Wide Stencil for the Monge Ampère Equation

by

Jessey Lin

A research paper
presented to the University of Waterloo
in partial fulfillment of the
requirement for the degree of
Master of Mathematics
in
Computational Mathematics

Supervisor: Prof. Justin Wan W. L.

Waterloo, Ontario, Canada, 2014

I hereby declare that I am the sole author of this report. This is a true copy of the report, including any required final revisions, as accepted by my examiners.

I understand that my report may be made electronically available to the public.

Abstract

We propose a new numerical scheme to solve the elliptic Monge Ampère Equation (MAE) with Dirichlet boundary condition. The problem is motivated from applications of the MAE to image registration modelling. The MAE is challenging to solve, it is fully nonlinear and has non-unique solutions and a general and efficient numerical scheme is difficult to construct. Our numerical algorithm solves the MAE by transforming it to a Hamilton-Jacobi-Bellman (HJB) equation, which has the form of a linear PDE coupled nonlinearly with two control parameters. The HJB equation is further discretized by a wide stencil method. We prove the Barles-Souganidis convergence of the numerical scheme to the viscosity solution by showing consistency, stability and monotonicity. The performance of the numerical method will be shown by examples of smooth and singular MAE problems.

Acknowledgements

I would like to thank my supervisor, Professor Justin W. L. Wan for his guidance and encouragement. I would like to thank my CM classmates who are all going to be future leaders of the 21st century! I have learnt a lot from them. In addition, I would like to thank all the teachers who taught me here at the university, I had a truly rewarding year.

Dedication

This is dedicated to my family and to all my teachers whose influence, is a life time.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The Monge Ampère equation belongs to the category of fully nonlinear second order PDEs. It was first studied by Gaspard Monge in 1784 and later by Andrè-Marie Ampère in 1820. It has wide applications in differential geometry problems such as the Minkowski problem and optimization problems such as the Monge-Kantorovich minimization problem. Areas such as astrophysics, medical image analysis and reflector design apply models using the MAE [11]. In this chapter, we will motivate our study of the MAE and review some of the existing methods.

1.1 Image Registration and the MAE

Our interest of the MAE stems from the image registration problem. Figure 1.1a and Figure 1.1b shows two images, the reference (R) and the template (T), of an MR scan of the human knee. They are to be aligned or *registered* for clinical purposes. However R is bent to an angle while T is not bent. One cannot apply simple linear transformations to align them so we need to find a good transformation $y : R \to T$, in such a way that their difference $T(y) - R$ in the resulting image is minimized.

To mathematically tackle the problem, we consider the model provided by the optimal transport problem [13].

The optimal transport problem seeks to find an optimal mapping y between two density functions R, T defined on $\Omega \subseteq \mathbb{R}^2$ with the constraint that mass is preserved, i.e.

$$\int_\Omega R(x)\,\mathrm{d}x = \int_\Omega T(y(x)) det(Dy(x))\,\mathrm{d}x \tag{1.1}$$

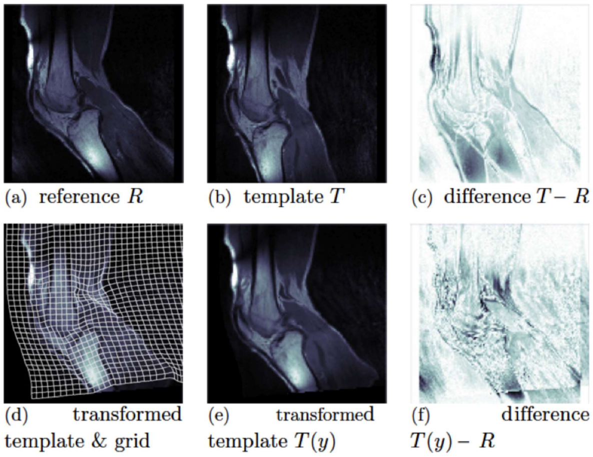


Figure 1.1: Registration of human knee. (a)-(c) : registration without transformation. (d)-(f): registration with transformation y. (Image courtesy of IOP Publishing, [9])

so that their Monge-Kantorovich distance metric,

$$d(R,T) = \min \int_{\Omega} \|x - y(x)\|^p R(x)\,\mathrm{d}x$$

is minimized.

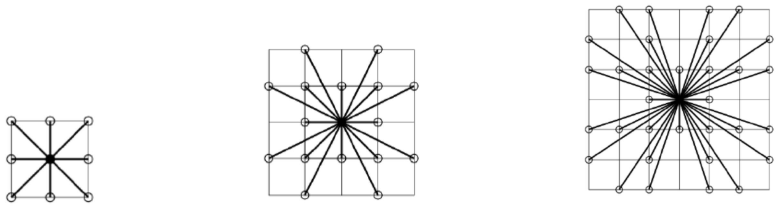
In particular, when $p = 2$, we can write the optimal map, $\bar{y}$, as:

$$\begin{gathered} \bar{y} = \nabla\phi, \\ \phi \text{ convex on } \Omega, \end{gathered} \tag{1.2}$$

where ϕ satisfies the Monge Ampère Equation (MAE):

$$det(D^2\phi(x)) = \frac{R(x)}{T(\nabla\phi)} =: f.$$

This is the mass preserving requirement and can be easily observed when subsituting (1.2) into (1.1) above.



(a) 9 point stencil (b) 17 point stencil (c) 33 point stencil

Figure 1.2: Reproduction of figures for the wide stencil scheme proposed by Oberman et al [11]. The number of stencil points (represented by circles) increases with decreasing mesh size.

1.2 Numerical methods

Due to the high non-linearity and non-uniqueness of the MAE, it poses a number of numerical challenges to set up with the right numerical scheme.

Fortunately, the frame work provided by Barles and Souganidis [3] allows one to study numerical schemes to overcome the aforementioned difficulties. It basically states that if the numerical scheme is consistent, stable and monotone in an appropriate sense, then it would converge to the viscosity solution of the MAE.

In reference to this framework, Oberman et al [11] has developed a monotone wide stencil finite difference scheme to approximate the fully nonlinear PDE. Note that a monotone finite difference scheme even for a linear elliptic PDE, using a narrow stencil does not in general exist [7]. However, Oberman's method may be computationally expensive since the number of stencil points quadruples as each dimension of the 2D domain increases two-folds; see Figure 1.2

Galerkin type methods based on approximating infinite spaces with finite ones such as the augmented Lagragian and least squares methods were developed by Dean and Glowinski [6]. Their method involves formulating the MAE into a constrained minimization problem and solving the non-linear systems numerically. However, when the solution is not smooth, convergence may not always be guaranteed.

Finite element methods such as the vanishing moment method was studied by Feng and Neilan [8]. The method involves approximating a fully non-linear second order PDE by a sequence of higher order quasi-linear PDEs. However, highly non-linear systems need to be computed and the boundary conditions are hard to approximate well [7].

1.3 Overview of The Essay

In this essay, we will study the degenerate elliptic fully nonlinear second order Monge Ampère equation on the square domain with a positive source term f and Dirichlet boundary conditions. This essay first applies a transformation to the MAE in Chapter 2 to a nonlinear HJB PDE with a linear objective function. Then it adopts a monotone wide stencil scheme [16] in Chapter 3 to approximate the numerical solution. In Chapter 4, we prove that the numerical scheme converges to the unique viscosity solution. Finally, in Chapter 5, we will present examples of numerical experiments and convergence results.

Chapter 2

The Monge Ampère Equations

2.1 The Monge-Ampère Equation

The Monge Ampère equation (MAE) belongs to the class of fully non-linear second order partial differential equations. Fully non-linear PDEs are the class of non-linear PDEs which are non-linear in the highest order derivatives. Formally the operator is of the form:

$$\mathcal{F}[u](x) = \mathcal{F}\left(D^2u(x), \nabla u(x), u(x), x\right) = 0,$$

where $\mathcal{F} \in C\left(\mathbb{R}^{d\times d} \times \mathbb{R}^d \times \mathcal{R} \times \Omega\right)$ and $\Omega \subseteq \mathbb{R}^d$ is a bounded domain. Here, $D^2u(x)$ denotes the hessian matrix of u at x. Moreover, it is *degenerate elliptic* if

$$\mathcal{F}\left(B, p.z, x\right) \leq \mathcal{F}\left(A, p.z, x\right)$$

for all $x \in \Omega, z \in \mathbb{R}, p \in \mathbb{R}^d$ and A, B are symmetric $d \times d$ matrices with $A \geq B$ which means that $A - B$ needs to be positive semi-definite. In this paper, we will consider the non-homogeneous Dirichlet Monge Ampère equations, i.e.

$$\begin{aligned} \mathcal{F}u = u_{xx}u_{yy} - u_{xy}^2 = f(x,y) \quad in \quad \Omega, \\ u = g \quad on \quad \partial\Omega. \end{aligned} \tag{2.1}$$

The MAE is a (degenerate) elliptic operator only if we impose the additional requirement that

1. u is strictly convex,

2. $f > 0$.

Interested readers can refer to [5] for the reference. The domain Ω will be any convex bounded region in the two dimensional Euclidean space $\mathbb{R}^2$. In this paper, we will consider the square domain given by

$$\Omega = [\,0, 1\,] \times [\,0, 1\,].$$

2.2 Viscosity Solutions of the Monge-Ampère Equation

In general, classical solutions of the MAE do not exist. We need weaker versions of the concept of 'solutions'.

Definition 2.1 (Viscosity solution)**.** *Let $F[\varphi] = det(D^2\varphi) - f$.The function $u \in C(\Omega)$ is a **viscosity subsolution (supersolution)** of F if whenever $\varphi \in C^2(\Omega)$ and $x_0 \in \Omega$ maximizes (minimizes) $u - \varphi$ for all x in a neighborhood of x_0, then we must have*

$$F[\varphi](x_0) \geq (\leq) 0.$$

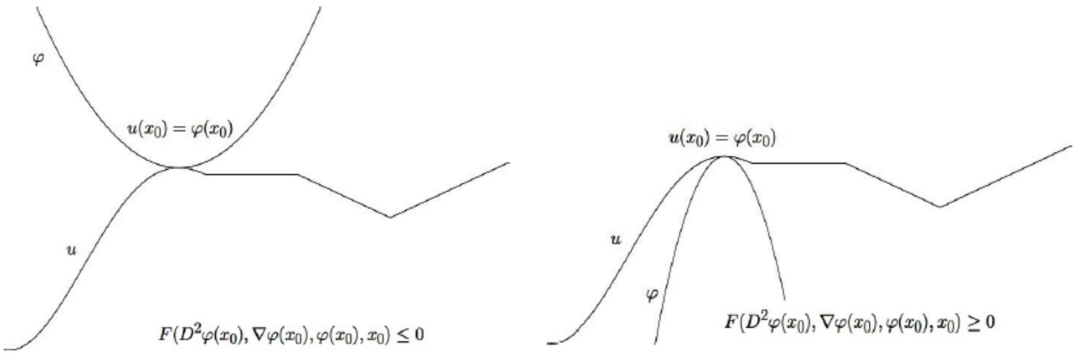
The function u is a viscosity solution if it is both a viscosity subsolution and supersolution.

Geometrically, u is a viscosity subsolution if for every test function $\phi \in C^2$ that touches the graph of u from above at x_0 in Figure 2.1a, there holds $F[\phi](x_0) \leq 0$ and if ϕ touches the graph from below at x_0 in Figure 2.1b, there holds $F[\phi](x_0) \geq 0$.

For the existence and uniqueness of the viscosity solution of problem (2.1), we need the following theorem.

Theorem 2.1. *[12] Let $\Omega \subseteq \mathbb{R}^d$ be bounded and strictly convex, $g \in C(\partial\Omega)$, $f \in C(\Omega)$ with $f \geq 0$. Then there exists a unique convex viscosity solution $u \in C(\bar{\Omega})$ of problem (2.1).*

In Chapter 4, we will see a general framework provided by Barles and Souganidis [3] allows one to show that their approximation schemes achieve convergence to the viscosity solutions.


φ
$u(x_0) = \varphi(x_0)$
u
$F(D^2\varphi(x_0), \nabla\varphi(x_0), \varphi(x_0), x_0) \leq 0$
$u(x_0) = \varphi(x_0)$
u
φ
$F(D^2\varphi(x_0), \nabla\varphi(x_0), \varphi(x_0), x_0) \geq 0$

(a) u subsolution

(b) u supersolution

Figure 2.1: Illustrations for viscosity solutions. (Image courtesy of SIAM Review, [7])

2.3 From the Monge-Ampère Equation to Hamilton-Jacobi-Bellman Equation

MAE has derivative terms which are quadratic and it is difficult to construct a numerical scheme without having to compute complex non-linear systems. If we can transform the MAE to HJB, which has a linear PDE objective function, then it would be much easier to discretize. In fact, we shall see that the Monge-Ampère equation can be formulated in the following form:

$$\min_{\alpha \in \mathrm{Z}} \left[\mathcal{L}^{\alpha} u - f\,\right] = 0, \qquad \mathcal{L}^{\alpha} \text{ is a linear operator,}$$

where $\alpha \in \mathrm{Z}$ is the set of admissible controls. The equivalent formulation above was first proved in [15] but we will present the version from [18] here. First we need the following lemma.

Lemma 2.1. *Let H be symmetric, $g > 0$ and S_1^+ be defined as above. Then H satisfies*

$$\max_{A \in S_1^+} \left[Tr(AH) + g\sqrt{det(A)} \right] = 0$$

if and only if

$$H \text{ is negative definite,}$$

$$2\sqrt{det(-H)} = g.$$

Interested readers may refer to [15], [18] for the proof. In fact it follows from a variant of the AM-GM inequality:

$$A, B \geq 0, \qquad 2\sqrt{det(AB)} \leq Tr(AB),$$

and the properties of the set S_1^+.

Theorem 2.2. *Let $u = u(x, y)$ and Ω be convex in $\mathbb{R}^2$. Then u solves the elliptic MAE:*

$$det(D^2u(x)) = f^2, \tag{2.2}$$
$$D^2u(x)\, positive\ definite\ on\, \Omega, \tag{2.3}$$

if and only if it solves the HJB:

$$\min_{A \in S_1^+} \left[Tr(AD^2u(x)) - 2f\sqrt{det(A)} \right] = 0, \tag{2.4}$$

where $S_1^+ = \{A \in \mathbb{R}^{d \times d} : A \geq 0, Tr(A) = 1\}$.

Proof. : Let $g = 2f$, $w = -u$ and $H = D^2w$. Applying Lemma 2.1, we have,

$$\begin{aligned} \left\{ \begin{array}{l} det(D^2u) = f^2 \\ D^2u(x) \text{ positive definite } \quad \text{on } \Omega, \end{array} \right. &\iff \left\{ \begin{array}{l} 2\sqrt{det(-D^2w(x))} = g \\ D^2w(x) \text{ negative definite } \quad \text{on } \Omega, \end{array} \right. \\ &\iff \max_{A \in S_1^+} \left[Tr(AD^2w(x)) + g\sqrt{det(A)} \right] = 0, \\ &\iff \min_{A \in S_1^+} \left[Tr(AD^2u(x)) - 2f\sqrt{det(A)} \right] = 0 \end{aligned}$$

□

Hence the MAE with a convex solution u is equivalent to the HJB equation (2.4). In order to compute (2.4) numerically in $\mathbb{R}^2$, we need an explicit form of S_1^+. For example, S_1^+ in $\mathbb{R}^2$ can be parametrized by the set below:

$$\left\{ \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} a & 0 \\ 0 & 1-a \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} : 0 \leq a \leq 1, 0 \leq \theta \leq 2\pi \right\}. \tag{2.5}$$

Finally, if we let $Z = [0, 1] \times [0, 2\pi]$ be the set of all admissible controls and $Q = (a, \theta) \in Z$, then (2.4) can be written as:

$$\min_{Q\in Z}\left\{ \mathcal{L}_H^Q u - 2\sqrt{a(1-a)f} \right\} = 0, \tag{2.6}$$

where

$$\mathcal{L}_H^Q u = d_{11}u_{xx} + 2d_{12}u_{xy} + d_{22}u_{yy}. \tag{2.7}$$

Here f is given in (2.1) and

$$\begin{aligned} d_{11} =&\ a\cos\theta + (1-a)\sin^2\theta, \\ d_{22} =&\ a\sin\theta + (1-a)\cos^2\theta, \\ d_{12} =&\ (1-2a)\cos\theta\sin\theta. \end{aligned} \tag{2.8}$$

Remark 2.1. *It is not hard to verify that*

$$D = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix}$$

is a semi-positive definite matrix.

Chapter 3

Discretization of the Monge-Ampère Equation

In this chapter, we will describe our discretization of the transformed Monge-Ampère Equation (2.6). As mentioned, we need the discretization to ensure a monotone scheme which guarantees the convergence to the desired viscosity solution. The discretization that we will adopt is a wide stencil method based on the rotation of the local grid. Additional problems are further addressed due to this method, such as the shrinking of stencil points that fall outside Ω.

3.1 Basic Set Up and Notations

The solution of the HJB equation (2.6) is a function defined in $\Omega \subseteq \mathbb{R}^2$, it will be solved on a set of $n \times n$ grid points. Let $\mathcal{U}_{i,j}$ be the approximate solution of $u(x_i, y_j)$ where $i, j = 0, 1, \ldots, n+1$. Note that since the Dirichlet boundary condition is imposed (2.1), $\mathcal{U}_{i,j}$ are given by the corresponding values of $g(x_i, y_j)$ when $i, j = 0$ or $n+1$. We will also be using a uniform grid, so the size of our grid at each dimension is

$$h = \frac{1}{n+1}.$$

For the purpose of implementing the approximation on a computer, we will compute (2.6) via constructing the linear system:

$$\mathbf{L}^{Q^*}\mathbf{U} = \mathbf{F}^{Q^*}, \tag{3.1}$$

where

$$\mathbf{U} = (\mathcal{U}_{1,1}, \mathcal{U}_{1,2}, \ldots, \mathcal{U}_{n,1}, \ldots, \mathcal{U}_{1,n}, \ldots, \mathcal{U}_{n,n})$$

is the solution vector. $\mathbf{L}^Q$ is the $n^2 \times n^2$ matrix consisting of the coefficients of the discretized HJB and $Q^* = (a^*, \theta^*)$ is the optimal control in the set of discrete admissible controls Z_h. For computational purposes, we have to use a single index to reference an entry of $\mathcal{U}$ above:

$$\mathbf{U}_l = \mathcal{U}_{i,j}, \quad l = i + (j-1)n \quad i, j = 1, \ldots, n.$$

We also let $\mathbf{L}^Q_{l,k}$ be the $(l,k) - th$ entry of the matrix where $k = 1, \ldots, n^2$. We give details on how $\mathbf{L}^Q$ and $\mathbf{F}^Q$ are constructed below.

3.2 The $\mathcal{L}^Q_H$ Operator Discretization

Observe that $\mathcal{L}^Q_H$, we see that it consists of second derivatives of u: u_{xx}, u_{xy}, u_{yy}. The standard approach is to approximate them by central differencing:

$$\begin{aligned}
u_{xx}(x_i, y_j) &\approx \frac{\mathcal{U}_{i-1,j} - 2\mathcal{U}_{i,j} + \mathcal{U}_{i+1,j}}{h^2}, \\
u_{yy}(x_i, y_j) &\approx \frac{\mathcal{U}_{i,j-1} - 2\mathcal{U}_{i,j} + \mathcal{U}_{i,j+1}}{h^2}, \\
u_{xy}(x_i, y_j) &\approx \frac{2\mathcal{U}_{i,j} + \mathcal{U}_{i+1,j+1} + \mathcal{U}_{i-1,j-1}}{2h^2} - \frac{\mathcal{U}_{i+1,j} + \mathcal{U}_{i-1,j} + \mathcal{U}_{i,j+1} + \mathcal{U}_{i,j-1}}{2h^2}, \\
&\text{or} \\
u_{xy}(x_i, y_j) &\approx -\frac{2\mathcal{U}_{i,j} + \mathcal{U}_{i+1,j-1} + \mathcal{U}_{i-1,j+1}}{2h^2} + \frac{\mathcal{U}_{i+1,j} + \mathcal{U}_{i-1,j} + \mathcal{U}_{i,j+1} + \mathcal{U}_{i,j-1}}{2h^2}. \qquad (3.2)
\end{aligned}$$

Substituting (3.2) into (2.6) and collecting terms, we will see that the coefficients of $\mathcal{U}_{p,q,\ p\neq i\ or\ q\neq j}$ are positive but the coefficient of $\mathcal{U}_{i,j}$ is given by the term

$$\frac{2d_{11} + 2d_{22} - 2d_{12}}{h^2},$$

which may be positive or negative. As a result, the scheme may not be monotone. This is due to the presence of the crossed derivative, creating a non-zero d_{12} term in the above discretization. If we can eliminate the u_{xy} term from $\mathcal{L}^Q_H$, we can guarantee a positive coefficient discretization and hence a monotone scheme.

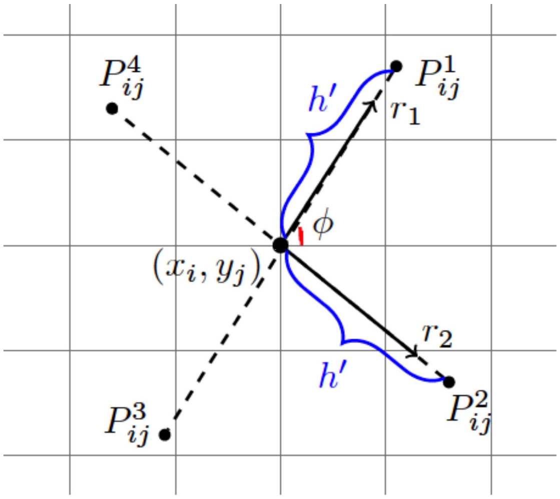


Figure 3.1: Local grid rotation of angle ϕ about (x_i, y_j). $P_{i,j}^m$, $m = 1, \dots 4$ are the new stencil points. r_1, r_2 are the new axes, h' is the new stencil length.

We can eliminate the u_{xy} term by a change of variables at each grid point. This is equivalent to an appropriate rotation ϕ about each (x_i, y_j)(Figure 3.1) such that when the corresponding transformation $X(\phi)$ is applied to $\mathcal{L}_H^Q$, only the terms u_{xx}, u_{yy} remain, i.e. find ϕ such that

$$X(\phi)^{-1}\mathcal{L}_H^Q X(\phi) = d'_{11}\tfrac{\partial^2}{\partial w^2} + d'_{22}\tfrac{\partial^2}{\partial z^2}, \tag{3.3}$$

where d'_{11}, d'_{22} are the corresponding coefficients of the transformed equation. Solving the above, we find that

$$\phi = \frac{1}{2}\arctan\left(\frac{2d_{12}}{d_{11} - d_{22}}\right),$$

$$X(\phi) = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix},$$

and

$$\begin{aligned} d'_{11} &= d_{11}\cos^2(\phi)d_{22}\sin^2(\phi) + 2d_{12}\cos(\phi)\sin(\phi), \\ d'_{22} &= d_{11}\sin^2(\phi)d_{22}\cos^2(\phi) - 2d_{12}\cos(\phi)\sin(\phi). \end{aligned} \tag{3.4}$$

Consider Figure 3.1. Denote the new axes by r_1 and r_2, where

$$r_1 = \begin{pmatrix} \cos\phi \\ \sin\phi \end{pmatrix}, \qquad r_2 = \begin{pmatrix} -\sin\phi \\ -\cos\phi \end{pmatrix}.$$

The new stencil points are $P_{i,j}^m$, $m = 1, \dots 4$, where

$$\begin{aligned} P_{i,j}^1 &= (x_i, y_j) + h' r_1, \\ P_{i,j}^2 &= (x_i, y_j) + h' r_2, \\ P_{i,j}^3 &= (x_i, y_j) - h' r_1, \\ P_{i,j}^4 &= (x_i, y_j) - h' r_2. \end{aligned}$$

Here h' is the new stencil length. A suitable value will be assigned to it in Chapter 4. As it will be made more clear in Chapter 4, the stencil length h' needs to be greater than h and hence the name *wide stencil* method. In contrast, the wide stencil method by Oberman et al [4], [11], [17] refers to the use of many neighbours, not all of which are the nearest ones.

Solving (2.6) is then equivalent to solving

$$\min_{Q \in Z} \left\{ \mathcal{L}_T^Q v - 2\sqrt{a(1-a)f} \right\} = 0, \tag{3.5}$$

where $v = v(w, z)$ is the representation of u in its new coordinates (w, z).

Let (w_i, z_j) be the grid points of the new coordinate plane. Note that since (x_i, y_j) is the rotation center, we have $(x_i, y_j) = (w_i, z_j)$. We discretize $\mathcal{L}_T^Q$ by central differencing:

$$\mathcal{L}_T^Q v(w_i, z_j) \approx \ d'_{11} \left[\frac{u(P_{i,j}^2) + u(P_{i,j}^4) - 2\mathcal{U}_{i,j}}{(h')^2} \right] + d'_{22} \left[\frac{u(P_{i,j}^1) + u(P_{i,j}^3) - 2\mathcal{U}_{i,j}}{(h')^2} \right]. \tag{3.6}$$

3.3 Bilinear Interpolation of points from the Wide Stencil Method

As shown in Figure 3.1, the stencil points $P_{i,j}^m$, $m = 1, \dots 4$ do not in general lie on the original grid points. Let us zoom into the point $P_{i,j}^4$ in Figure 3.2. $P_{i,j}^4$ lies in the grid square of its neighbours $(p_4 + s, q_4 + t)$, $s, t = 0, 1$. The value of u at $P_{i,j}^4$ can be interpolated from the values at the four points. Let $\mathcal{I}_h$ be a bilinear interpolation operator on a domain with grid size h. To approximate $u(P_{i,j}^4)$, we have the following form

$$\mathcal{I}_h u(P_{i,j}^4) = \sum_{\substack{s=0,1 \\ t=0,1}} \omega_{ij}^{p_4+s, q_4+t} u(p_4 + s, q_4 + t). \tag{3.7}$$

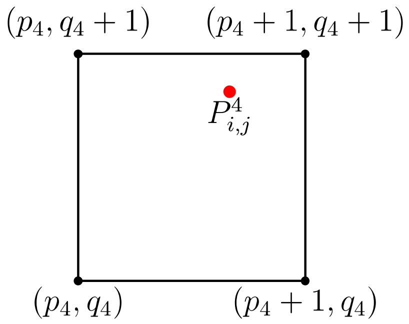


Figure 3.2: Location of $P^4_{i,j}$ and its 4 nearest neighbours.

Combining the above, we have:

$$\begin{aligned} u(P^4_{i,j}) &\approx \mathcal{J}_h u(P^4_{i,j}) \\ &= \sum_{\substack{s=0,1\\t=0,1}} \omega_{ij}^{p_4+s,q_4+t} u(p_4+s, q_4+t) \\ &= \sum_{\substack{s=0,1\\t=0,1}} \omega_{ij}^{p_4+s,q_4+t} \mathcal{U}_{p_4+s,q_4+t}. \end{aligned} \tag{3.8}$$

From (3.6), we have

$$\begin{aligned} \mathcal{L}^Q_T v(w_i, z_j) \approx \; & d'_{11}\left[\frac{\mathcal{J}_h u(P^2_{i,j}) + \mathcal{J}_h u(P^4_{i,j}) - 2\mathcal{U}_{i,j}}{(h')^2}\right] + d'_{22}\left[\frac{\mathcal{J}_h u(P^1_{i,j}) + \mathcal{J}_h u(P^3_{i,j}) - 2\mathcal{U}_{i,j}}{(h')^2}\right] \\ = & \frac{d'_{11}}{(h')^2}\mathcal{J}_h u(P^1_{i,j}) + \frac{d'_{11}}{(h')^2}\mathcal{J}_h u(P^3_{i,j}) + \frac{d'_{22}}{(h')^2}\mathcal{J}_h u(P^2_{i,j}) + \frac{d'_{22}}{(h')^2}\mathcal{J}_h u(P^4_{i,j}) \\ & - 2\frac{d'_{11}+d'_{22}}{(h')^2}\mathcal{U}_{i,j}. \end{aligned} \tag{3.9}$$

3.4 Points near the Boundary

The previous section has covered the case where $P^m_{i,j}$ is inside the computational domain. When it falls outside, we would need to shrink the point back to the boundary by an appropriate distance h^*, i.e. find h^* such that $(x_i, y_j) \pm h^* r_k \in \partial\Omega$.

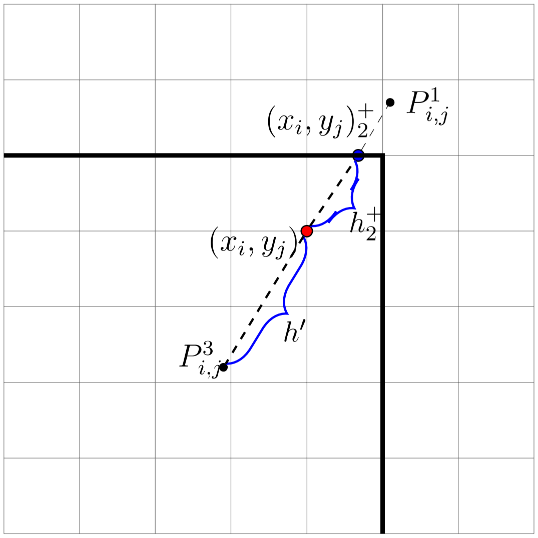


Figure 3.3: Shrinking of stencil point $P^1_{i,j}$

For $k = 1, 2$, define $(x_i, y_j)^+_k = (x_i, y_j) + h^+_k r_k$ to be the updated stencil point positions in the direction of r_k and $(x_i, y_j)^-_k = (x_i, y_j) - h^-_k r_k$, in the negative direction. Here $h^\pm_k$, $k = 1, 2$ are the values assigned to h^*.

Consider Figure 3.3. Let $k = 1$, $P^3_{i,j}$ lies in the computational grid, so $(x_i, y_j)^-_1 = P^3_{i,j}$. However $P^1_{i,j}$ falls outside. We will shrink the stencil distance from the default h' to $h^* = h^+_2$ such that the new stencil point $(x_i, y_j)^+_1$ lies on the boundary. Other cases are treated similarly, see Algorithm 1.

After obtaining h^+_k, h^-_k from Algorithm 1, we use central differencing to approximate the second derivatives $\frac{\partial^2 v}{\partial w^2}$ or $\frac{\partial^2 v}{\partial z^2}$ of $\mathcal{L}^Q_T$ to get (3.10). It is easy to see that (3.10) reduces to (3.9) when

$$h^+_k = h^-_k = h'$$

is satisfied. If we let L^Q_W be the discretized form of the operator $\mathcal{L}^Q_T$, it can be written compactly as

Algorithm 1 Shrink points to boundary for grid point $(x_i, y_j), k = 1$ or 2

```
Let $(x_i, y_j)_k^+ = (x_i, y_j) + h' r_k$ and $h_k^+ = h'$
if $(x_i, y_j)_k^+ \notin \Omega$ then
    solve $(x_i, y_j) + h^* r_k = (x_b, y_b)$ for $h^*$, such that $(x_b, y_b) \in \partial\Omega$
    $h_k^+ = h^*$
    $(x_i, y_j)_k^+ = (x_b, y_b)$
end if
Let $(x_i, y_j)_k^- = (x_i, y_j) - h' r_k$ and $h_k^- = h'$
if $(x_i, y_j)_k^- \notin \Omega$ then
    solve $(x_i, y_j) - h^* r_k = (x'_b, y'_b)$ for $h^*$, such that $(x'_b, y'_b) \in \partial\Omega$
    $h_k^- = h^*$
    $(x_i, y_j)_k^- = (x'_b, y'_b)$
end if
The second derivative terms $\frac{\partial^2 v}{\partial w^2}$ or $\frac{\partial^2 v}{\partial z^2}$ are approximated as
```

$$\frac{\frac{\mathcal{J}_h u((x_i,y_j)_k^-)-\mathcal{U}_{i,j}}{h_k^-} + \frac{\mathcal{J}_h u((x_i,y_j)_k^+)-\mathcal{U}_{i,j}}{h_k^+}}{\frac{h_k^+ + h_k^-}{2}} \tag{3.10}$$

$$
\begin{aligned}
L_W^Q v(w_i, z_j) = {} & \frac{2d'_{11}}{(h_1^+ + h_1^-)h_1^-}\mathcal{J}_h u((x_i, y_j)_1^-) + \frac{2d'_{11}}{(h_1^+ + h_1^-)h_1^+}\mathcal{J}_h u((x_i, y_j)_1^+) \\
& + \frac{2d'_{22}}{(h_2^+ + h_2^-)h_2^-}\mathcal{J}_h u((x_i, y_j)_2^-) + \frac{2d'_{22}}{(h_2^+ + h_2^-)h_2^+}\mathcal{J}_h u((x_i, y_j)_2^+) \\
& - 2\left(\frac{d'_{11}}{(h_1^+ h_1^-)} + \frac{d'_{22}}{(h_2^+ h_2^-)}\right)\mathcal{U}_{i,j},
\end{aligned}
\tag{3.11}
$$

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

In addition, it is worth noting that if (x_i, y_j) is a point near the corners of the square domain, then more than one stencil point will fall outside the domain. This creates problems in the consistency of L_W^Q approximation. However, in Chapter 4, we will see that by choosing an appropriate h' and together with Algorithm 1, we can still retain consistency.

3.5 The Matrix form of the Discretized Equations

From(3.5) and (3.11), (3.5) has the following discretized form for all grid points $(x_i, y_j) \in \Omega$:

$$
\min_{Q \in Z_h} \left\{ L_W^Q \mathcal{U}_{i,j} - 2\sqrt{a(1-a)f(x_i, y_j)} \right\} = 0,
\tag{3.12}
$$

where Z_h is the set of controls discretized to order h to retain consistency. From this equation, we will compute numerically the values of $\mathcal{U}_{i,j}$ through assembling a matrix where each row represents one grid point. This is shown in detail below.

Let us fix the grid point (x_i, y_j), observing from (3.11), we see that its corresponding entry, $\mathbf{L}_{l,l}^Q, l = i + (j-1)n$ is the coefficient of the term $\mathcal{U}_{i,j}$ above, i.e.,

$$
\mathbf{L}_{l,l}^Q = -2\left(\frac{d'_{11}}{(h_1^+ h_1^-)} + \frac{d'_{22}}{(h_2^+ h_2^-)}\right).
$$

From the previous two sections, stencil points of point (x_i, y_j) is either inside Ω or outside it. It falls into the two cases below and for simplicity, let us consider the stencil point $P_{i,j}^4$. Recall that its updated position is $(x_i, y_j)_2^-$. The other cases are treated similarly.

Case 1: $(x_i, y_j)_2^- \in \Omega$.

From (3.8) and (3.9), we have

$$\mathbf{L}^{Q}_{l,k} = \frac{d'_{11}}{(h')^2}\omega_{ij}^{p_4+s,q_4+t}, \qquad l = i + (j-1)n, \quad k = p_4 + s + (q_4 + t - 1)n, \quad s,t = 0,1.$$

In general, we have

$$\mathbf{L}^{Q}_{l,k} = \begin{cases} \frac{d'_{11}}{(h')^2}\omega_{ij}^{p_4+s,q_4+t} & \text{if } k = p_m + s + (q_m + t - 1)n, \quad s,t = 0,1, \quad m = 1,3, \\ \frac{d'_{22}}{(h')^2}\omega_{ij}^{p_4+s,q_4+t} & \text{if } k = p_m + s + (q_m + t - 1)n, \quad s,t = 0,1, \quad m = 2,4, \\ 0 & \text{otherwise.} \end{cases} \tag{3.13}$$

To handle cases where we have to use a boundary value (instead of the computational domain) from the Dirichlet condition, we need to define the vector $\mathbf{B}^Q$. First, let

$$\chi_{ij}^m = \begin{cases} 1 & \text{if } P_{i,j}^m \notin \Omega, \\ 0 & \text{otherwise.} \end{cases}$$

be the indicator function of whether the stencil points $P_{i,j}^m$ fall outside Ω. Define

$$\begin{aligned} \mathbf{B}^Q_I = {} & \chi_{ij}^1 \tfrac{d'_{11}}{(h_1^+ + h_1^-)h_1^-} u(P_{i,j}^1) + \chi_{ij}^2 \tfrac{d'_{11}}{(h_1^+ + h_1^-)h_1^+} u(P_{i,j}^2) + \chi_{ij}^3 \tfrac{d'_{22}}{(h_2^+ + h_2^-)h_2^-} u(P_{i,j}^3) \\ & + \chi_{ij}^4 \tfrac{d'_{22}}{(h_2^+ + h_2^-)h_2^+} u(P_{i,j}^4). \end{aligned} \tag{3.14}$$

Only when $P_{i,j}^m$ falls outside Ω will its corresponding term be added to B_I^Q.

Case 2: $(x_i, y_j)_2^- \in \partial\Omega$

In this case, we have $\chi_{ij}^4 = 1$ in (3.14) above. Other instances of m are treated similarly and B_I^Q is updated.

Finally we are now in the position to describe our matrix system explicitly. Let

$$\begin{aligned} Q^* &= (\, a^*, \theta^* \,) \\ &\in arg \min_{Q \in Z} \left\{ [\, \mathbf{L}^Q \mathbf{U} + \mathbf{B^Q} \,]_I \right\}. \end{aligned}$$

Then, the matrix entries on the l-th row, $[\mathbf{L}^{Q^*}\mathbf{U}]_l$ where $l = i + (j-1)n, \ i,j = 1,\ldots,n$ is given by the terms of $L_W^{Q^*}\mathcal{U}_{i,j}$, i.e. (3.13) above. The right hand side of (3.1) is given by

$$\mathbf{F}_I^{Q^*} = \mathbf{B}_I^{Q^*} + 2\sqrt{a^*(1-a^*)f(x_i,y_j)}, \tag{3.15}$$

which is the sum of vector $\mathbf{B}_I^Q$ and the constant term in (3.5).

Hence we have constructed the linear system (3.1) as desired at the beginning of this chapter.

Chapter 4

Convergence to the Viscosity Solution

As mentioned in previous chapters, we are interested in computing the viscosity solution to (2.1). In [3], a sufficient condition which guarantees convergence to the viscosity solution is provided. We give a proof that our numerical scheme given in chapter 3 satisfies all of the requirements.

For clarity, let us first define the following notations.

$$\begin{aligned} \mathbf{x} &= (x, y), \\ D^2u(\mathbf{x}) &= \left(\begin{array}{cc} \frac{\partial^2 u}{\partial x^2} & \frac{\partial^2 u}{\partial x \partial y} \\ \frac{\partial^2 u}{\partial x \partial y} & \frac{\partial^2 u}{\partial y^2} \end{array} \right). \end{aligned}$$

Then the value function (3.5) is denoted by

$$\mathcal{F}u = \mathcal{F}(\mathbf{x}, u(\mathbf{x}), D^2u(\mathbf{x})) = 0, \tag{4.1}$$

where

$$\mathcal{F}u = \min_{Q \in Z} \left\{ \mathcal{L}_T^Q u - 2\sqrt{a(1-a)f} \right\}. \tag{4.2}$$

The discrete value function (3.12) will be denoted by

$$\mathrm{K}\Big(h, (x_i, y_j), \mathcal{U}_{i,j}, \{\mathcal{U}_{p,q}\}_{p \neq i \, or \, q \neq j}\Big) = 0, \tag{4.3}$$

for all $i, j = 1, \dots n$, where

$$\mathrm{K}\bigg(h, (x_i, y_j), \mathcal{U}_{i,j}, \{\mathcal{U}_{p,q}\}_{p\neq i\, or\, q\neq j}\bigg) = \min_{Q\in Z_h} \left\{ L_W^Q \mathcal{U}_{i,j} - 2\sqrt{a(1-a)f_{i,j}} \right\}. \tag{4.4}$$

Here $f_{i,j} = f(x_i, y_j)$ and h is the mesh size.

The convergence theorem in [3] that allows us to guarantee convergence to the viscosity solutions is given below. Interested readers can refer to [3].

Theorem 4.1. *Consider a degenerate elliptic equation for which there exist unique viscosity solutions. A consistent (in the viscosity sense), l_∞ stable and monotone approximation scheme converges on compact subsets to the viscosity solution.*

4.1 Consistency

To be able to state our analysis in a rigorous way, we need the following definition of the supremum and infimum of a function f which may not always be a continuous function.

Definition 4.1. *Let f be a real-valued function on $\Omega \subseteq \mathbb{R}^2$. The* ***upper semi-continuous envelope*** *of f, f^*, is*

$$f^* = \lim_{r\to 0} \sup \{ f(y) \,|\, y \in B(x, r) \cap \Omega \},$$

where $B(x, r)$ denotes the open ball centered at x with radius r.

The **lower semi-continuous envelope** of f, denoted by f_*, can be defined similarly. Based on the meaning of a viscosity solution in Definition 2.1, we give below the corresponding definition of *consistency*:

Definition 4.2. *[2] A numerical scheme is* ***consistent in the viscosity sense*** *if for any function $\phi \in \mathrm{C}^\infty$ with $\phi_{i,j} = \phi((x_i, y_j))$ and for all $\mathbf{x} \in \Omega, (x_i, y_j) \in \Omega$, we have*

$$\limsup_{\substack{h\to 0\\ \xi\to 0\\ (x_i,y_j)\to\mathbf{x}}} \mathrm{K}\bigg(h, (x_i, y_j), \phi_{i,j} + \xi, \{\phi_{p,q} + \xi\}_{p\neq i\, or\, q\neq j}\bigg) \leq \mathcal{F}^*(\mathbf{x}, \phi(\mathbf{x}), D^2\phi(\mathbf{x})) \tag{4.5}$$

and

$$\liminf_{\substack{h\to 0\\ \xi\to 0\\ (x_i,y_j)\to\mathbf{x}}} \mathrm{K}\Big(h,(x_i,y_j),\phi_{i,j}+\xi,\{\phi_{p,q}+\xi\}_{p\neq i\,or\,q\neq j}\Big)\geq \mathcal{F}_*(\mathbf{x},\phi(\mathbf{x}),D^2\phi(\mathbf{x})), \tag{4.6}$$

where h and ξ are arbitrary small constants independent of $\mathbf{x}$.

It has similar meanings to the standard consistency definition, that is, the discretization error will be negligible as mesh size h decreases.

To prove that our numerical scheme is consistent, we first prove that it is so locally.

Lemma 4.1. *(local consistency conditions) Suppose the mesh size is h, and the control discretization is of order $\mathrm{O}(h)$, and if we take the stencil length h' (defined in section 3.2), to be $\sqrt{h}$, then for any function $\phi\in \mathrm{C}^\infty$ and using the notations above, we have that*

$$\mathrm{K}\Big(h,(x_i,y_j),\phi_{i,j}+\xi,\{\phi_{p,q}+\xi\}_{p\neq i\,or\,q\neq j}\Big) = \begin{cases}\mathcal{F}\phi_{i,j}+\mathrm{O}(h)+\mathrm{O}(\xi), & P^m_{i,j}\in\Omega\ \forall m=1,\ldots,4.\\ \mathcal{F}\phi_{i,j}+\mathrm{O}(\sqrt{h})+\mathrm{O}(\xi), & \textit{otherwise.}\end{cases} \tag{4.7}$$

Proof. <u>**Case 1:** $P^m_{i,j}\in\Omega$ for all m.</u>

In this case, we have $h_1^+=h_1^-=h_2^+=h_2^-=\sqrt{h}$. For all (x_i,y_j) belonging to Ω, we will show:

$$\mathrm{L}^Q_W\phi_{i,j}=\mathcal{L}^Q_T\phi_{i,j}+\mathrm{O}(h).$$

$$
\begin{aligned}
\mathrm{L}_W^Q\phi_{i,j} - \mathcal{L}_T^Q\phi_{i,j} =\ & d'_{11}\left[\frac{\mathcal{J}_h\phi((w_i,z_j)+\sqrt{h}r_1)+\mathcal{J}_h\phi((w_i,z_j)-\sqrt{h}r_1)-2\phi_{i,j}}{h}\right] \\
&+ d'_{22}\left[\frac{\mathcal{J}_h\phi_{i,j}(\,(w_i,z_j)+\sqrt{h}r_2\,)+\mathcal{J}_h\phi_{i,j}(\,(w_i,z_j)-\sqrt{h}r_2\,)-2\phi_{i,j}}{h}\right] \\
&- d'_{11}\,\frac{\partial^2\phi_{i,j}}{\partial w^2} - d'_{22}\,\frac{\partial^2\phi_{i,j}}{\partial z^2} \\
=\ & d'_{11}\left[\frac{\phi((x_i,y_j)+\sqrt{h}e_1)+\mathrm{O}(h^2)+\phi((x_i,y_j)-\sqrt{h}e_1)+\mathrm{O}(h^2)-2\phi_{i,j}}{h}\right] \\
&+ d'_{22}\left[\frac{\phi((x_i,y_j)+\sqrt{h}e_2)+\mathrm{O}(h^2)+\phi((x_i,y_j)-\sqrt{h}e_2)+\mathrm{O}(h^2)-2\phi_{i,j}}{h}\right] \\
&- d'_{11}\,\frac{\partial^2\phi_{i,j}}{\partial x^2} - d'_{22}\,\frac{\partial^2\phi_{i,j}}{\partial y^2} \\
=\ & \mathrm{O}(h)+\mathrm{O}(h) \\
=\ & \mathrm{O}(h),
\end{aligned}
$$

where e_1, e_2 are the canonical axes in the new coordinate grid. Note that the second equality follows from the $\mathrm{O}(h^2)$ accuracy of the bilinear interpolation and the second last equality follows from the error when central differencing with stencil length $\sqrt{h}$ was used to approximate the second derivatives. It remains to show that (4.7) is true. For all (x_i, y_j) such that $P^m_{i,j} \in \Omega$,

$$
\begin{aligned}
\mathrm{K}\Big(h,(x_i,y_j),\phi_{i,j}+\xi,\{\phi_{p,q}+\xi\}_{p\neq i\,or\,q\neq j}\Big) &= \min_{Q\in Z_h}\left\{\mathcal{L}_T^Q\phi_{i,j}-2\sqrt{a(1-a)f_{i,j}}\right\}+\mathrm{O}(h)+\mathrm{O}(\xi) \\
&= \min_{Q\in Z}\left\{\mathcal{L}_T^Q\phi_{i,j}-2\sqrt{a(1-a)f_{i,j}}\right\}+\mathrm{O}(h)+\mathrm{O}(h)+\mathrm{O}(\xi) \\
&= \mathcal{F}\phi_{i,j}+\mathrm{O}(h)+\mathrm{O}(\xi).
\end{aligned}
\tag{4.8}
$$

The first equality follows from the above analysis and the second, from the discretization of Z.

Case 2: $\exists P^m_{i,j} \notin \Omega$.

For all (x_i, y_j) belonging to Ω, we will show:

$$\mathrm{L}_W^Q \phi_{i,j} = \mathcal{L}_T^Q \phi_{i,j} + \mathrm{O}(\sqrt{h}).$$

Consider the approximation for $\frac{\partial^2 \phi_{i,j}}{\partial w^2}$ and suppose $m = 1$ (or 3), ($\frac{\partial^2 \phi_{i,j}}{\partial z^2}$ is just the same but $m = 2$ or 4 replaced). With an abuse of notation here, write $\phi(w_i)$ as $\phi(w_i, z_j)$ and similarly for $\phi(x_i)$. Let us first consider the following analysis based on the Taylor series expansion:

$$\begin{aligned}
\frac{\phi(x_i - h_1^- e_1) - \phi(x_i)}{h_1^-} &= -\phi'(x_i) + \frac{h_1^-}{2!}\phi''(x_i) - \frac{(h_1^-)^2}{3!}\phi^{(3)}(x_i) + \frac{(h_1^-)^3}{4!}\phi^{(4)}(x_i), \\
\frac{\phi(x_i + h_1^+ e_1) - \phi(x_i)}{h_1^+} &= \phi'(x_i) + \frac{h_1^+}{2!}\phi''(x_i) + \frac{(h_1^+)^2}{3!}\phi^{(3)}(x_i) + \frac{(h_1^+)^3}{4!}\phi^{(4)}(x_i).
\end{aligned}$$

If we sum the above, the term on the left hand side is in fact the discretization of the second order derivatives from Algorithm 1, i.e. (3.10) which is the local truncation error:

$$\frac{\frac{\phi(x_i - h_1^- e_1) - \phi(x_i)}{h_1^-} + \frac{\phi(x_i + h_1^+ e_1) - \phi(x_i)}{h_1^+}}{\frac{h_1^- + h_1^+}{2}} - \phi''(x_i) = \frac{h_1^+ - h_1^-}{3}\phi^{(3)}(x_i) + \frac{(h_1^+)^2 - h_1^- h_1^+ + (h_1^-)^2}{12}\phi^{(4)}(x_i).$$

There are several cases to consider. Firstly, suppose that $h_1^+ = h_1^-$, then from the above equation, the local truncation error is $\mathrm{O}((h_1^-)^2)$, but since h_1^- is order h, we have the error to be $\mathrm{O}(h^2)$. The second case is when $h_1^+ \neq h_1^-, h_1^+ \,(\text{or}\, h_1^-) = \sqrt{h}$. In this case, $h_1^- = \mathrm{O}(h)$, so the local error is $\mathrm{O}(h - \sqrt{h})$, which is just $\mathrm{O}(\sqrt{h})$. Finally, suppose $h_1^+ \neq h_1^-, h_1^- \neq \sqrt{h}, h_1^+ \neq \sqrt{h}$, then $h_1^- = \mathrm{O}(h)$ and $h_1^+ = \mathrm{O}(h)$ which makes the local error to be order $\mathrm{O}(h)$. Combining the three cases, we have:

$$\frac{\frac{\phi(x_i - h_1^- e_1) - \phi(x_i)}{h_1^-} + \frac{\phi(x_i + h_1^+ e_1) - \phi(x_i)}{h_1^+}}{\frac{h_1^- + h_1^+}{2}} - \phi''(x_i) = \mathrm{O}(\sqrt{h}).$$

The intermediate steps above can all be verified if one works out the details of Algorithm 1. From the above analysis, we have the following:

$$
\begin{aligned}
\mathrm{L}^Q_W\phi(w_i) - \mathcal{L}^Q_T\phi(w_i) &= d'_{11}\left[\frac{\mathcal{J}_h\phi(w_i + h_1^+ r_1) + \mathcal{J}_h\phi(w_i - h_1^- r_2) - 2\phi(w_i)}{\frac{h_1^- + h_1^+}{2}}\right] - d'_{11}\,\frac{\partial^2\phi(w_i)}{\partial w^2} \\
&= d'_{11}\left[\frac{\phi(x_i + h_1^+ e_1) + \mathrm{O}(h^2) + \phi(x_i - h_1^- e_1) + +\mathrm{O}(h^2) - 2\phi(x_i)}{\frac{h_1^- + h_1^+}{2}}\right] - d'_{11}\,\frac{\partial^2\phi(x_i)}{\partial x^2} \\
&= \mathrm{O}(h) + \mathrm{O}(\sqrt{h}) \\
&= \mathrm{O}(\sqrt{h}).
\end{aligned}
$$

To show that (4.7) is true, we follow similar steps as in (4.8) above. Hence, our lemma is proved. □

Proposition 4.1. *Suppose the numerical scheme 4.3 satisfies the conditions in Lemma 4.1, then it is consistent in the viscosity sense.*

Proof. One may follow similar steps in [14] . □

4.2 Stability

Stability of a numerical scheme is when it produces an approximiate solution that is bounded independent of the mesh size h. It turns out here in this case that stability has large relations with the M-matrix property of the matrix $\mathbf{L}^Q$ constructed in the previous chapter.

Definition 4.3. *Let A be an $n \times n$ matrix. It is an **M-matrix** if*

i. $a_{ii} > 0$ *for all* i,

ii. $a_{ij} \leq 0$ *for all* $i \neq j$,

iii. A *is nonsingular,*

iv. $A^{-1} \geq 0$.

A sufficient condition will be given below, let us first define the following:

Definition 4.4. *An $n \times n$ matrix A is irreducible if there exists an $n \times n$ permutation matrix P such that*

$$PAP^T = \begin{pmatrix} A_{11} & A_{12} \\ \mathrm{O} & A_{22} \end{pmatrix},$$

where A_{11} is an $r \times r$ submatrix and A_{22} is an $(n-r)\times(n-r)$ submatrix, where $1 \leq r < n$. If no such permutation matrix exists, then A is irreducible.

Proposition 4.2. *(Axelsson, 1996) [1] If A is a real $n \times n$ matrix and satisfies*

1. *$a_{ii} > 0$ for all i,*
2. *$a_{ij} \leq 0$ for all $i \neq j$,*
3. *A is irreducible,*
4. *diagonally dominant with at least one i strictly diagonally dominant.*

then A is an M-matrix.

Lemma 4.2. *If L_W^Q is defined as in (3.11) and if a linear interpolation operator $\mathcal{J}_h$ is used in (3.7), such that*

$$\begin{aligned} \omega_{ij}^{p_m+s,q_m+t} &\geq 0 \qquad \forall m = 1, \ldots, 4 \ , s,t = 0,1, \\ \sum_{\substack{s=0,1 \\ t=0,1}} \omega_{ij}^{p_m+s,q_m+t} &= 1, \end{aligned} \tag{4.9}$$

then $\mathbf{L}^Q$ in (3.15) is an M-matrix for all $Q \in Z$.

Proof. We prove that L_W^Q is an M-matrix by verifying the four conditions in the proposition above. From (3.13) and (3.5), properties 1, 2 can be verified if we can show that $\mathbf{L}_{l,l}^Q$ never vanishes for all l. This is a bit technical and we shall outline the idea here instead. Notice that if the optimal control parameter a^* (c.f. 2.5) is neither 0 nor 1, then the matrix D in 2.1 is strictly positive definite, hence $d'_{11} \neq 0$ and $d'_{22} \neq 0$ by definition of positive definiteness (for example, to verify the case for d'_{11}, note that $d'_{11} = \begin{pmatrix} \sin(\phi) \\ \cos(\phi) \end{pmatrix}' D \begin{pmatrix} \sin(\phi) \\ \cos(\phi) \end{pmatrix}$). However, when $a = 0$, we claim that if $d'_{11} = 0$ then $d'_{22} \neq 0$ and vice versa. For $a = 1$, it is similar. If what we have just claimed is true, then by (3.5), $\mathbf{L}_{l,l}^Q$ never vanishes. So when

$a = 0$, it is not hard to verify that $d_{11} = \sin(\theta)^2, d_{22} = \cos(\theta)^2$ and $d_{12} = \cos(\theta)\sin(\theta)$. If it occurs that $d'_{11} = 0$, then by (3.4), we have

$$-2\cos(\theta)\sin(\theta)\cos(\phi)\sin(\phi) = \sin(\theta)^2\cos(\phi)^2 + \cos(\theta)^2\sin(\phi)^2,$$

then also by (3.4),

$$d'_{22} = \sin(\theta)^2\sin(\phi)^2 + \cos(\theta)^2\cos(\phi)^2 + \sin(\theta)^2\cos(\phi)^2 + \cos(\theta)^2\sin(\phi)^2 = 1 \neq 0.$$

We will concentrate on problems such that the resulting matrix L_W^Q satisfies property 3. Let $l = i + (j-1)n$, now we prove property 4. We have the following two cases:

Case 1: $P_{i,j}^m \in \Omega$ for all m.

From (3.13),

$$\begin{aligned}
|\left[\mathbf{L}^Q\right]_{l,l}| - \sum_{k\neq l} |\left[\mathbf{L}^Q\right]_{l,k}| &= 2\Big(\frac{d'_{11}}{h} + \frac{d'_{22}}{h}\Big) - \frac{d'_{11}}{h}\sum_{\substack{s=0,1\\t=0,1}} \omega_{ij}^{p_1+s,q_1+t} - \frac{d'_{11}}{h}\sum_{\substack{s=0,1\\t=0,1}} \omega_{ij}^{p_2+s,q_2+t} \\
&\quad - \frac{d'_{22}}{h}\sum_{\substack{s=0,1\\t=0,1}} \omega_{ij}^{p_3+s,q_3+t} - \frac{d'_{22}}{h}\sum_{\substack{s=0,1\\t=0,1}} \omega_{ij}^{p_4+s,q_4+t} \\
&= 0.
\end{aligned}$$

Case 2: $\exists P_{i,j}^m \notin \Omega$.

In general, we have

$$\begin{aligned}
|\left[\mathbf{L}^Q\right]_{l,l}| - \sum_{k\neq l} |\left[\mathbf{L}^Q\right]_{l,k}| &= 2\Big(\frac{d'_{11}}{h_1^+h_1^-} + \frac{d'_{22}}{h_2^+h_2^-}\Big) - \chi_{ij}^1 \tfrac{2d'_{11}}{(h_1^+ + h_1^-)h_1^-} - \chi_{ij}^2 \tfrac{2d'_{11}}{(h_1^+ + h_1^-)h_1^+} \\
&\quad - \chi_{ij}^3 \tfrac{2d'_{22}}{(h_2^+ + h_2^-)h_2^-} - \chi_{ij}^4 \tfrac{2d'_{22}}{(h_2^+ + h_2^-)h_2^+} \\
&\geq 2\Big(\frac{d'_{11}}{h_1^+h_1^-} + \frac{d'_{22}}{h_2^+h_2^-}\Big) - 2\Big(\frac{d'_{11}}{h_1^+h_1^-} + \frac{d'_{22}}{h_2^+h_2^-}\Big) \\
&= 0.
\end{aligned}$$

However, note that in the boundary points (e.g. when $l = 1, n^2$), there must be a stencil point falling outside the computational domain and hence $\chi_{ij}^m = 0$ for some $m = 1, \ldots, 4$, e.g. let $m = 1$:

$$\begin{aligned}
|\,[\mathbf{L}^Q\,]_{l,l}\,| - \sum_{k\neq l} |\,[\mathbf{L}^Q\,]_{l,k}\,| &= 2(\frac{d'_{11}}{h_1^+ h_1^-} + \frac{d'_{22}}{h_2^+ h_2^-}) - \chi_{ij}^2 \tfrac{2d'_{11}}{(h_1^+ + h_1^-)h_1^+} \\
&\quad - \chi_{ij}^3 \tfrac{2d'_{22}}{(h_2^+ + h_2^-)h_2^-} - \chi_{ij}^4 \tfrac{2d'_{22}}{(h_2^+ + h_2^-)h_2^+} \\
&> 2(\frac{d'_{11}}{h_1^+ h_1^-} + \frac{d'_{22}}{h_2^+ h_2^-}) - 2(\frac{d'_{11}}{h_1^+ h_1^-} + \frac{d'_{22}}{h_2^+ h_2^-}) \\
&= 0.
\end{aligned}$$

So in cases like that , we have strict diagonal dominance on that row. Combining the two cases, property 4 is verified. □

From the proposition below we see that our scheme is a stable scheme. For the proof, interested readers may refer to [16].

Proposition 4.3. *If the conditions for Lemma 4.2 are satisfied then the discretization scheme (4.3) is l_∞ stable. And as mesh size $h \to 0$, we have*

$$\|\mathbf{U}\|_\infty \leq \max\left(\left\|\mathbf{U}^0\right\|_\infty, \, \|g\|_\infty\right),$$

where g is the given Dirichlet boundary condition of (2.1).

4.3 Monotonicity

Monotonicity is an essential requirement for our scheme to converge to the viscosity solution as mentioned earlier. We give the definition based on [16].

Definition 4.5. *The discrete scheme is* ***monotone*** *if $\mathcal{V}_{i,j} \geq \mathcal{U}_{i,j}$ for all i, j, we have*

$$\mathrm{K}\Big(h, (x_i, y_j), \mathcal{V}_{i,j}, \{\mathcal{V}_{p,q}\}_{p\neq i\, or\, q\neq j}\Big) \geq \mathrm{K}\Big(h, (x_i, y_j), \mathcal{U}_{i,j}, \{\mathcal{U}_{p,q}\}_{p\neq i\, or\, q\neq j}\Big).$$

Proposition 4.4. *If the scheme (4.3) satisfies the condition of Lemma 4.2, then our discretization consists of positive coefficients only and thus results in a monotone scheme.*

Proof. From (3.13), all the coefficients involved in the discretization are positive for all $Q \in Z$, since linear interpolation satisfying (4.9) was used and the coefficients d'_{11} and

d_{22}' are positive for all $Q \in Z$, $\phi \in [\frac{-\pi}{2}, \frac{\pi}{2}]$ (remark 2.1). Hence monotonicity follows from similar derivations in [10]. $\square$

Chapter 5

Numerical Results

In this chapter, we apply the proposed discretization method for the MAE on three examples with smooth to mildly singular solutions. The computations were performed on a Mac desktop with 2.8GHz Intel Core Duo processor and 4GB memory, using MATLAB running in Mac OS X.6.

All the examples were used by [4] and [6]. The method used in [6] converged for the first two but not for the last one. We will follow [4] and use the approximate solution to

$$u_{xx} + u_{yy} = \sqrt{2f},$$

as the initial estimation for all the examples. This is to minimize the time needed to obtain the final convergence result. In each iteration, we use policy iteration to numerically solve the HJB equation (3.5). It is an iterative process and at each iteration, the value function from the previous policy is updated and then an improved policy is found via the new value function. In theory, if enough iterations are evaluated, the optimal policy should converge to the optimal control of the HJB and the value function should converge towards its solution. In our numerical experiments, we will just fix the number of iterations due to the computational time using MATLAB.

Error is measured as the L_2 norm of the difference between the computed and the exact solutions. Formally, let $\mathbf{U}^h$ be the approximate solution and e^h be the error at the grid level h, then

$$e_h = h \left\| \mathbf{U}^h - u \right\|_2 .$$

The loglog plot of the graph of error versus mesh size h is plotted to analysis the convergence rate and the slope of the best fitted line is computed which represents this rate. In nearly all the examples, we observed convergence rates of linear convergence $O(h)$ or near linear convergence.

5.1 Smooth Examples

5.1.1 Example 1

Consider the problem

$$\mathcal{L}u = (1 + x^2 + y^2)\exp(x^2 + y^2) \quad in \quad \Omega, \tag{5.1}$$

$$u = \exp\left(\tfrac{1}{2}(x^2 + y^2)\right) \quad on \quad \partial\Omega, \tag{5.2}$$

where Ω is the square domain $[-1/2, 1/2] \times [-1/2, 1/2]$.

An exact solution is

$$u(x, y) = \exp\left(\tfrac{1}{2}(x^2 + y^2)\right). \tag{5.3}$$

Our method converges to the same numerical solution (5.3). A plot of the computed and exact solutions is given in Figures 5.1a and 5.1b respectively.

Figure 5.1c is a loglog plot of mesh size h versus the L_2 error and the slope of the best fitted line was found to be 1.02 which implies that the convergence rate is linear.

5.1.2 Example 2

Consider the problem

$$\mathcal{L}u = 2 \quad in \quad \Omega, \tag{5.4}$$

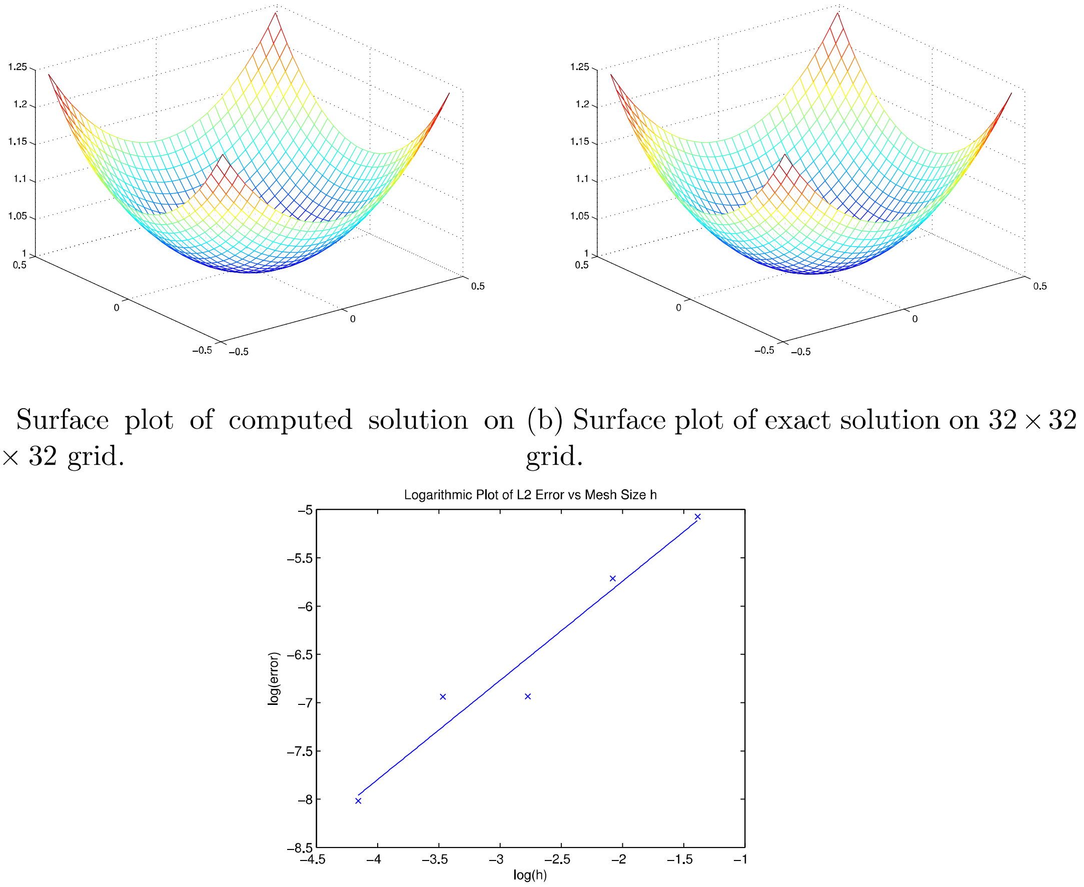
$$u = x^2 + y^2 \quad on \quad \partial\Omega,$$

where Ω is the square domain $[0, 1] \times [0, 1]$.

An exact solution is

$$u(x, y) = x^2 + y^2. \tag{5.5}$$

A plot of the computed and exact solutions is given in Figures 5.2a and 5.2b respectively. The computed solution closely approximates the exact solution (5.5). Figure 5.2c is a loglog



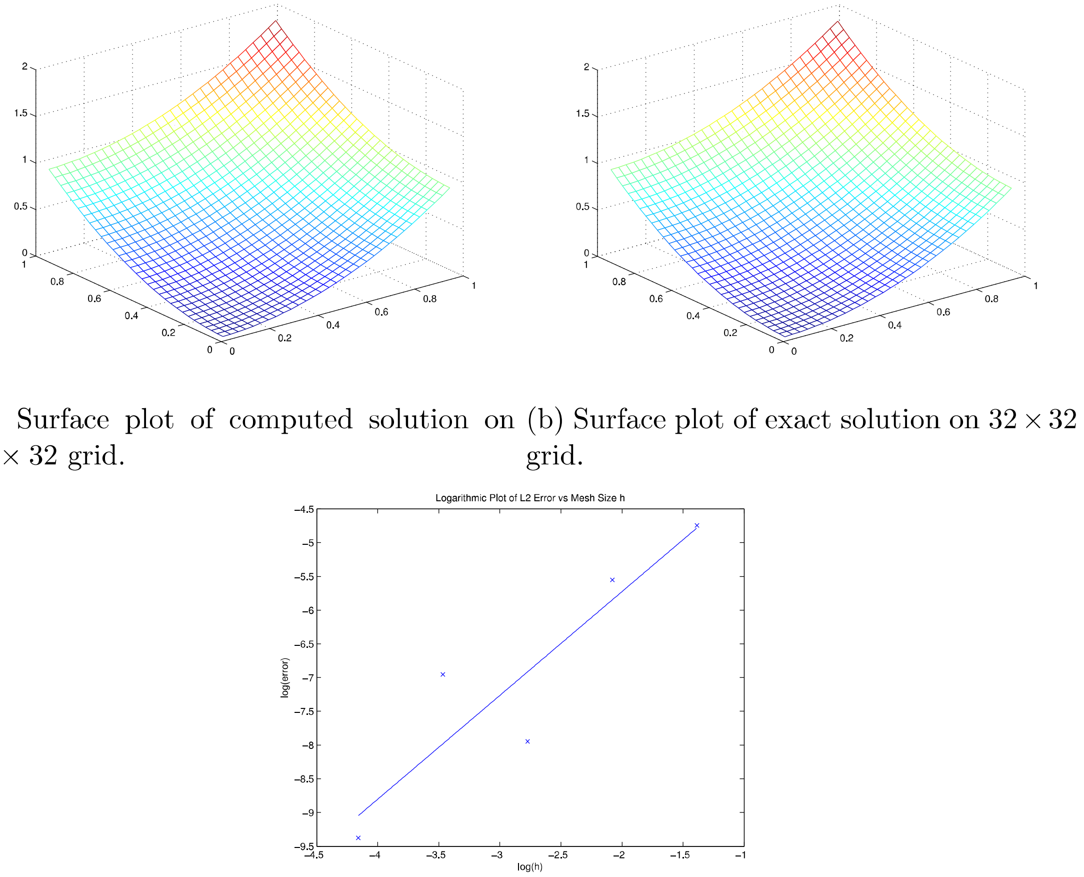
(a) Surface plot of computed solution on 32×32 grid.

(b) Surface plot of exact solution on 32×32 grid.

(c) Best fit line in the loglog plot of mesh size h vs error, slope of line is 1.02.

Figure 5.1: Plots of example 1

plot of the mesh size h versus the L_2 error and the slope of the line was found to be 1.51. Although the convergence was oscillatory, the overall trend implies that the convergence is approximately linear.



(a) Surface plot of computed solution on 32×32 grid.

(b) Surface plot of exact solution on 32×32 grid.

(c) Best fit line in the loglog plot of mesh size h vs error, slope of line is 1.51.

Figure 5.2: Plots of example 2

5.2 Non-smooth Solutions

5.2.1 Example 3

Consider the problem

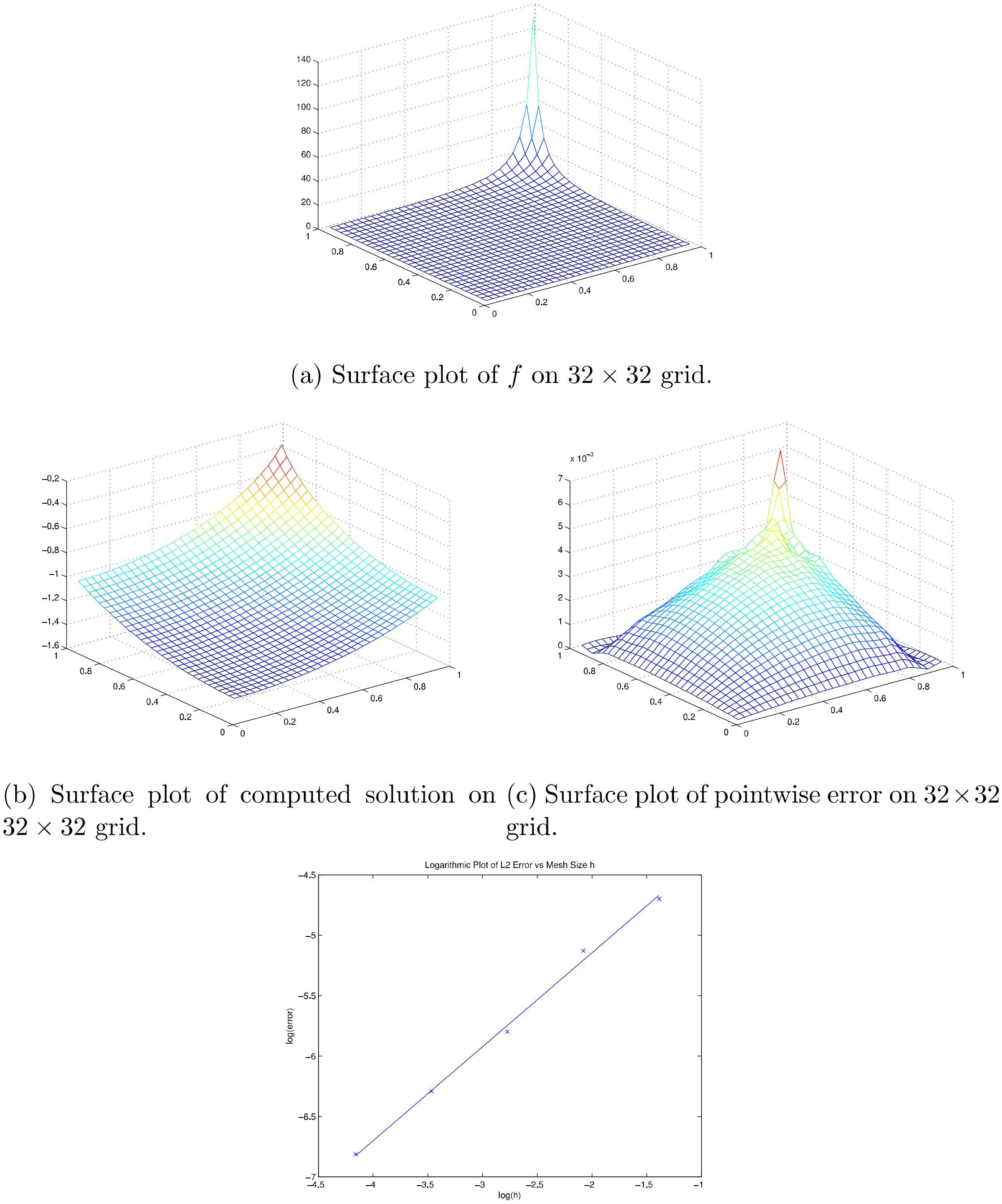
$$\begin{aligned} \mathcal{L}u &= \frac{2}{(2-x^2-y^2)^2} \quad in \quad \Omega, \\ u &= -\sqrt{2-x^2-y^2} \quad on \quad \partial\Omega, \end{aligned}$$

where Ω is the square domain $[0,1] \times [0,1]$.

An exact solution is

$$u(x,y) = -\sqrt{2-x^2-y^2}. \tag{5.6}$$

From Figure 5.3a, the gradient of f is unbounded at $(1,1)$, making the equation moderately singular. This example was also used by [6] and their method was known to diverge. A solution plot on grid 32×32 is plotted in Figure 5.3b. Figure 5.3d is a loglog plot of the mesh size h versus the L_2 error and the slope of the best fitted line was found to be approximately 0.82 which implies that the convergence is near linear. The error is mainly located at the region of blow up, see Figure 5.3c and does not affect the overall convergence of our solution as much as the method used in [6].



(a) Surface plot of f on 32×32 grid.

(b) Surface plot of computed solution on 32×32 grid.

(c) Surface plot of pointwise error on 32×32 grid.

(d) Best fit line in the loglog plot of mesh size h vs error, slope of line is 0.82.

Figure 5.3: Plots of example 3

Chapter 6

Conclusions

In this paper, we proposed a numerical scheme to compute the unique viscosity solution of the elliptic MAE with Dirichlet boundary condition. We first transformed the MAE to a Hamilton-Jacobian-Bellman equation whose objective function is a linear second order PDE coupled with non-linear control parameters. This allows ease of discretization. To obtain a monotone scheme, further work was done to transform it by a local grid rotation to eliminate the crossed derivatives terms, thus resulting in a wide stencil discretization. This caused new challenges such as the stencil points lying outside the computational domain, however we modified the algorithm in [16] to shrink the points back to its boundary and at the same time retained consistency of the discretization.

Our numerical method is also stable and monotone in addition to being consistent with the original MAE, hence by the Barles Souganidis convergence theorem, we proved that it converged to the unique viscosity solutions. In the analysis of consistency, we showed that the convergence rate was either linear or order $O(\sqrt{h})$.

The numerical experiments performed had solutions ranging from smooth to moderately singular. They all showed convergence and at a linear or near linear rate ($O(h)$) and the worser case of convergence at the boundary had minimal effect.

Directions for further work will be to design more efficient algorithms to solve the nonlinear HJB as policy iteration alone took too long for practical applications. To apply our numerical scheme to real life image registration problems especially medical imaging is also one direction of pursuit.

References

[1] Owe Axelsson. *Iterative solution methods.* Cambridge University Press, 1996. 26

[2] G. Barles. *Convergence of numerical schemes for degenerate parabolic equations arising in finance.* Cambridge University Press, Cambridge, 1997. 21

[3] G. Barles and P.E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Anal.*, 4:271–283, 1991. 3, 6, 20, 21

[4] Jean-David Benamou, Brittany D Froese, and Adam M Oberman. Two numerical methods for the elliptic monge-ampere equation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44(04):737–758, 2010. 13, 30

[5] Shiu-Yuen Cheng and Shing-Tung Yau. On the regularity of the monge-ampère equation det (2 u/ xi xj)= f (x, u). *Communications on Pure and Applied Mathematics*, 30(1):41–68, 1977. 6

[6] E.J. Dean and R. Glowinski. Numerical methods for fully nonlinear elliptic equations of the monge-ampere type. *Computer Methods in Applied Mechanics and Engineering*, 195:1344–1386, 2006. 3, 30, 34

[7] Glowinski R. Feng, X. and M. Neilan. Recent developments in numerical methods for fully nonlinear second order partial differential equations. *SIAM Review*, 55:205–267, 2013. ix, 3, 7

[8] X. Feng and M. Neilan. Mixed finite element methods for the fully nonlinear mongeampre equation based on the vanishing moment method. *SIAM Journal on Numerical Analysis*, 47(2):1226–1250, 2009. 3

[9] Bernd Fischer and Jan Modersitzki. Ill-posed medicinean introduction to image registration. *Inverse Problems*, 24(3):034008, 2008. ix, 2

[10] P. A. Forsyth and G. Labahn. Numerical methods for controlled Hamilton-Jacobi-Bellman partial differential equations in finance. *Journal of Computational Finance*, 11(2):1, 2007. 29

[11] Brittany D Froese and Adam M Oberman. Convergent finite difference solvers for viscosity solutions of the elliptic monge-ampere equation in dimensions two and higher. *SIAM Journal on Numerical Analysis*, 49(4):1692–1714, 2011. ix, 1, 3, 13

[12] C. E. Gutierrez. *The Monge Ampère Equation*. Birkhuser Mathematics, Basel, 2001. 6

[13] Steven Haker, Lei Zhu, Allen Tannenbaum, and Sigurd Angenent. Optimal mass transport for registration and warping. *International Journal of Computer Vision*, 60(3):225–240, 2004. 1

[14] Y. Huang and P. A. Forsyth. Analysis of a penalty method for pricing a guaranteed minimum withdrawal benefit (GMWB). *Journal of Numerical Analysis*, 32(1):320–351, 2012. 25

[15] N. V. Krylov. On control of the solution of a stochastic integral equation with degeneration. *Math. USSR Izv.*, 6(1):249, 1972. 7

[16] K. Ma and P.A. Forsyth. An unconditionally monotone numerical scheme for the two factor uncertain volatility model. 2014. 4, 28, 36

[17] A. M. Oberman. Wide stencil finite difference schemes for the elliptic monge-ampere equation and functions of the eigenvalues of the hessian. *Discrete Contin. Dyn. Syst. Ser. B*, 10(1):221–238, 2008. 13

[18] I. Smears. *Hamilton-Jacobi-Bellman Equations. Analysis and Numerical Analysis.* 7