(r)\sin(k\theta)k^2\frac{y^2}{(x^2+y^2)^2}\right) \\ &+ \varepsilon^2\left(\eta_{rr}^2(r)\sin^2(k\theta)\frac{x^2y^2}{(x^2+y^2)^2} - \eta^2(r)\sin^2(k\theta)k^4\frac{x^2y^2}{(x^2+y^2)^4}\right). \end{aligned}$$

We are looking for $\varepsilon_1 > 0$ such that

$$v_{xx}v_{yy} > 0$$

which happens if

$$u_{xx}u_{yy} > \varepsilon_1||u_{xx}||(||\eta_{rr}|| + ||\eta||k^2) + \varepsilon_1||u_{yy}||(||\eta_{rr}|| + ||\eta||k^2) + \varepsilon_1^2||\eta_{rr}||^2$$

for every (x, y) in $sopp(\phi)$. Since $sopp(\phi)$ is compact and $u_{xx}u_{yy} > 0$ on this set, then such ε_1 exists. On the other hand,

$$v_{xx} + v_{yy} = u_{xx} + u_{yy} \pm \varepsilon\left(\eta_{rr}(r)\sin(k\theta) + \eta(r)\sin(k\theta)k^2\frac{x^2-y^2}{(x^2+y^2)^2}\right).$$

We are looking for $\varepsilon_2 > 0$ such that

$$v_{xx} + v_{yy} < 0.$$

This happens if

$$-(u_{xx} + u_{yy}) > \varepsilon_2(||\eta_{rr}|| + ||\eta||k^2)$$

for every (x, y) in $sopp(\phi)$. Since $sopp(\phi)$ is compact and $u_{xx} + u_{yy} < 0$ on this set, then such ε_2 exists.

Let r_1, r_2 be such that $sopp(\eta) = [r_1, r_2]$. We know that $r_0 < r_1 < r_2 < R$. Lets define $M_1 = u(r_1)$ and $M_2 = u(r_2)$. First notice that u is a decreacing function of r. Then, for every $r \in [r_1, r_2]$ we can find $\varepsilon_r > 0$ such that

$$M_1 > u(r) \pm \varepsilon_r ||\eta|| > M_2.$$

Using compacity we can find $\varepsilon_3 > 0$ such that

$$M_1 > u(r) \pm \varepsilon_3 ||\eta|| > M_2.$$

for every $r \in [r_1, r_2]$. Finally we can define $\varepsilon_0 = \min\{\varepsilon_1, \varepsilon_2, \varepsilon_3\}$.

With the last theorem we can conclude that the Newton problem has no unique solution. This is because we can always take a solution, rotate it and get a new one.

The next theorem give us a necessary condition for a function u, in order to solve the Newton problem.

Theorem 8 Let D be a convex domain and u a solution of the Newton problem. Assume that in an open set ω the function u is of class C^2 and does not attain its maximum M. Then

$$detD^2u = 0 \text{ in } \omega. \tag{1.14}$$

Proof Let us proceed by contradiction. Fix $x_0 \in \omega$ and let a be a unit vector orthogonal to $\nabla u(x_0)$. Assume that the equation (1.14) is not satisfied. Then, D^2u is negative definite, so we can use the inequality (1.13) for every function ϕ with support in a small neighborhood of x_0. Take

$$\phi(x) = \eta(x) \sin(ka \cdot x),$$

where the support of η is in a small neighborhood of x_0. Then

$$\nabla\phi(x) = \sin(ka \cdot x)\nabla\eta(x) + \cos(ka \cdot x)\eta(x)ka.$$

Notice that

$$\begin{aligned}(\nabla u(x)\nabla\phi(x))^2 &= \sin^2(ka \cdot x)(\nabla\eta(x)\nabla u(x))^2 \\ &+ 2k \sin(ka \cdot x)\cos(ka \cdot x)\eta(x)(\nabla u(x)\nabla\eta(x))(\nabla u(x) \cdot a) \\ &+ k^2(\nabla u(x) \cdot a)^2 \cos^2(ka \cdot x)\eta^2(x)\end{aligned}$$

and

$$
\begin{aligned}
|\nabla\phi(x)|^2 = &\sin^2(ka\cdot x)|\nabla\eta(x)|^2 \\
&+ 2k\sin(ka\cdot x)\cos(ka\cdot x)\eta(x)(\nabla\eta(x)\cdot a) \\
&+ k^2|a|^2\cos^2(ka\cdot x)\eta^2(x).
\end{aligned}
$$

Then, for k large enough, using (1.13) we have

$$
\begin{aligned}
\delta^2 J(u,\phi,\phi) = &\int_\omega \frac{2\cos^2(ka\cdot x)\eta^2(x)}{(1+|\nabla u(x)|^2)^3}(4(a\cdot\nabla u(x))^2 - (1+|\nabla u(x)|^2))dx + o(\frac{1}{k}) \\
\geq &\, 0
\end{aligned}
$$

As $\cos^2(ka\cdot x)\in[0,1]$ for every x, then

$$
\begin{aligned}
\int_\omega \frac{2\eta^2(x)}{(1+|\nabla u(x)|^2)^3}(4(a\cdot\nabla u(x))^2 - (1+|\nabla u(x)|^2))dx + o(\frac{1}{k}) &\geq \delta^2 J(u,\phi,\phi) \\
&\geq 0.
\end{aligned}
$$

Making the support of η tend to x_0 then

$$\eta(x)(\nabla u(x)\cdot a)\to 0,$$

and so

$$4(a\cdot\nabla u(x))^2 - (1+|\nabla u(x)|^2) < 0$$

then

$$\delta^2 J(u,\phi,\phi) < 0,$$

which is a contradiction. ■

Chapter 2

The Dirichlet problem for the Monge-Ampere equation

In the previous chapter we saw that a required optimality condition is the existence of a subset $\omega \subset \Omega$ where the Monge-Ampère equation is satisfied. This motivates the study of this equation that we carry out in this chapter. Here we are content with establishing existence and uniqueness of generalized, as well as, viscosity solutions for the Dirichlet problem. Most of the theorems in this chapter can be found in [5].

2.1 Generalized solutions

Lets start with a couple of definitions.

Definition 9 Let $\Omega \subset \mathbb{R}^N$ be open and $u : \Omega \to \mathbb{R}$. Given x_0, a supporting hyperplane to the function u at the point $(x_0, u(x_0))$ is an affine function $l(x) = u(x_0) + p \cdot (x - x_0)$ such that $u(x) \geq l(x)$ for every $x \in \Omega$.

Definition 10 The normal mapping of u, or subdifferential is the set-valued function $\partial : \Omega \to \mathcal{P}(\mathbb{R}^n)$ defined by

$$\partial u(x_0) = \{p : u(x) \geq u(x_0) + p \cdot (x - x_0), \text{ for every } x \in \Omega\}.$$

Given $E \subset \Omega$, we define $\partial u(E) = \cup_{x \in E} \partial u(x)$.

Lets note that the set $\partial u(x_0)$ may be empty.

The normal mapping has the following properties.

Proposition 11 Let $S = \{x \in \Omega : \partial u(x) \neq \emptyset\}$.

(a) If $u \in C^1(\Omega)$ and $x \in S$ then $\partial u(x) = Du(x)$, in other words, when u is differentiable the normal mapping is the gradient.

(b) If $u \in C^2(\Omega)$ and $x \in S$, the Hessian of u is positive semi-definite, that is $D^2u(x) \geq 0$. This means that if $u \in C^2(\Omega)$, S is the set where the graph of u is concave up.

The following lemma is useful to proof some of the results in this chapter.

Lemma 12 If $\Omega \subset \mathbb{R}^N$ is open, $u \in C(\Omega)$ and $K \subset \Omega$ is compact, then $\partial u(K)$ is compact.

Proof Let $\{p_m\} \subset \partial u(K)$ a sequence. We claim that $\{p_m\}$ is bounded. For every m there exist $x_m \in K$ such that $p_m \in \partial u(x_m)$, that is

$$u(x) \geq u(x_m) + p_m \cdot (x - x_m)$$

for every $x \in \Omega$. Since K is compact, the set $K_\delta = \{x : dist(x, K) \leq \delta\}$ is compact, and it is contained in Ω for δ small. Without loss of generality we can assume that there exist $x_0 \in K$ such that $x_m \to x_0$. Then $x_m + \delta w \in K_\delta$ and

$$u(x_m + \delta w) \geq u(x_m) + \delta p_m \cdot w$$

for every $|w| = 1$ and m. If $p_m \neq 0$ and $w = p_m/|p_m|$, then

$$\max_{K_\delta} u(x) \geq \min_K u(x) + \delta |p_m|$$

for every m. Since u is locally bonded, the claim is proved. Consequently there exist p_0 and a sequence $\{p_{m_k}\}$ such that $p_{m_k} \to p_0$. We claim that $p_0 \in \partial u(K)$. We shall prove that $p_0 \in \partial u(x_0)$. We have

$$u(x) \geq u(x_{m_k}) + p_{m_k} \cdot (x - x_{m_k})$$

for every $x \in \Omega$, and since u is continuous, by letting $m \to \infty$ we obtain

$$u(x) \geq u(x_0) + p_0 \cdot (x - x_0)$$

for all $x \in \Omega$. ■

The following lemmas are used in the proofs of following results.

Lemma 13 If u is a convex function in Ω and $K \subset \Omega$ is compact, then u is uniformly Lipschitz on K, that is, there exists a constant $c = c(u, K)$ such that $|u(x) - u(y)| \leq c|x - y|$ for every $x, y \in K$.

Lemma 14 If Ω is open and u is Lipschitz in Ω, then u is differentiable a.e. in Ω.

Lemma 15 If u is concave up (concave down) in Ω, then u is differentiable in all Ω.

Definition 16 The Legendre transform of the function $u : \Omega \to \mathbb{R}$ is the function $u^* : \mathbb{R}^N \to \mathbb{R}$ defined as

$$u^*(p) = \sup_{x \in \Omega} \{x \cdot p - u(x)\}.$$

Remark If Ω is bounded and u is bounded in Ω then u^* is finite. Also, u^* is convex in $\mathbb{R}^N$.

The next lemma helps us to show some properties of the Monge-Ampère measure, which is introduced in the following theorem.

Lemma 17 If Ω is open and u is continuous in Ω, then the set of points in $\mathbb{R}^N$ that belong to the image by the normal mapping of more than one point of Ω has Lebesgue measure zero. That is, the set

$$S = \{p \in \mathbb{R}^N : \text{ there exist } x, y \in \Omega, x \neq y \text{ and } p \in \partial u(x) \cap \partial u(y)\}$$

has measure zero. This also means that the set of supporting hyperplanes that touch the graph of u at more than one point has measure zero.

In the following theorem we define the Monge-Ampère measure and show some of its properties.

Theorem 18 If Ω is open and $u \in C(\Omega)$, then the class

$$S = \{E \subset \Omega : \partial u(E) \text{ is Lebesgue measurable}\}$$

is a Borel σ-algebra. The set function $M_u : S \to \overline{\mathbb{R}}$ defined by

$$M_u(E) = |\partial u(E)|,$$

is a measure, finite on compacts, that is called the Monge-Ampére measure associated with the function u.

Proof By lemma 12, S contains every compact subset of Ω. By definition, for every $E \subset \Omega$,

$$\partial u(E) = \cup_{x \in E} \partial u(x).$$

Then, if $\{E_m\}$ is a sequence of subsets in Ω,

$$\partial u(\cup_m E_m) = \cup_m \cup_{x \in E_m} u(x) = \cup_m \partial u(E_m).$$

Hence, if $E_m \in S$ for $m = 1, 2, ...$, then $\cup_m E_m \in S$. In particular, we can write $\Omega = \cup_m K_m$ with K_m compact, and then $\Omega \in S$.To show that S is a σ-algebra, it remains to show that if $E \in S$ then $\Omega \backslash E \in S$. For every $E \subset \Omega$ we have that

$$\partial u(\Omega \backslash E) = (\partial u(\Omega) \backslash \partial u(E)) \cup (\partial u(\Omega \backslash E) \cap \partial u(E)). \tag{2.1}$$

By lemma 17, $|\partial u(\Omega \backslash E) \cap \partial u(E)| = 0$ for every E. Then, from (2.1), $\Omega \backslash E \in S$ if $E \in S$.
We now show that M_u is σ-additive. Let $\{E_i\}_{i=1}^{\infty}$ be a sequence of disjoint sets in S and set $\partial u(E_i) = H_i$. We must show that

$$|\partial u(\cup_{i=1}^{\infty} E_i)| = \sum_{i=1}^{\infty} |H_i|.$$

Since $\partial u(\cup_{i=1}^{\infty} E_i) = \cup_{i=1}^{\infty} H_i$, we shall show that

$$| \cup_{i=1}^{\infty} H_i| = \sum_{i=1}^{\infty} |H_i|. \tag{2.2}$$

We have $E_i \cap E_j$ for $i \neq j$. Then, by lemma 17 $|H_i \cap H_j| = 0$ for $i \neq j$. Let us write

$$\cup_{i=1}^{\infty} H_i = H_1 \cup (H_2 \backslash H_1) \cup (H_3 \backslash (H_1 \cup H_2)) \cup (H_4 \backslash (H_1 \cup H_2 \cup H_3)) \cup ...,$$

where the sets on the right-hand side are disjoint. Now

$$H_n = [H_n \cap (H_1 \cup H_2 \cup ... \cup H_n)] \cup [H_n \backslash (H_1 \cup H_2 \cup ... \cup H_n)].$$

Then by lemma 17 $|H_n \cap (H_1 \cup H_2 \cup ... \cup H_n)| = 0$ and we obtain

$$H_n = [H_n \backslash (H_1 \cup H_2 \cup ... \cup H_n)].$$

Consequently (2.2) follows, and the proof of the theorem is complete. ■

The next theorem is useful to proof an important property of the Monge-Ampère measure.

Sard's Theorem 19 Let $\Omega \subset \mathbb{R}^N$ be an open set and $g : \Omega \to \mathbb{R}^N$ a C^1 function in Ω. If $S_0 = \{x \in \Omega : \det g'(x) = 0\}$, then $|g(S_0)| = 0$.

Proposition 20 Let $u \in C^2(\Omega)$ be a convex function, then the Monge-Ampère measure M_u associated with u satisfies

$$M_u(E) = \int_E detD^2u(x)dx, \tag{2.3}$$

for every Borel set $E \subset \Omega$.

Proof First lets notice that since u is convex and $u \in C^2(\Omega)$, then Du is 1-1 in the set $A = \{x \in \Omega : D^2u(x) > 0\}$. Indeed, let $x_1, x_2 \in A$ with $Du(x_1) = Du(x_2)$. By convexity

$$u(z) \geq u(x_i) + Du(x_i) \cdot (z - x_i), \;\; z \in \Omega, i = 1, 2.$$

Hence

$$u(x_1) - u(x_2) = Du(x_1) \cdot (x_1 - x_2) = Du(x_2) \cdot (x_1 - x_2).$$

By the Taylor formula we can write

$$u(x_1) = u(x_2) + Du(x_2) \cdot (x_1 - x_2) + \int_0^1 t < D^2u(x_2 + t(x_1 - x_2))(x_1 - x_2), x_1 - x_2 > dt.$$

Therefore the integral is zero and the interand must vanish for $0 \leq t \leq 1$ because D^2u is positive definite and $t \geq 0$. Since $x_2 \in A$, we have that

$x_2 + t(x_1 - x_2) \in A$ for t small. Then $x_1 = x_2$. If $u \in C^2(\Omega)$, then $g = Du \in C^1(\Omega)$. We have $M_u(E) = |Du(E)|$ and

$$Du(E) = Du(E \cap S_0) \cup Du(E \backslash S_0).$$

Since $E \subset \mathbb{R}^n$ is a Borel set, $E \cap S_0$ and $E \backslash S_0$ are also Borel sets. Then, by the formula of change of variables and Sard's theorem,

$$M_u(E) = M_u(E \cap S_0) + M_u(E \backslash S_0) = \int_{E \backslash S_0} \det D^2 u(x) dx = \int_E \det D^2 u(x) dx$$

which shows (2.3). ■

Now we can introduce the notion of generalized solutions for the Monge-Ampère equation.

Definition 21 Let ν a Borel measure defined in Ω, an open and convex subset of $\mathbb{R}^n$. The convex function $u \in C(\Omega)$ is a generalized solution to the Monge-Ampère equation

$$det D^2 u = \nu$$

if the Monge-Ampère measure M_u associated with u equals ν.

Remark In the above definition, the equality $det D^2 u = \nu$ means that, for every Borel set $E \subset \Omega$, $M_u(E) = v(E)$.

Next we have a few properties of the normal mapping.

Lemma 22 Let $u_n \in C(\Omega)$ be convex functions such that $u_n \to u$ uniformly on compact subsets of Ω. Then,

(i) If $K \subset \Omega$ is compact, then

$$\overline{\lim}\, \partial u_n(K) \subset \partial u(K),$$

and by Fatou's lemma

$$\overline{\lim}\, |\partial u_n(K)| \leq |\partial u(K)|,$$

(ii) If $A \subset \Omega$ is open, then

$$\underline{\lim}\, \partial u_n(A) \subset \partial u(A),$$

and by Fatou's lemma

$$\underline{\lim}\, |\partial u_n(A)| \leq |\partial u(A)|.$$

Lemma 23 If u_n are convex functions in Ω such that $u_n \to u$ on compact subsets of Ω, then the Monge-Ampère measure M_{u_n} tend to M_u weakly, that is,

$$\int_\Omega f(x) dM_{u_n}(x) \to \int_\Omega f(x) dM_u(x)$$

for every f continuous with compact support in Ω.

2.1.1 Viscosity solutions

Generalized solutions are not the only kind of solutions that we can study. The viscosity solutions are defined as follows.

Definition Let $u \in C(\Omega)$ be convex and $f \in C(\Omega)$, $f \geq 0$. The function u is a viscosity subsolution (supersolution) of the equation $detD^2u = f$ in Ω if for every $\phi \in C^2(\Omega)$ convex and $x_0 \in \Omega$ such that

$$(u - \phi)(x) \leq (\geq)(u - \phi)(x_0)$$

for every x in a neighborhood of x_0, then we must have

$$\det D^2\phi(x_0) \geq (\leq) f(x_0).$$

The next proposition relates viscosity solutions with generalized solutions.

Proposition 24 If u is a generalized solution to $M_u = f$ with f continuous, then u is a viscosity solution.

2.2 Maximum principles

In order to prove uniqueness of solutions we want to get first a maximum principle.

Lemma 25 Let $\Omega \subset \mathbb{R}^N$ be open and bounded, $u, v \subset C(\overline{\Omega})$. If $u = v$ in $\partial\Omega$ and $v \geq u$ in Ω, then

$$\partial v(\Omega) \subset \partial u(\Omega).$$

Proof Let $p \in \partial v(\Omega)$. There exist $x_0 \in \Omega$ such that

$$v(x) \geq v(x_0 + p \cdot (x - x_0)),\ x \in \Omega.$$

Let

$$a = \sup_{c \in \Omega} \{v(x_0 + p \cdot (x - x_0)) - u(x)\}.$$

Since $v(x_0) \geq u(x_0)$, it follows that $a \geq 0$. We claim that $v(x_0)+p\cdot(x-x_0)-a$ is a supporting hyperplane to the function u at some point in Ω. Since Ω is bounded, there exist $x_1 \in \overline{\Omega}$ such that $a = v(x_0) + p \cdot (x_1 - x_0) - u(x_1)$ and then

$$u(x) \geq v(x_0 + p \cdot (x - x_0)) = u(x_1) + p \cdot (x - x_1),\ x \in \Omega.$$

We have

$$v(x_1) \geq v(x_0) + p \cdot (x_1 - x_0) = u(x_1) + a.$$

Then, if $a > 0$, $x_1 \notin \Omega$ and so the claim holds in this case. If $a = 0$ then

$$u(x) \geq v(x_0) + p \cdot (x - x_0) \geq u(x_0) + p \cdot (x - x_0)$$

and Consequently $u(x_0)+p\cdot(x-x_0)$ is a supporting hyperplane for u at x_0. ■

Theorem (Aleksandrov's maximum principle)26 If $\Omega \subset \mathbb{R}^N$ is open, bounded and convex with diameter Δ, and $u \in C(\overline{\Omega})$ is convex with $u = 0$ in $\partial\Omega$, then

$$|u(x_0)|^n \leq c_N \Delta^{N-1} d(x_0, \partial\Omega) |\partial u(\Omega)|,$$

for every $x_0 \in \Omega$ where c_N is a constant depending only on the dimension N.

2.3 Well-posedness of the Dirichlet Problem

In this section we proof a uniqueness theorem for generalized solutions of the Monge-Ampère equation.

Lets start with a few results about the subdifferential and the Monge-Ampère measure.

Theorem 27 Let $u, v \in C(\overline{\Omega})$, with v convex, such that

$$|\partial u(E)| \leq |\partial v(E)| \text{ for every Borel set } E \subset \Omega.$$

Then

$$\min_{x \in \Omega}\{u(x) - v(x)\} = \min_{x \in \partial\Omega}\{u(x) - v(x)\}.$$

Lemma 28 If v and ϕ are convex functions in Ω, then

$$M_{(v+\phi)}(E) \geq M_v(E) + M_\phi(E)$$

for each Borel set $E \subset \Omega$.

Corollary 29 If $u, v \in C(\overline{\Omega})$ are convex functions such that $|\partial u(E)| = |\partial v(E)|$ for every Borel set $E \subset \Omega$ and $u = v$ in $\partial\Omega$, then $u = v$ in Ω.

Definition 30 The open set $\Omega \subset \mathbb{R}^n$ is strictly convex if for all $x, y \in \overline{\Omega}$, the open segment joining x and y lies in Ω.

We can finally state a uniqueness theorem for the Monge-Ampère equation.

Theorem 31 Let $\Omega \subset \mathbb{R}^n$ be bounded and strictly convex, and $g : \partial\Omega \to \mathbb{R}$ a continuous function. There exist a unique convex function $u \in C(\overline{\Omega})$ generalized solution of the problem

$$\det D^2 u = 0 \text{ in } \Omega,$$
$$u = g \text{ in } \partial\Omega.$$

Proof Let $\mathcal{F} = \{a : a \text{ is an affine function and } a \leq g \text{ on } \partial\Omega\}$. As g is continuous, $\mathcal{F} \neq \emptyset$. Lets define

$$u(x) = \sup\{a(x) : a \in \mathcal{F}\}.$$

As affine functions are convex and u is the supremum of convex functions, u is also convex and $u(x) \leq g(x)$ for every $x \in \partial\Omega$.
The first step is to show that $u = g$ on $\partial\Omega$. Let $\xi \in \partial\Omega$; we show that $u(\xi) \geq g(\xi)$. Given $\varepsilon > 0$ there exist $\delta > 0$ such that $|g(x) - g(\xi)| < \varepsilon$ for $|x-\xi| < \delta$, $x \in \Omega$. Let $P(x) = 0$ be the equation of the supporting hyperplane to Ω at the point ξ, and assume that $\Omega \subset \{x : P(x) \geq 0\}$. Since Ω is strictly convex, there exist $\eta > 0$ such that $S = \{x \in \overline{\Omega} : P(x) \leq \eta\} \subset B_\delta(\xi)$. Let

$$M = \min\{g(x) : x \in \partial\Omega, P(x) \geq \eta\}$$

and consider

$$a(x) = g(\xi) - \varepsilon - AP(x)$$

where A is a constant such that

$$A \geq \max\left\{\frac{g(\xi) - \varepsilon - M}{\eta}, 0\right\}.$$

We have $a(\xi) = g(\xi) - \varepsilon - AP(\xi) = g(\xi) - \varepsilon$, and if $x \in \partial\Omega$ we claim that $a(x) \leq g(x)$. Indeed, if $x \in \partial\Omega \cap S$, then $g(\xi) - \varepsilon \leq g(x) \leq g(\xi) + \varepsilon$, so $g(x) \geq g(\xi) - \varepsilon - AP(x) + AP(x) \geq g(\xi) - \varepsilon - AP(x) = a(x)$. If $x \in \partial\Omega \cap S^c$, then $P(x) > \eta$ and by definition of M and A we have

$$\begin{aligned} g(x) &\geq M = a(x) + M - g(\xi) + \varepsilon + AP(x) \\ &\geq M = a(x) + M - g(\xi) + \varepsilon + A\eta \\ &\geq a. \end{aligned}$$

Therefore $a \in \mathcal{F}$, and in particular $u(\xi) \geq a(\xi) = g(\xi) - \varepsilon$ for every $\varepsilon > 0$ and therefore $u(\xi) \geq g(\xi)$.

The second step is to show that u is continuous in $\overline{\Omega}$. Since u is convex, in Ω, u is continuous in Ω. To proof the continuity on $\partial\Omega$, let $\xi \in \partial\Omega$, $\{x_n\} \subset \overline{\Omega}$ with $x_n \to \xi$. We show that $u(x_n) \to g(\xi)$. If a is the function constructed before, then $u(x) \geq a(x)$, in particular $u(x_n) \geq a(x_n)$ and then

$$\liminf u(x_n) \geq \liminf a(x_n) = \liminf(g(\xi) - \varepsilon - AP(x)) = g(\xi) - \varepsilon,$$

for all $\varepsilon > 0$. Then $\liminf u(x_n) \geq g(\xi)$. We now proof that $\limsup u(x_n) \leq g(\xi)$. Since Ω is convex, there exist h harmonic in Ω such that $h \in C(\overline{\Omega})$ and and $h\big|_{\partial\Omega} = g$. If a is any affine function so that $a \leq g$ on $\partial\Omega$, then a is

harmonic and by the maximum principle $a \leq h$ in Ω. By taking supremum over a we obtain $u(x) \leq h(x)$ for $x \in \Omega$. In particular, $u(x_n) \leq h(x_n)$ and therefore $\limsup u(x_n) \leq \limsup h(x_n) = g(\xi)$.

The third step is to proof that

$$\partial u(\Omega) \subset \{p \in \mathbb{R}^n : \text{ there exist } x, y \in \Omega, x \neq y \text{ and } p \in \partial u(x) \cap \partial u(y)\}, \tag{2.4}$$

and by lemma 17 $|\partial u(\Omega)| = 0$.

If $p \in \partial u(\Omega)$, then there exist $x_0 \in \Omega$ such that $u(x) \geq u(x_0) + p \cdot (x - x_0) = a(x)$ for all $x \in \Omega$. Since $u = g$ on $\partial\Omega$, we have $g(x) \geq a(x)$ for all $x \in \partial\Omega$. There exist $\xi \in \partial\Omega$ such that $g(\xi) = a(\xi)$. Otherwise, there exist some $\varepsilon > 0$ such that $g(x) > a(x) + \varepsilon$ for all $x \in \partial\Omega$ and then $u(x) \geq a(x) + \varepsilon$ for all $x \in \Omega$, and in particular $u(x_0) \geq a(x_0) + \varepsilon = u(x_0) + \varepsilon$, a contradiction. Since Ω is convex, the segment I joining x_0 and ξ is contained in Ω. Now $u(x_0) = a(x_0)$ and $u(\xi) = a(\xi)$. If $z \in I$, then $z = tx_0 + (1-t)\xi$ and by convexity,

$$u(z) \leq tu(x_0) + (1-t)u(\xi) = ta(x_0) + (1-t)a(\xi) = a(z).$$

But $u(x) \geq a(x)$ for all $x \in \Omega$ so a is a supporting hyperplane to u at any point of the segment I, therefore $p \in \partial u(z)$ for all $z \in I$ and (2.4) is then proved. Uniqueness follows from Corollary 29. ■

Chapter 3

Numerical solution of the M-A equation

In this chapter we develop a numerical method to solve the M-A equation with Dirichlet conditions. Research on this subject is very active. Since we seek smooth solutions as required in the Newton problem of minimal resistance, we follow the meshless approach in [1]. therein a polynomial basis is used. Alternatively, we explore the classic Radial Basis Functions (RBF) interpolation. We remark that this is a preliminary, albeit satisfactory, study.

3.1 Computational Model

Let us consider the problem

$$\begin{aligned} u_{xx}u_{uyy} - u_{xy}^2 &= g \text{ in } \Omega, \\ u &= f \text{ en } \partial\Omega, \end{aligned}$$

in a domain Ω.

The solution scheme is as follows:

1. Choose a trial space that should approximate the true solution u^* well,
2. Select sets of test points on which the differential operator and the boundary conditions are directly sampled,

3. Form a nonlinear system of collocation equations, possibly overdetermined,

4. Apply a nonlinear optimizer to minimize residuals of the system.

With this method, the consistency is guaranteed by choosing a sufficiently rich trial space. Stability requires choosing sufficiently many well-posed collocation points. It is also important to be able to easily compute up to second derivatives of the trial functions.

In order to solve the Dirichlet Problem, we define a functional (**Res**(c,x,y,n,f,g)) as

```
sum = 0
for i=0 to n do
    sum+ = (Ψ(x_i, y_i, c, x, y) − f(x_i, y_i))^2
end for
for i=n to x.shape do
    sum+ = (Ψ_xx(x_i, y_i, c, x, y) · Ψ_yy(x_i, y_i, c, x, y) − Ψ_xy(x_i, y_i, c, x, y)^2) −
    g(x_i, y_i))^2
end for
sum = sum/x.shape
sum = sqrt(sum)
return sum,
```

where n is the number of points in the boundary of the domain. With this process we get a non-negative functional, which minimal value is reached by solving the Monge-Ampère problem. The function **scipy.optimize.minimize** of Python is used to minimize the previous functional, and the method used is BFGS.

For the trial space, we consider a set of functions ψ_i for $i = 1, ..., n$ in $C^\infty(D)$ such that the function ψ_i is centered on (x_i, y_i). The objective is, given a function f defined on D, find a set of coefficients c_i for $i = 1, ..., n$ such that the function $\Psi := \sum c_i \psi_i$ approaches well enough to the function f. Our results are illustrated with trial spaces formed by Radial Basis Functions (RBF).

3.1.1 Radial Basis Functions interpolation

The ideal numerical method for PDE problems should be high-order accurate, flexible with respect to the geometry, computationally efficient, and easy to implement. The methods that are commonly used usually fulfill one or two of the criteria, but not all. Finite difference methods can be made high-order accurate, but require a structured grid (or a collection of structured grids). Spectral methods are even more accurate, but have severe restrictions on geometry. Finite element methods are highly flexible, but it is hard to achieve high-order accuracy, and both coding and mesh generation become increasingly difficult when the number of the space dimensions increases.

A fairly new approach to solving PDEs is through radial basis functions. An RBF depends only on the distance to a center point x_j and is of the form $\phi(||x - x_j||)$. The RBF may also have a shape parameter ε, in which case $\phi(r)$ is replaced with $\phi(r, \varepsilon)$.

In this section we review interpolation by radial basis functions, see [8]. Later on next section we display some results of solving the M-A equation by RBF.

Our Objective is to approximate a function f using a given set of points in a domain $\Omega \subset \mathbb{R}^n$, using a finite number of evaluations of f. More formaly, let $X \subset \Omega$ be the set of points $X = \{x_1, x_2, ..., x_N\}$ and let $\{y_1, y_2, ..., y_N\}$ be such that $f(x_i) = y_i$ for $i = 1, ..., N$. We look for a function $\Phi_{f,X}$ such that $\Phi_{f,X}(x_i) = y_i$, which will be an approximation for our unknown function f.

By a function $\phi : \mathbb{R}^n \to \mathbb{R}$, we form the interpolant

$$\Phi_{f,X}(x) = \sum_{j=1}^{N} \alpha_j \phi(x - x_j),$$

where the coefficients α_j are determined by the interpolation conditions

$$\begin{aligned} \Phi_{f,X}(x_j) = y_j,\ 1 \leq j \leq N. \\ A_{\phi,X}\alpha = y. \\ [A_{\phi,X}]_{j,k} = \phi(x_j - x_k). \end{aligned}$$

The solution of the linear system depends on some technical properties of the trial function. Let us recall the basics.

Definition 32 A continuous function $\phi : \mathbb{R}^n \to \mathbb{C}$ is called positive semi-definite if, for every $N \in \mathbb{N}$ and every pairwise distinct centers $X = \{x_1, x_2, ..., x_N\} \subset \mathbb{R}^n$ and every $\alpha \in \mathbb{C}^N$, the cuadratic form

$$\sum_{j=1}^{N} \sum_{k=1}^{N} \alpha_j \overline{\alpha}_k \phi(x_j - x_k)$$

is nonnegative. The function ϕ is called positive definite if the cuadratic form is positive for every $\alpha \in \mathbb{C}^N \backslash \{0\}$.

For positive definite functions, the interpolating system is uniquely solvable.

Examples 33

1. The Gaussian $\phi(x) = \exp(-\alpha|x|^2)$, $\alpha > 0$, is positive on $\mathbb{R}^n$.
2. The inverse multiquadrics $\phi(x) = (c^2 + |x|^2)^{-\beta}$, $x \in \mathbb{R}^n$, with $c > 0$ and $\beta > n/2$.

Definition 34 We say that a function $\phi : \mathbb{R}^n \to \mathbb{R}$ is radial if there exist a function $\psi : [0, \infty) \to \mathbb{R}$ such that $\phi(x) = \psi(|x|)$, for every $x \in \mathbb{R}^n$. We say that ψ is positive definite on $\mathbb{R}^n$ if $\phi(x) = \psi(|x|)$ is positive definite.

Examples 35

1. The Gaussian $\psi(r) = \exp(-\alpha r^2)$, $\alpha > 0$.
2. The inverse multiquadrics $\phi(r) = (c^2 + r^2)^{-\beta}$, with $c > 0$ and $\beta > n/2$.
3. The truncated power function

   $$\psi_l(r) = (1 - r)_+^l$$

   is positive definite on $\mathbb{R}^n$ if $l \in \mathbb{N}$ satisfies $l \geq \lfloor n/2 \rfloor + 1$.

Definition 36 We say that a continuous function $\phi : \mathbb{R}^n \to \mathbb{C}$ is Conditionally positive semi-definite of order m if, for every $N \in \mathbb{N}$, for every set of pairwise distinct centers $X = \{x_1, x_2, ..., x_n\} \subset \mathbb{R}^n$ and every $\alpha \in \mathbb{C}^N$, such that

$$\sum_{j=1}^{N} a_j p(x_j) = 0$$

for every complex-valued polynomial of degree less than m, the cuadratic form

$$\sum_{j=1}^{N}\sum_{k}^{N}\alpha_j\overline{\alpha}_k\phi(x_j-x_k)$$

is nonnegative. The function ϕ is called Conditionally positive definite if the cuadratic form is positive for every $\alpha\in\mathbb{C}\backslash\{0\}$.

The conditional positive definiteness of order m of a function ϕ can also be interpreted as the positive definiteness of the matrix $A_{\phi,X}$ on the space of vectors α such that

$$\sum_{j=1}^{N}\alpha_j p_l(x_j)=0, 1\leq l\leq Q=\dim\pi_{m-1}(\mathbb{R}^n).$$

Thus, in this case, $A_{\phi,X}$ is positive definite on the space of vectors α "perpendicular" to polynomials.

Examples 37 For $\phi(x)=\psi(|x|)$

1. The multiquadrics $\psi(r)=(-1)^{\lceil\beta\rceil}(c^2+r^2)^\beta$ with $c,\beta>0$, and $\beta\notin\mathbb{N}$, are conditionally positive definite of order $m=\lceil\beta\rceil$ on $\mathbb{R}^d$.

2. The Thin-plate splines $\psi(r)=(-1)^{k+1}r^{2k}\log(r)$ are positive definite of order $m=k+1$ on $\mathbb{R}^n$.

3. The function $\psi(r)=(-1)^{\lceil\beta/2\rceil}r^\beta$ $\beta>0$, $\beta\notin\not\models\mathbb{N}$ is conditionally positive definite of order $m=\lceil\beta/2\rceil$ on $\mathbb{R}^n$.

Here we only work with positive definite functions.

In the set displayed on figure 3.1.1 we show an interpolation using the Gaussian functions and the truncated power functions and as we can see in figure 3.1, both approximations fit the function $u^*(x,y)=(x^2+y^2)^{3/2}$.

3.1.2 Estimates for basis functions

In this subsection we will briefly study the estimates for the Gaussians and the truncated power functions. The theory on this subsection can be found in [8].

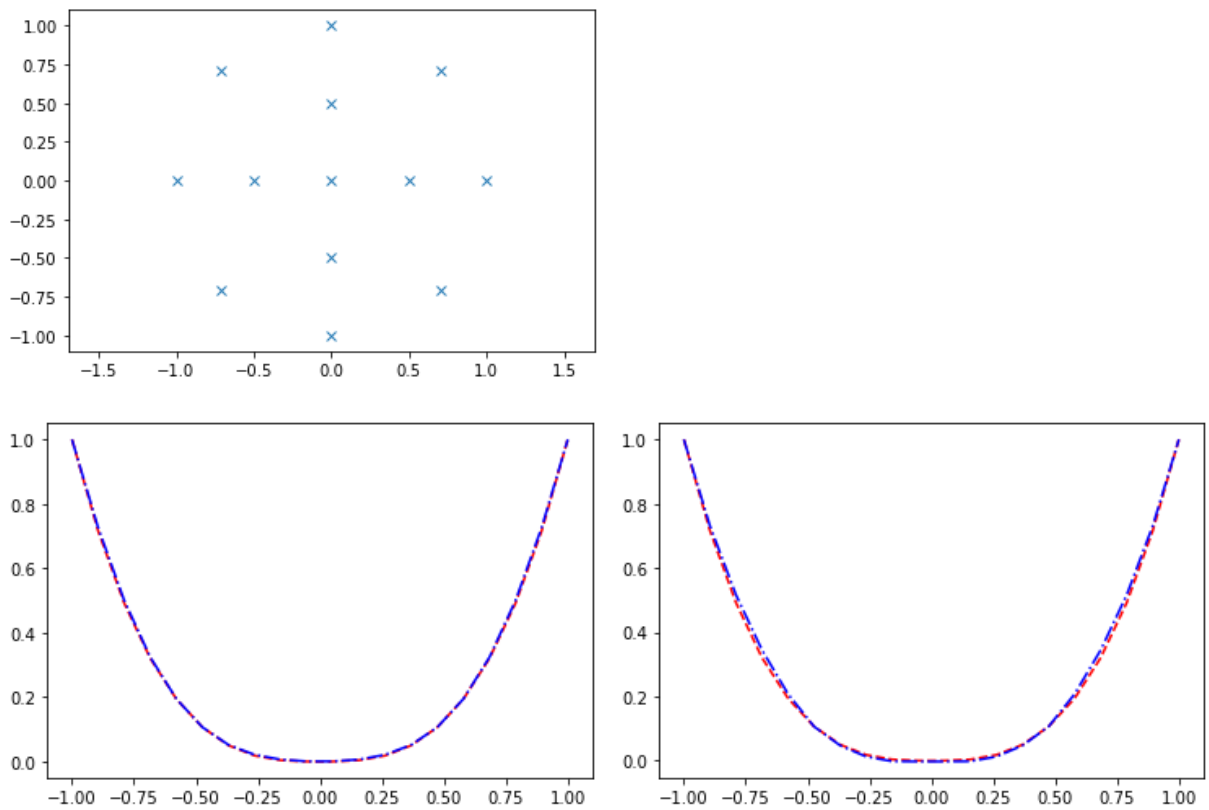


Figure 3.1: The blue graphs are the approximations and the red graph is the graph of u^*.

Definition 38 Let $\mathcal{F}$ be a real Hilbert space of functions $f : \Omega \to \mathbb{R}$. A function $\Phi : \Omega \times \Omega \to \mathbb{R}$ is called a reproducing kernel for $\mathcal{F}$ if

1. $\Phi(\cdot, y) \in \mathcal{F}$ for all $y \in \Omega$,
2. $f(y) = (f, \Phi(\cdot, y))_{\mathcal{F}}$ for all $f \in \mathcal{F}$ and all $y \in \Omega$.

We define the $\mathbb{R}$-linear space

$$F_\Phi(\Omega) := \{\Phi(\cdot, y) : y \in \Omega\}$$

and equip this space with the bilinear form

$$\left(\sum_{j=1}^{N} \alpha_j \Phi(\cdot, x_j), \sum_{k=1}^{M} \beta_k \Phi(\cdot, y_k) \right)_\Phi := \sum_{j=1}^{N} \sum_{k=1}^{M} \alpha_j \beta_k \Phi(x_j, y_k).$$

Theorem 39 If $\Phi : \Omega \times \Omega \to \mathbb{R}$ is a symmetric positive definite kernel then $(\cdot,\cdot)_\Phi$ defines an inner product on $F_\Phi(\Omega)$. Furthermore, $F_\Phi(\Omega)$ is a pre-Hilbert space with reproducing kernel Φ.

We define the completion $\mathcal{F}_\Phi(\Omega)$ of this pre-Hilbert space with respect to the $||\cdot||_\Phi$-norm. We also define a linear mapping

$$R : \mathcal{F}_\Phi(\Omega) \to C(\Omega),\ R(f)(x) := (f, \Phi(\cdot, x))_\Phi.$$

Definition 40 The native Hilbert function space corresponding to the symmetric positive definite kernel $\Phi : \Omega \times \Omega \to \mathbb{R}$ is defined by

$$\mathcal{N}_\Phi(\Omega) := R(\mathcal{F}_\Phi(\Omega)).$$

It carries the inner product

$$(f, g)_{\mathcal{N}_\Phi(\Omega)} := (R^{-1}f, R^{-1}g)_\Phi.$$

Theorem 41 Let Φ be the Gaussians. Suppose that $\Omega \subset \mathbb{R}^d$ is bounded and satisfies an interior cone condition. Denote the radial basis function interpolant to $f \in \mathcal{N}_\Phi(\Omega)$ based on Φ and $X = \{x_1, ..., x_N\}$ by $S_{f,X}$. Fix $\alpha \in \mathbb{N}_0^d$. For every $\in \mathbb{N}$ with $l \geq \alpha$ there exist constants $h_0(l)$, $C_l > 0$ such that

$$|D^\alpha f(x) - D^\alpha S_{f,X}(x)| \leq C_l h_{X,\Omega}^{l-|\alpha|} |f|_{\mathcal{N}_\Phi(\Omega)}$$

for all $x \in \Omega$, provided that $h_{X,\Omega} \leq h_0(l)$.

Definition 42 With $\phi_l(r) = (1-r)_+^l$ we define

$$\phi_{d,k} = (I)^k \phi_{\lfloor d/2 \rfloor + k + 1}.$$

Theorem 43 Within its support $[0, 1]$ the function $\phi_{d,k}$ has the representation

$$p_{d,k}(r) = \sum_{j=0}^{l+2k} d_{j,k}^{(l)} r^j$$

with $l = \lfloor d/2 \rfloor + k + 1$.

Theorem 44 The functions $\phi_{d,k}$ are positive definite on $\mathbb{R}^d$ and are of the form

$$\phi_{d,k}(r) = \begin{cases} p_{d,k}(r), & \text{if } 0 \leq r \leq 1, \\ 0, & \text{if } r > 1, \end{cases}$$

with a univariable polynomial $p_{d,k}$ of degree $\lfloor d/2 \rfloor + 3k + 1$.

Theorem 45 Let $\Phi_{d,k} = \phi_{d,k}(|| \cdot ||_2)$ be the functions from the previous theorem. Suppose that $\Omega \subset \mathbb{R}^d$ is bounded and satisfies an interior cone condition. Denote the radial basis function interpolant of $f \in \mathcal{N}_{\Phi_{d,k}}(\Omega)$ based on $\Phi_{d,k}$ and $X?\{x_1, ..., x_N\} \subset \Omega$ by $S_{f,X}$. Then there exist constants $C, h_0 > 0$ such that

$$|D^\alpha f(x) - D^\alpha S_{f,X}(x)| \leq C h_{X,\Omega}^{k+1/2-|\alpha|} ||f||_{\mathcal{N}_\Phi(\Omega)}$$

for every $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \leq k$ and every $x \in \Omega$, provided that $h_{X,\Omega} \leq h_0$.

Theorem 46 Let Ω be a cube in $\mathbb{R}^d$. Suppose that $\Phi = \phi(|| \cdot ||_2)$ is a conditionally positive definite function such that $f := \phi(\sqrt{\cdot})$ satisfies $|f^{(l)}(r)| \leq l! M^l$ for all integers $l \leq l_0$ and all $r \in [0, \infty)$, where $M > 0$ is a fixed constant. Then there exist a constant $c > 0$ such that the error between a function $f \in \mathcal{N}_\Phi(\Omega)$ and its interpolant $S_{f,X}$ can be bounded by

$$||f - S_{f,X}||_{L_\infty(\Omega)} \leq e^{-c/h_{X,\Omega}} |f|_{\mathcal{N}_\Phi(\Omega)}$$

for all data sites X with sufficiently small $h_{X,\Omega}$.

From theorem 45 we get that the truncatec power functions have algebraic convergence. On the other hand, from theorem 46 we get that the Gaussian functions have spectral convergence.

3.2 A Dirichlet Problem in the disk

Here we consider the unit disc centered on the origin, as motivated by the Newton Problem.

As trial space, consider the radial basis functions (RBF) of the form $\psi(r) = \exp(-\varepsilon r^2)$.

3.2.1 Numerical Results

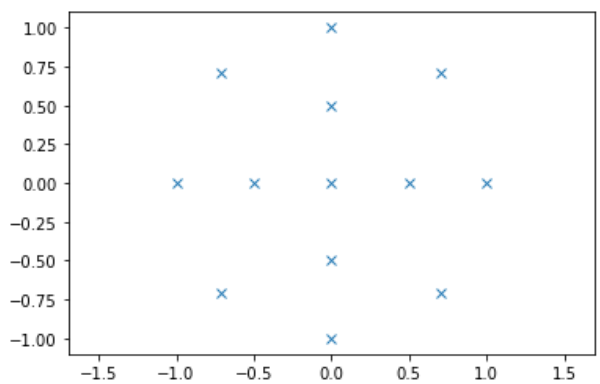
In this thesis we are not following the typical process of making a sample of points with uniform distribution on the domain. The goal is to show that it is possible to solve the M-A equation with a meshfree method.

The nonlinear system of equations generated by this method is

$$\Psi(x_i) = f(x_i),\ 1 \leq i \leq n$$
$$\Psi_{xx}(x_i)\Psi_{yy}(x_i) - \Psi_{xy}(x_i)^2 = g(x_i),\ n+1 \leq i \leq x.shape$$

where n is the number of points in the boundary.

The following graphs are obtained with the next set of points



The graphs on figure 3.2 are the results after solving the M-A equation given by $\det D^2u = 18(x^2+y^2)$. On the left side we have the graphs for $y = 0$ and then $x = 0$ for the Gaussian functions. On the right side we have the graphs for $y = 0$ and then $x = 0$ for the truncated power functions.

It is shown in figure 3.1 that accurate interpolation can be achieved with both RBF but, as we can see in figure 3.2, the numerical solution of PDE is a different matter.

In our numerical experiments, the same conclusion is reached, the Gaussian function outperforms other choices. This makes sense because, as we said in the previous section, the Gaussians have an exponential convergence. In what follows we only show the satisfactory results obtained with the Gaussian functions.

In figure 3.3 we choose a different initial guess to run the iteration. We also obtain a close approximation to the real solution.

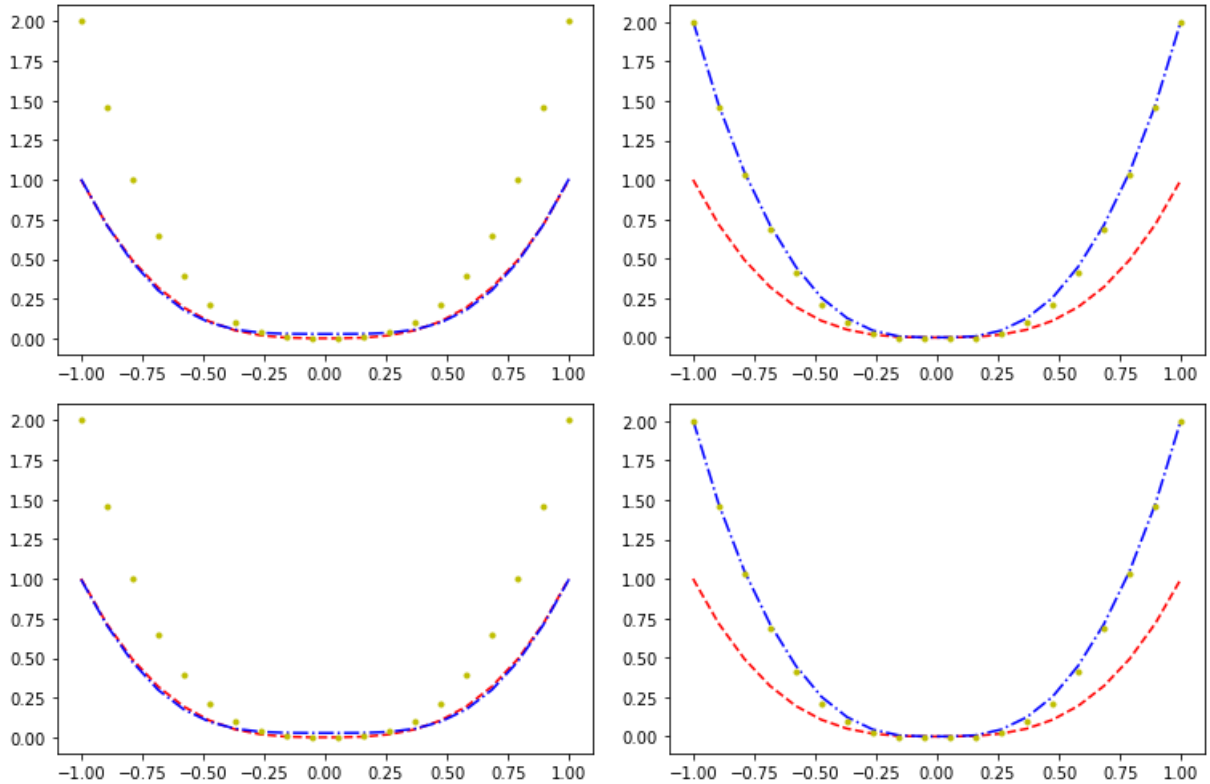


Figure 3.2: The yellow dots represent the initial guess, obtained doubling the coefficients from the interpolations. The red graph is the real solution and the blue graph is the approximation.

We also have an example in figure 3.4 where the real solution is the function $u^*(x, y) = \exp((x^2 + y^2)/2)$.

3.3 A Dirichlet Problem in the unit square

Here we solve the examples in [1]. Consider the real solution $u^*(x, y) = \exp((x^2 + y^2)/2)$ in the unit square. We get the results on figure 3.5.

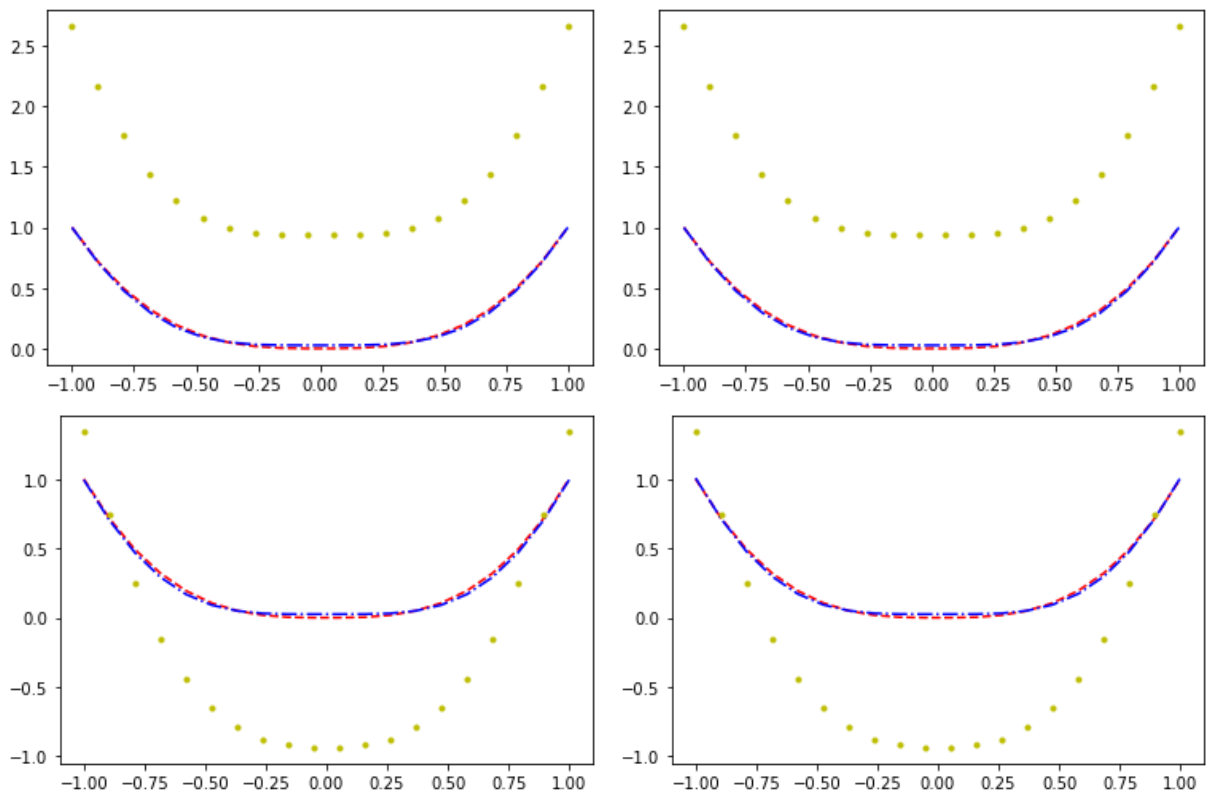


Figure 3.3: Graphs for the solutions on $y = 0$ and $x = 0$ using the Gaussian functions on two different initial guess. The value of the error function is respectively $5.53e - 07$ and 0.00016.

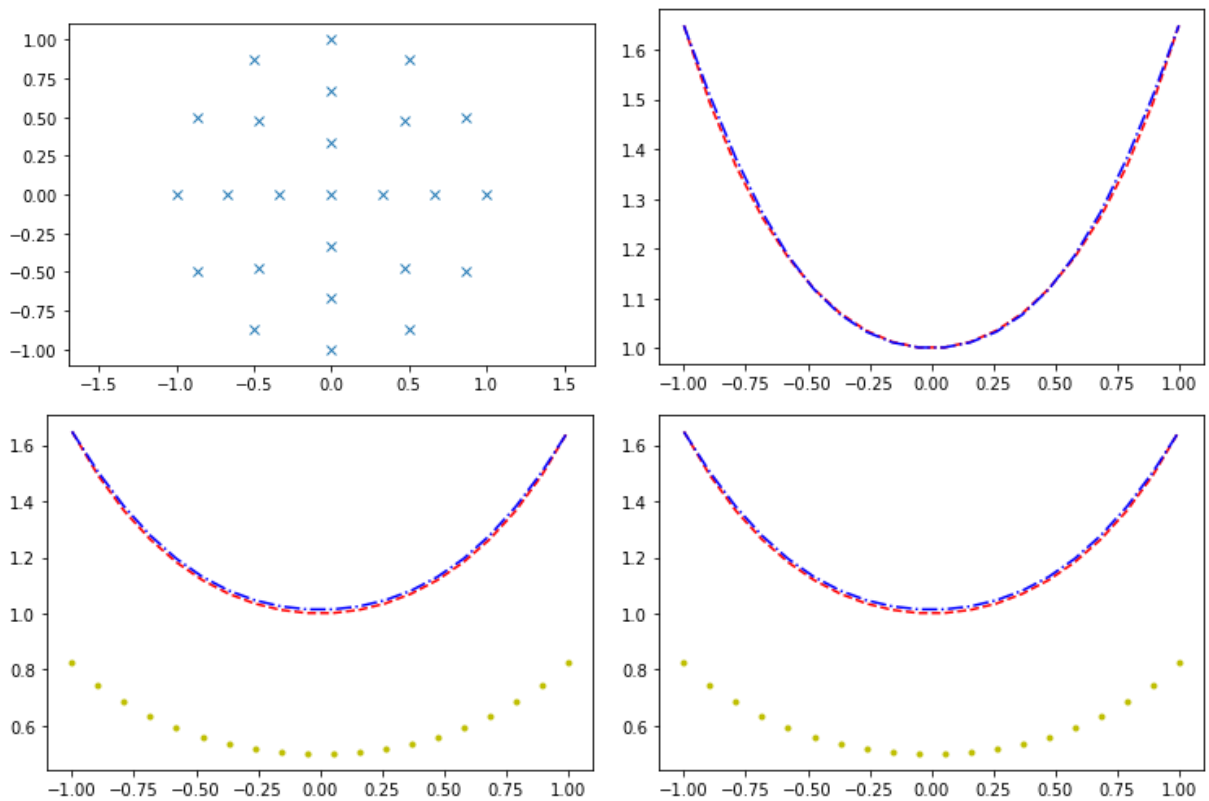


Figure 3.4: Set of points, interpolation and graphs for the solutions on $y = 0$ and $x = 0$ using the Gaussian functions. The value of the error function is 0.011.

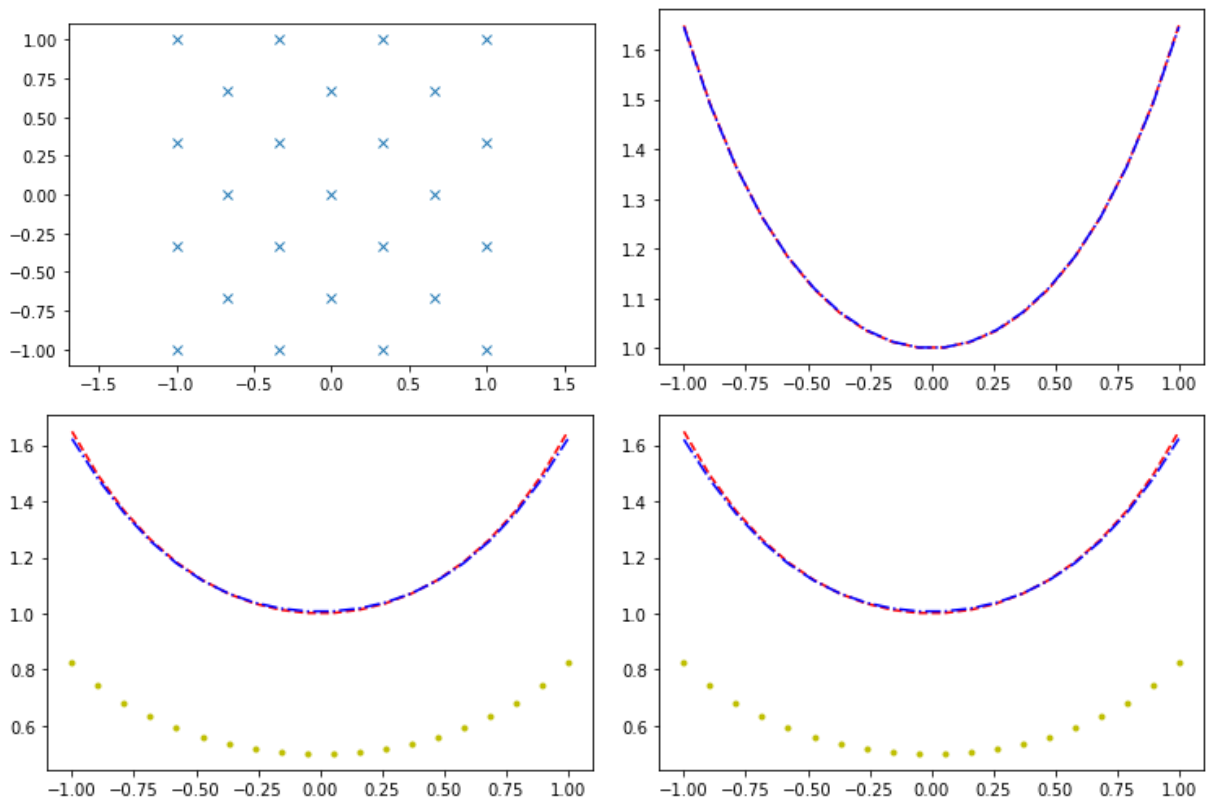


Figure 3.5: Set of points, interpolation and graphs for the solutions on $y = 0$ and $x = 0$ using the Gaussian functions. The value of the error function is 0.0706.

Chapter 4

Conclusions and future work

In this work we have presented several aspects of the Monge-Ampère equation. In the Newton problem of minimal resistance, it is shown that an optimal function is a smooth local solution of the M-A equation.

Once the M-A equation is introduced, the classical question of well posedness is addressed. Existence and uniqueness are established for generalized and viscosity solutions.

In applications, the numerical solution of the M-A equation is required. Motivated by the meshless approach of [1] with polynomial basis, an alternative is developed using Gaussian Radial Basis Functions. The benchmark problems in [1] are solved in both the unit disc and the unit square. The former in line with the Newton problem of minimal resistance in its original formulation.

These three problems are of intensive research activity. In the case of the minimal resistance problem, appropiate admissible sets are sought for minimization. The geometry of the domain is also a topic of interest, see [6]. Actual solutions of the Newton problem are of practical interest, and numerical methods are being developed to solve the minimization problem.

The numerical solution of the Monge-Ampère equation is an unresolved issue. From our preliminary numerical explorations, meshfree methods seem promising.

In practice, the M-A equation is descretized leading to a nonlinear high dimensional algebraic problem. As in [1], we have used a general use implementation of quasi-Newton methods. For improvement, a specific modification needs to be developed.

Appendix

The results on this appendix can be found on [7].

Definition 47 We say that J is strongly continuous (we say simply continuous if there is no ambiguity), if

$$v^k \to v \text{ (strongly) } \Rightarrow J(v^k) \to J(v) \text{ (in } \mathbb{R}).$$

Similarly, we say that J is weakly continuous if

$$v^k \to v \text{ (weakly) } \Rightarrow J(v^k) \to J(v) \text{ (in } \mathbb{R}).$$

Lets notice that

$$J \text{ weakly continuous } \Rightarrow J \text{ strongly continuous.}$$

Recall that a set $U \subset V$ is strongly (respectively weakly) compact if from every sequence $\{v^k\}$ of elements of U we can extract a sub-sequence which converges strongly (respectively weakly) in U.

The following theorem gives a sufficient condition for the minimization problem

$$\min_{v \in U \subset V} \{J(v)\} \tag{4.1}$$

to have an optimal solution in U.

Theorem (Weierstrass) 48 If the subset $U \subset V$ is strongly (respectively weakly) compact, and if J is strongly (respectively weakly) continuous on U, then the problem (4.1) has an optimal solution in U.

Let V be a normed vector space and let J be a functional on V.

Definition 49 We say that J has a directional derivative (or a diferential in the sense of Gateaux) at $v \in V$ in the direction $\varphi \in V$, if

$$\frac{J(v+\theta\varphi) - J(v)}{\theta}$$

has a limit when $\theta \to 0$ (in $\mathbb{R}$). This limit is denoted $\delta J(v, \varphi)$.
If $\forall \varphi \in V$: $\delta J(v, \varphi)$ exists, then J is said to be differentiable in the sense of Gateaux (G-differentiable) at $v \in V$.

Definition 50 Let V be a Hilbert space with the scalar product $< \cdot, \cdot >$. If J is G-differentiable at $v \in V$, and if $\delta J(v, \varphi)$ is a continuous linear form w.r. to φ, then (by the representation theorem of Riesz), there exist an element $J'(v) \in V$ such that

$$\forall \varphi \in V : \ \delta J(v, \varphi) =< J'(v), \varphi >$$

$J'(v)$ is called the gradient of J at v.

Proposition 51 If J is G-differentiable at $v + \alpha\varphi$ whatever $\alpha \in [0, 1]$ in the direction φ, then there exist $\theta \in (0, 1)$ such that

$$J(v + \varphi) = J(v) + J(v + \theta\varphi, \varphi).$$

Definition 52 We say that J has a second derivative in the sense of Gateaux at the point v in the directions φ and ψ if the ratio

$$\frac{\delta J(v + \theta\psi, \varphi) - \delta J(v, \varphi)}{\theta}$$

has a limit when $\theta \to 0$ in $\mathbb{R}$. This limit is denoted $\delta^2 J(v, \varphi, \psi)$. If $\delta^2 J(v, \varphi, \psi)$ exist $\forall \varphi \in V$, $\forall \psi \in V$, then we say that J is twice G-differentiable at the point $v \in V$.
If, moreover, for $v \in V$, $\delta^2 J(v, \varphi, \psi)$ is continuous and linear in φ and ψ, then there exist a linear operator $\kappa(v) : V \to V$ such that

$$\delta^2 J(v, \varphi, \psi) =< \kappa(v) \cdot \psi, \varphi >$$

$\kappa(v)$ is called the Hessian of J at v.

Proposition 53 If, for all $\alpha \in [0,1]$, J is twice G-differentiable at v in the directions φ and $\psi = \varphi$, then there exist $\theta \in (0,1)$ such that

$$J(v+\varphi) = J(v) + \delta J(v,\varphi) + \frac{1}{2}\delta^2 J(v+\theta\varphi,\varphi,\varphi).$$

Theorem 54 Let $J(v)$ be a functional on V, G-differentiable at $v^0 \in V$. A necessary condition for v^0 to be an optimum of J is to have

$$\delta J(v^0,\varphi) = 0 \ (\forall \varphi \in V).$$

Theorem 55 Let $J(v)$ be a functional on V, twice G-differentiable at $v^0 \in V$. A necessary condition for v^0 to be an optimum of J is that for all $\varphi \in V$ we have

$$\begin{cases} \delta J(v^0,\varphi) = 0 \\ \delta^2 J(v^0,\varphi) \geq 0. \end{cases}$$

Bibliography

[1] K. Böhmer, R. Schaback, *A meshfree method for solving the Monge-Ampére equation*, Springer, 2018.

[2] Buttazzo, G. *A survey on the Newton problem of optimal profiles. In Variational Analysis and Aerospace Engineering* (pp. 33-48). Springer, New York, NY., 2009.

[3] D. Bucur, G. Buttazzo, *Variational Methods in Shape Optimization Problems*, Birkhauser, 2005.

[4] Lawrence C. Evans, *Partial Differential Equations*, Department of Mathematics University of California, Berkeley, 2010.

[5] Cristian E. Gutiérrez, *The Monge-Ampére Equation*, Birkhauser, 2016.

[6] Mainini, E., Monteverde, M., Oudet, E., & Percivale, D. *The minimal resistance problem in a class of non convex bodies. ESAIM: Control, Optimisation and Calculus of Variations*, 25, 27, 2019.

[7] M. Minoux *Mathematical Programming: Theory and Algorithms*, University of Michigan, 1986.

[8] Wendland, H. *Scattered data approximation* (Vol. 17). Cambridge university press, 2004.