FINITE-DIFFERENCE AND FINITE-ELEMENT SOLUTION OF BOUNDARY VALUE AND OBSTACLE PROBLEMS FOR THE HESTON OPERATOR

BY EDUARDO OSORIO

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Mathematics

Written under the direction of

Paul M. N. Feehan

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2014

ABSTRACT OF THE DISSERTATION

Finite-difference and finite-element solution of boundary value and obstacle problems for the Heston operator

by Eduardo Osorio

Dissertation Director: Paul M. N. Feehan

We develop finite-element and finite-difference methods for boundary value and obstacle problems for the elliptic Heston operator. For the finite-element method we first review existence and uniqueness results for these problems on weighted Sobolev spaces, where their variational formulations are formulated, and finite-dimensional subspaces are chosen to find approximating solutions, and obtain error estimates and numerical results. Similarly, for the finite-difference method, we start by reviewing the existence, uniqueness and regularity results on boundary value and obstacle problems on weighted Hölder spaces, then consider finite-difference operators, establish discrete maximum principles for them, and obtain error estimates and numerical results.

Acknowledgements

First, I would like to express my sincere gratitude to my thesis advisor Paul Feehan for his encouragement and support. I couldn't have done it without his guidance. Second, I would like to thank my benefactor Sergio Rodriguez for his generosity while I was in college. Without your financial help throughout my high level education I couldn't have got this far.

This work was done across five years, mostly during my weekends, while I kept a full time job and because of that I may have missed many special life moments of the ones that surround me, hence I would like to thank my family, friends and loved ones for being by my side with their words and actions of inspiration and motivation.

Dedication

I dedicate this work to my mom, the best example I have in my life of persistence and hard work, and to my aunt who I wish gets to share with me the feeling of accomplishment of completing this chapter of my life.

Table of Contents

Chapter 1

Introduction

Stochastic volatility processes are used in mathematical finance as models for asset prices when valuing derivative securities, such as options. In these models, the volatility of the underlying security is itself a random process.

Stochastic volatility processes are one way to address a defect in the Black-Scholes model [6], where the underlying volatility is constant over the life of the derivative and hence they cannot explain observed features of the implied volatility surface, such as volatility smile. Although the simpler diffusion model of Black-Scholes is still widely used, mostly due to the availability of closed-form analytical pricing formulas for many types of derivatives, by assuming that the volatility of the underlying price is a stochastic process rather than a constant, it becomes possible to value and hedge derivative prices more accurately.

The Heston process [27] is one example of a stochastic volatility process where the randomness of the variance, that is, the square of the volatility, obeys a Cox-Ingersoll-Ross [10] stochastic differential equation. Let $S(t)$ and $V(t)$ be the price of an asset and its instantaneous variance, respectively, to be determined by the system of stochastic differential equations

$$\begin{cases} dS(t) = \mu S(t)dt + \sqrt{V(t)}S(t)dW^S(t) \\ dV(t) = \kappa(\theta - V(t))dt + \sigma\sqrt{V(t)}dW^V(t) \end{cases} \tag{1.1}$$

where $W^S(t)$ and $W^V(t)$ are Wiener processes with constant correlation $\rho \in (-1, 1)$. This process is widely used as an asset price model in mathematical finance, but it is a degenerate diffusion process where the degeneracy in the diffusion coefficient is proportional to the square root of the distance to the boundary of the upper half-plane. The

generator of this process with killing [32, Section 8.2] is called the *elliptic Heston operator* and is a second-order elliptic, but not strictly elliptic, partial differential operator whose coefficients have linear growth in the spatial variable y.

The Feynman-Kac theorem, [32, Section 8.2] for strictly elliptic PDEs and [22] for the elliptic Heston operator, relates stochastic differential equations to a parabolic partial differential equation, where its elliptic part is precisely given by the generator of the process. This parabolic partial differential equation can be used to calculate the price of a derivative security whose underlying asset follows a stochastic volatility process, such as (1.1), but a closed-form solution usually does not exist, hence a numerical approximation is needed. Solving the associated elliptic problem numerically becomes a stepping-stone towards this goal, and that is the case which we consider in our thesis. We could have considered parabolic versions of the boundary value and obstacle problems for the elliptic Heston operator, but we focus on their elliptic versions as there are no new essential difficulties.

A recent citation search revealed that over 1000 articles* in scientific journals, not including books, reference the stochastic volatility model proposed by Heston in [27]. An example of this is Hilber-Schwab-Reichmann-Winter's book Computational Methods for Quantitative Finance where they are able to prove existence and uniqueness of solutions in suitable weighted Sobolev spaces (different from the ones we will consider) to the boundary value problem for the elliptic Heston operator, but only for restricted values of the constant parameters appearing in (1.1) ($\beta = 2\kappa\theta/\sigma^2 < 1$). By combining Feehan and Daskalopoulos results [12] with a coercivity result of our own, we can prove this existence and uniqueness for all values of β. Furthermore, despite of the widespread use of this degenerate stochastic process in financial engineering, we don't think, besides Kluge's diploma thesis [28], there has been very much good work addressing unresolved fundamental questions concerning the numerical solution of boundary value and obstacle problems for it, where convergence proofs are properly developed, as we do it in this thesis.

---

*A Thompson-Reuters Web of Knowledge [34] citation search performed on December 2, 2013 yielded 1034 references.

In Chapter 2 we start by reviewing results obtained by Feehan and Daskalopoulos [12] about existence and uniqueness of solutions to both boundary value and obstacle problems in weighted Sobolev spaces, together with our own results concerning the coercivity of the bilinear form in their variational formulations. We also extend our coercivity result to slightly more general operators than the Heston operator. Chapter 3 and Chapter 4 describe the Galerkin method and, for a collection of functions that we prove to be a basis, we also formulate the finite-element methodology to solve both the boundary value problem and the obstacle problem on a bounded open subset of the upper half plane. We prove convergence and obtain rates of convergence of the finite-element solutions for both problems, but numerical results are obtained only for the boundary value problem. Our MATLAB code implementation is not efficient enough to allow us to get numerical results for the obstacle problem within a reasonable computation time.

In Chapter 5 we review existence, uniqueness and regularity results, derived by Feehan [16], for both boundary value and obstacle problems on weighted Hölder spaces, followed by our own results in Chapter 6 with the analysis of some finite-difference operators, where we establish conditions on their parameters so that they satisfy discrete maximum principles, and we provide convergence results for the finite-difference solutions in the case of the boundary value problem. Finally, in Chapter 7 we combine the penalty and projection methods with the finite-difference methods of Chapter 6 to obtain numerical results for the obstacle problem.

Chapter 2

Review of weighted Sobolev spaces, and existence and uniqueness of solutions to elliptic Heston boundary value and obstacle problems

2.1 Elliptic Heston boundary value problem

Definition 2.1.1 (Spatial domain for the Heston partial differential equation)**.** *Let* $\mathscr{O} \subset \mathbb{H} := \mathbb{R} \times [0, \infty)$ *be a possibly unbounded open subset with boundary* $\partial\mathscr{O}$*, let* $\Gamma_1 := \mathbb{H} \cap \partial\mathscr{O}$*, let* Γ_0 *denote the interior of* $\{y = 0\} \cap \partial\mathscr{O}$*, and require* Γ_0 *to be non-empty. Notice that* $\partial\mathscr{O} = \Gamma_0 \cup \overline{\Gamma_1} = \overline{\Gamma_0} \cup \Gamma_1$*.*

We consider questions of existence, uniqueness, and regularity of solutions, $u : \mathscr{O} \to \mathbb{R}$, to a boundary value problem

$$Au = f \quad \text{a.e. in } \mathscr{O}, \quad u = g \quad \text{on } \Gamma_1, \tag{2.1}$$

where $f : \mathscr{O} \to \mathbb{R}$ is a source function, $g : \Gamma_1 \to \mathbb{R}$ is a function that prescribes a Dirichlet boundary condition along Γ_1, and A is an elliptic differential operator on $\mathscr{O}$ which is degenerate along Γ_0. Notice that no boundary condition is prescribed along Γ_0 - the reason why will be clear once we formulate a variational formulation to the problem of solving the equation $Au = f$. Throughout this thesis we set

$$Au := -\frac{y}{2}\left(u_{xx} + 2\rho\sigma u_{xy} + \sigma^2 u_{yy}\right) - (r - q - y/2)u_x - \kappa(\theta - y)u_y + ru, \tag{2.2}$$

and notice that $-A$ is the generator of the two-dimensional Heston stochastic volatility process with killing [27].

2.1.1 Hypotheses for the boundary value problem

The coefficients of A are required to obey

Assumption 2.1.2 (Strict ellipticity condition)**.** *The coefficients defining A in Equation* (2.2) *are constants obeying*

$$\sigma \neq 0 \text{ and } -1 < \rho < 1,$$

and $\kappa > 0$, $\theta > 0$, $r \geq 0$, and $q \in \mathbb{R}$. Define the constants $\beta := 2\kappa\theta/\sigma^2 > 0$ and $\mu := 2\kappa/\sigma^2 > 0$.

As in Daskalopoulos and Feehan [12], we will consider open subsets, $\mathscr{O}$, that satisfy some key hypotheses.

Hypothesis 2.1.3 (Hypothesis on the domain near Γ_0)**.** *For $\mathscr{O}$ as in Definition 2.1.1, there is a positive constant, δ_0, such that for all $0 < \delta \leq \delta_0$,*

$$\mathscr{O}_\delta^0 := \mathscr{O} \cap (\mathbb{R} \times (0,\delta)) = \Gamma_0 \times (0,\delta),$$
$$\Gamma_1 \cap (\mathbb{R} \times (0,\delta)) = \partial\Gamma_0 \times (0,\delta),$$

where $\Gamma_0 \subsetneqq \mathbb{R}$ is a finite union of open intervals.

Notice that if Γ_0 was empty, then standard methods [5,24] would apply to all of the problems considered in this section since the operator A would be strictly elliptic on $\mathscr{O}$. To state the next hypothesis on $\mathscr{O}$, let us recall the definition of an extension operator:

Definition 2.1.4 (Extension operator)**.** *For a domain $\mathscr{U} \subset \mathbb{H}$ and an integer $k \geq 1$, we call a bounded linear map $E : H^k(\mathscr{U}) \to H^k(\mathbb{R}^d)$ a* simple k-extension operator for $\mathscr{U}$ *if $Eu = u$ a.e. on $\mathscr{U}$ and $\|Eu\|_{H^k(\mathbb{R}^d)} \leq K\|u\|_{H^k(\mathscr{U})}$ for some constant $K > 0$ depending only on $\mathscr{U}$ and k.*

Hypothesis 2.1.5 (Extension operator property of the domain)**.** *For a domain, $\mathscr{O}$, as in Definition 2.1.1 and an integer $k \geq 1$, we require that there is a* simple k-extension operator from $\mathscr{O}$ to $\mathbb{H}$ *[12, Definition A.24].*

Hypothesis 2.1.5, with $k \leq 2$, is required when we consider traces of functions on Γ_1.

Feehan and Daskalopoulos [12] proved that Problem 2.1 is well-posed when solutions are sought in suitable function spaces which describe their qualitative behavior near the

boundary portion Γ_0: for example, integrability of derivatives in a neighborhood of Γ_0 via suitable weighted Sobolev spaces (by analogy with [29]).

Remark 2.1.6 (Interpretation of coefficients)**.** *In mathematical finance, the constants κ, θ, r, q, and σ have the interpretation described in [27]. Assumption 2.1.2 ensures that $y^{-1}A$ is strictly elliptic on $\mathbb{H}$, that is,*

$$\frac{1}{2}(\xi_1^2 + 2\rho\sigma\xi_1\xi_2 + \sigma^2\xi_2^2) > \nu(\xi_1^2 + \xi_2^2), \quad \forall (\xi_1, \xi_2) \in \mathbb{R}^2 - \{(0,0)\}, \tag{2.3}$$

where

$$0 < \nu := \frac{1}{4}\left(1 + \sigma^2 - \sqrt{(1-\sigma^2)^2 + 4\rho^2\sigma^2}\right) \leq 1/2,$$

by Assumption 2.1.2. Indeed, ν is the smallest eigenvalue of the symmetric matrix,

$$\begin{pmatrix} 1/2 & \rho\sigma/2 \\ \rho\sigma/2 & \sigma^2/2 \end{pmatrix},$$

which is positive because of Assumption 2.1.2.

2.1.2 Variational formulation

Consider Problem 2.1. We will study the boundary value problem

$$\begin{cases} Aw = f \quad \text{a.e. in } \mathscr{O}, \\ w = g \quad \text{on } \Gamma_1 \end{cases} \tag{2.4}$$

Assume we have a function $\bar{g} : \mathscr{O} \to \mathbb{R}$ smooth enough, such that $\bar{g} \restriction_\Gamma = g$ on Γ_1. Making the change of variable $\bar{w} := w - \bar{g}$ then Problem 2.4 is equivalent to

$$\begin{cases} Au = f \quad \text{a.e. in } \mathscr{O} \\ u = 0 \quad \text{on } \Gamma_1 \end{cases} \tag{2.5}$$

for some other function f. We will call Problem 2.5 the boundary value problem for the Heston operator with homogeneous Dirichlet condition along Γ_1, and we will restrict our analysis to it.

2.1.2.1 Weighted Sobolev spaces

We follow the same notation and definitions as in Feehan and Daskalopoulos [12].

Definition 2.1.7 (Spaces of continuous functions)**.** *Let $\mathscr{U} \subsetneqq \mathbb{R}^2$ be a domain with boundary $\partial\mathscr{U}$ and closure $\overline{\mathscr{U}} = \mathscr{U} \cup \partial\mathscr{U}$.*

1. *Let $T \subsetneqq \partial\mathscr{U}$ be relatively open. For any integer $\ell \geq 0$, then $C^{\ell}_{\mathrm{loc}}(\mathscr{U} \cup T)$ denotes the vector space of functions u on $\mathscr{U}$ with partial derivatives, $D^{\alpha}u$, for $0 \leq |\alpha| \leq \ell$, which are continuous on $\mathscr{U}$ and have continuous extensions to $\mathscr{U} \cup T$. (Compare [24, Section 4.4]) When $T = \partial\mathscr{U}$ (respectively, $T = \varnothing$), we abbreviate $C^{\ell}_{\mathrm{loc}}(\mathscr{U} \cup \partial\mathscr{U})$ by $C^{\ell}_{\mathrm{loc}}\left(\overline{\mathscr{U}}\right)$ (respectively, $C^{\ell}_{\mathrm{loc}}(\mathscr{U} \cup \varnothing)$ by $C^{\ell}(\mathscr{U})$). When $\ell = 0$, we abbreviate $C^{0}_{\mathrm{loc}}(\mathscr{U} \cup T)$ by $C_{\mathrm{loc}}(\mathscr{U} \cup T)$.*

2. *Denote $C^{\infty}_{\mathrm{loc}}(\mathscr{U} \cup T) := \cap_{\ell \geq 0} C^{\ell}_{\mathrm{loc}}(\mathscr{U} \cup T)$.*

3. *Let $C^{\infty}_{0}(\mathscr{U} \cup T)$ denote the subspace of C^{∞} functions with compact support in $\mathscr{U} \cup T$. When $T = \partial\mathscr{U}$ (respectively, $T = \varnothing$), we abbreviate $C^{\infty}_{0}(\mathscr{U} \cup \partial\mathscr{U})$ by $C^{\infty}_{0}\left(\overline{\mathscr{U}}\right)$ (respectively, $C^{\infty}_{0}(\mathscr{U} \cup \varnothing)$ by $C^{\infty}_{0}(\mathscr{U})$).*

4. *As in [2, Section 1.26], let $C^{\ell}\left(\overline{\mathscr{U}}\right)$ denote the Banach space of functions u on $\mathscr{U}$ with partial derivatives, $D^{\alpha}u$, for $0 \leq |\alpha| \leq \ell$, which are* bounded *and* uniformly *continuous on $\mathscr{U}$.*

5. *As in [31, Section 3.10], denote $C^{\infty}\left(\overline{\mathscr{U}}\right) := \cap_{\ell \geq 0} C^{\ell}\left(\overline{\mathscr{U}}\right)$.*

Remark 2.1.8. *Because we consider* unbounded *domains in this review, it is important to note the following:*

1. *Compare the definition of $C^{\ell}\left(\overline{\mathscr{U}}\right)$ and related vector spaces in [24, p. 10, Section 4.1, and p. 73], where it is only assumed that the derivatives $D^{\alpha}u$ are continuous on U, with continuous extensions to $\overline{\mathscr{U}}$. We emphasize the distinction here because in [24] the authors typically assume that $\mathscr{U}$ is* bounded *whereas we wish to include the case where $\mathscr{U}$ is* unbounded*. In other words, the definition of $C^{\ell}\left(\overline{\mathscr{U}}\right)$ in [24, p. 10] coincides with our definition of $C^{\ell}_{\mathrm{loc}}(\overline{\mathscr{U}})$.*

2. *We could have equivalently defined* $C^{\ell}_{\mathrm{loc}}\left(\overline{\mathscr{U}}\right)$ *as the vector space of functions* u *on* $\mathscr{U}$ *with partial derivatives,* $D^{\alpha}u$*, for* $0 \leq |\alpha| \leq \ell$*, which are bounded and uniformly continuous on* bounded subsets *of* $\mathscr{U}$*.*

3. *When* $\mathscr{U}$ *is bounded, then* $C^{\ell}_{\mathrm{loc}}\left(\overline{\mathscr{U}}\right) = C^{\ell}\left(\overline{\mathscr{U}}\right)$*.*

Definition 2.1.9 (First-order weighted Sobolev spaces)**.** *Let* $\mathscr{O} \subseteq \mathbb{H}$ *be a domain. Consider the positive weight function*

$$\mathfrak{w}(x,y) := y^{\beta-1}e^{-\gamma|x|-\mu y}, \quad (x,y) \in \mathbb{H}, \tag{2.6}$$

for a suitable non-negative constant* γ*. Recall* $\beta = 2\kappa\theta/\sigma^2$ *and* $\mu = 2\kappa/\sigma^2$*. Let* $L^2(\mathscr{O},\mathfrak{w})$ *be the space of all measurable functions* $u : \mathscr{O} \to \mathbb{R}$ *for which*

$$\|u\|^2_{L^2(\mathscr{O},\mathfrak{w})} := \int_{\mathscr{O}} u^2\,\mathfrak{w}\,dx\,dy < \infty,$$

and denote $H^0(\mathscr{O},\mathfrak{w}) := L^2(\mathscr{O},\mathfrak{w})$*.*

1. *Define the vector space of functions*

$$H^1(\mathscr{O},\mathfrak{w}) := \{u \in L^2(\mathscr{O},\mathfrak{w}) : (1+y)^{1/2}u \text{ and } y^{1/2}|Du| \in L^2(\mathscr{O},\mathfrak{w})\},$$

with norm

$$\|u\|_{H^1(\mathscr{O},\mathfrak{w})} := \left(\int_{\mathscr{O}} \left(y|Du|^2 + (1+y)u^2\right)\,\mathfrak{w}\,dx\,dy\right)^{1/2}$$

2. *Let* $T \subseteq \partial\mathscr{O}$ *be relatively open and let* $H^1_0(\mathscr{O}\cup T,\mathfrak{w})$ *be the closure in* $H^1(\mathscr{O},\mathfrak{w})$ *of* $C^{\infty}_0(\mathscr{O}\cup T)$*.*

By a straightforward modification of the proof of [2, Theorem 3.2], one can show that the spaces $H^k(\mathscr{O},\mathfrak{w})$, $k = 0,1$, and $H^1_0(\mathscr{O}\cup T,\mathfrak{w})$ are Banach spaces. Furthermore, the spaces $H^k(\mathscr{O},\mathfrak{w})$, $k = 0,1$, and $H^1_0(\mathscr{O}\cup T,\mathfrak{w})$ are Hilbert spaces with the inner products,

$$(u,v)_{L^2(\mathscr{O},\mathfrak{w})} := \int_{\mathscr{O}} uv\,\mathfrak{w}\,dx\,dy,$$
$$(u,v)_{H^1(\mathscr{O},\mathfrak{w})} := \int_{\mathscr{O}} \left(y\langle Du, Dv\rangle + (1+y)uv\right)\,\mathfrak{w}\,dx\,dy.$$

We let $H^{-1}(\mathscr{O},\mathfrak{w})$ denote the dual space of $H^1_0(\mathscr{O}\cup\Gamma_0,\mathfrak{w})$.

---

* *This constant will be determined later depending on* $\mathscr{O}$ *and the Heston operator coefficients. In the case of* $\mathscr{O}$ *bounded we will set* γ *as zero.*

Definition 2.1.10 (Second-order weighted Sobolev spaces)**.** *Let $\mathscr{O} \subseteq \mathbb{H}$ be a domain and define the vector space of functions*

$$H^2(\mathscr{O}, \mathfrak{w}) := \left\{ u \in L^2(\mathscr{O}, \mathfrak{w}) : (1+y)^{1/2}u, (1+y)|Du|, y|D^2u| \in L^2(\mathscr{O}, \mathfrak{w}) \right\},$$

with norm

$$\|u\|_{H^2(\mathscr{O},\mathfrak{w})} := \left(\int_{\mathscr{O}} \left(y^2|D^2u|^2 + (1+y)^2|Du|^2 + (1+y)u^2 \right) \mathfrak{w}\, dx\, dy \right)^{1/2}$$

The space $H^2(\mathscr{O}, \mathfrak{w})$ is a Banach space (again, by modification of the proof of [2, Theorem 3.2]) and a Hilbert space with the inner product,

$$(u, v)_{H^2(\mathscr{O},\mathfrak{w})} := \int_{\mathscr{O}} \left(y^2\langle D^2u, D^2v\rangle + (1+y)^2\langle Du, Dv\rangle + (1+y)uv \right) \mathfrak{w}\, dx\, dy.$$

2.1.2.2 Bilinear form associated with the elliptic Heston operator

Define the constants

$$a_1 := \frac{\kappa\rho}{\sigma} - \frac{1}{2} \quad \text{and} \quad b_1 := r - q - \frac{\kappa\theta\rho}{\sigma}. \tag{2.7}$$

We define the bilinear form associated to the Heston boundary value problem.

Definition 2.1.11 (Heston bilinear form)**.** *We call*

$$\begin{aligned} \mathfrak{a}(u, v) := {} & \frac{1}{2} \int_{\mathscr{O}} \left(u_x v_x + \rho\sigma u_y v_x + \rho\sigma u_x v_y + \sigma^2 u_y v_y \right) y\, \mathfrak{w}\, dx\, dy \\ & - \frac{\gamma}{2} \int_{\mathscr{O}} \left(u_x + \rho\sigma u_y \right) v\, \mathrm{sign}(x) y\, \mathfrak{w}\, dx\, dy \\ & - \int_{\mathscr{O}} (a_1 y + b_1) u_x v\, \mathfrak{w}\, dx\, dy + \int_{\mathscr{O}} ruv\, \mathfrak{w}\, dx\, dy, \quad \forall u, v \in H^1(\mathscr{O}, \mathfrak{w}), \end{aligned} \tag{2.8}$$

the bilinear form associated with the Heston operator*, A, in* (2.1)*.*

The following result is shown in [12],

Lemma 2.1.12 (Integration by parts for the Heston operator, [12])**.** *Suppose $u \in H^2(\mathscr{O}, \mathfrak{w})$ and $v \in H^1(\mathscr{O}, \mathfrak{w})$. Then $Au \in L^2(\mathscr{O}, \mathfrak{w})$ and*

$$(Au, v)_{L^2(\mathscr{O},\mathfrak{w})} = \mathfrak{a}(u, v) - \frac{1}{2} \int_{\Gamma_1} \left(n^x(u_x + \rho\sigma u_y) + n^y(\rho\sigma u_x + \sigma^2 u_y) \right) vy\mathfrak{w}\, dS, \tag{2.9}$$

where $\mathbf{n} := (n^x, n^y)$ is the outward-pointing unit normal vector field along Γ_1, dS is the curve measure on Γ_1 induced by Lebesgue measure on $\mathbb{R}^2$, and the integrand on Γ_1 is defined in the trace sense.

Notice that this integration-by-parts formula does not involve any integral along Γ_0.

Remark 2.1.13. *Equation* (2.9) *does not necessarily hold if the hypothesis* $u \in H^2(\mathscr{O}, \mathfrak{w})$ *is relaxed to* $u \in H^2_{\mathrm{loc}}(\mathscr{O}, \mathfrak{w}) \cap H^1(\mathscr{O}, \mathfrak{w})$ *and* $Au \in L^2(\mathscr{O}, \mathfrak{w})$. *Examples [12, C.1] and [1, Sections 13.4.21 and 13.5.8] show that there are functions* $u \in H^2_{\mathrm{loc}}(\mathscr{O}, \mathfrak{w}) \cap H^1(\mathscr{O}, \mathfrak{w})$ *with* $Au = 0$ *on* $\mathscr{O}$ *but* $y^\beta u_y > 0$ *along* Γ_0 *and so the* Γ_0*-boundary integral expected to be part of the integration-by-parts formula for* $(Au, v)_{L^2(\mathscr{O},\mathfrak{w})}$ *is non-zero for such a function* u.

2.1.2.3 Classical, strong and weak solutions

Following Feehan and Daskalopoulos [12], the integration by parts formula of Lemma 2.1.12 motivates the following definitions.

Definition 2.1.14 (Classical solution)**.** *Given a function* $f \in C^\alpha(\mathscr{O})$, *for some* $0 < \alpha < 1$, *we call a function* $u \in C^{2,\alpha}(\mathscr{O}) \cap C_{\mathrm{loc}}(\mathscr{O} \cup \Gamma_1)$ *a* classical solution *to the boundary value problem for the Heston operator with homogeneous Dirichlet condition along* Γ_1 *if*

$$\begin{cases} Au = f \quad \text{in } \mathscr{O}, \\ u = 0 \quad \text{on } \Gamma_1, \\ \lim_{y \downarrow 0} y^\beta (\rho u_x + \sigma u_y) = 0 \quad \text{on } \Gamma_0. \end{cases} \tag{2.10}$$

Definition 2.1.15 (Strong solution)**.** *Given a function* $f \in L^2(\mathscr{O}, \mathfrak{w})$, *we call a function* $u \in H^2(\mathscr{O}, \mathfrak{w})$ *a* strong solution *to the boundary value problem for the Heston operator with* homogeneous *Dirichlet boundary condition on* Γ_1 *if* u *obeys*

$$\begin{cases} Au = f \quad \text{a.e. in } \mathscr{O} \\ u = 0 \quad \text{on } \Gamma_1 \end{cases} \tag{2.11}$$

Definition 2.1.16 (Weak solution)**.** *Given a function* $f \in L^2(\mathscr{O}, \mathfrak{w})$, *we call a function* $u \in H^1_0(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$ *a* solution to the variational equation *for the Heston operator with* homogeneous *Dirichlet boundary condition on* Γ_1 *if*

$$\mathfrak{a}(u, v) = (f, v)_{L^2(\mathscr{O},\mathfrak{w})}, \quad \forall v \in H^1_0(\mathscr{O} \cup \Gamma_0, \mathfrak{w}). \tag{2.12}$$

Feehan and Daskalopoulos [12] proved the well-posedness of Problem 2.11. They showed that if $u \in H^2(\mathscr{O}, \mathfrak{w})$, then u is a strong solution if and only if u is a weak solution, and proves the well-posedness of Problem 2.10 by regularity arguments.

2.1.2.4 Continuity and coercivity of the bilinear form

One of their results that will be used frequently is the continuity estimate for $\mathfrak{a}$. Namely,

Proposition 2.1.17. *[12, Proposition 2.40] Assume $b_1 = 0$. For all $u, v \in H^1(\mathscr{O}, \mathfrak{w})$,*

$$|\mathfrak{a}(u, v)| \leq C\|u\|_{H^1(\mathscr{O},\mathfrak{w})}\|v\|_{H^1(\mathscr{O},\mathfrak{w})}$$

for a positive constant C that depends only on coefficients $r, q, \kappa, \theta, \rho, \sigma$, and γ.

Remark 2.1.18. *The assumption $b_1 = 0$ is not restrictive. An affine change of variables on independent and dependent variables can be done to achieve that [12, Lemma 2.2]. We note that this change of variables sends Γ_0 to Γ_0, $\mathbb{H}$ to $\mathbb{H}$, and preserves the boundedness (or unboundedness) of $\mathscr{O}$.*

Feehan and Daskalopoulos [12] provide a framework for examining existence and uniqueness of solutions to the infinite-dimensional Problem 2.12. When the bilinear form $\mathfrak{a}$ is non-coercive, a finite-element method implementation is more challenging. Fortunately, the existence and uniqueness of solutions to the finite-dimensional variational problem, given by the Galerkin method [26], will follow by classical arguments since the Heston bilinear form, $\mathfrak{a}$, is continuous on $H^1(\mathscr{O}, \mathfrak{w})$ (in particular on $H^1_0(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$), by Proposition 2.1.17, and coercive on $H^1_0(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$ when $\mathscr{O}$ is bounded in the x-direction and satisfies the regularity hypothesis we introduced earlier, as we prove below.

Proposition 2.1.19 (Coercivity of $\mathfrak{a}$)**.** *The Heston bilinear form of Definition 2.1.11, for $\gamma = 0$ and $\mathscr{O}$ satisfying Hypothesis 2.1.3 and 2.1.5, is coercive on $H^1_0(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$. That is, for all $u \in H^1_0(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$,*

$$|\mathfrak{a}(u, u)| \geq \alpha\|u\|^2_{H^1(\mathscr{O},\mathfrak{w})},$$

where α is a positive constant that depends only on ρ, σ and r.

Proof. Let $u \in H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$. From Definition 2.1.11, with $\gamma = 0$,

$$\mathfrak{a}(u,u) = \frac{1}{2}\int\limits_{\mathscr{O}} \left(u_x^2 + 2\rho\sigma u_x u_y + \sigma^2 u_y^2\right) y \, d\mathfrak{w} - \int\limits_{\mathscr{O}} (a_1 y + b_1) u_x u \, d\mathfrak{w} + \int\limits_{\mathscr{O}} r u^2 \, d\mathfrak{w}$$

where $d\mathfrak{w} := \mathfrak{w}\, dx\, dy$. We will show that the second integral is actually zero, and we will estimate the first integral from below. Indeed, by integration-by-parts (we take $u \in C^1\left(\overline{\mathscr{O}}\right)$ by a density argument as in [12, Lemma 2.23] since we are assuming Hypothesis 2.1.3),

$$\begin{aligned}
\int\limits_{\mathscr{O}} (a_1 y + b_1) u_x u \, d\mathfrak{w} &= \frac{1}{2}\int\limits_{\mathscr{O}} (a_1 y + b_1) \left(u^2\right)_x \, d\mathfrak{w} \\
&= \frac{1}{2}\int\limits_{\partial\mathscr{O}} u^2 (a_1 y + b_1) n^x \mathfrak{w} \, dS - \frac{1}{2}\int\limits_{\mathscr{O}} u^2 (a_1 y + b_1)_x \, d\mathfrak{w} \\
&= \frac{1}{2}\int\limits_{\partial\mathscr{O}} u^2 (a_1 y + b_1) n^x \mathfrak{w} \, dS,
\end{aligned}$$

where dS represents the (Lebesgue) surface differential along the boundary, and n^x is the x-component of $n = (n^x, n^y)$, the outer unit normal vector to $\partial\mathscr{O}$. Now, given that $\partial\mathscr{O} = \overline{\Gamma}_0 \cup \Gamma_1$, $n^x \equiv 0$ on Γ_0, and since $u = 0$ on Γ_1 (in trace sense), then

$$\int\limits_{\mathscr{O}} (a_1 y + b_1) u_x u \, d\mathfrak{w} = \frac{1}{2}\int\limits_{\Gamma_1} u^2 (a_1 y + b_1) n^x \mathfrak{w} d\Gamma = 0.$$

Thus,

$$\mathfrak{a}(u,u) = \frac{1}{2}\int\limits_{\mathscr{O}} (u_x^2 + 2\rho\sigma u_x u_y + \sigma^2 u_y^2) y \, d\mathfrak{w} + \int\limits_{\mathscr{O}} r u^2 \, d\mathfrak{w}. \tag{2.13}$$

To prove that $\mathfrak{a}$ is coercive, given Equation (2.13), we essentially just need to bound from below the term $u_x^2 + 2\rho\sigma u_x u_y + \sigma^2 u_y^2$ by an expression of the form $C(u_x^2 + u_y^2)$ with C a positive constant. In fact, since

$$2\rho\sigma u_x u_y = \pm 2|\rho|\sigma u_x u_y \geq -\left(\sqrt{|\rho|} u_x\right)^2 - \left(\sqrt{|\rho|}\sigma u_y\right)^2,$$

then,

$$u_x^2 + 2\rho\sigma u_x u_y + \sigma^2 u_y^2 \geq (1 - |\rho|) u_x^2 + (1 - |\rho|)\sigma^2 u_y^2 = (1 - |\rho|) \min\left\{1, \sigma^2\right\} \left(u_x^2 + u_y^2\right).$$

From this find and Equation (2.13) follow that there exists $C(\rho, \sigma) > 0$ such that

$$\mathfrak{a}(u,u) \geq \frac{1}{2} C(\rho,\sigma) \int\limits_{\mathscr{O}} (u_x^2 + u_y^2) y \, d\mathfrak{w} + \int\limits_{\mathscr{O}} r u^2 \, d\mathfrak{w},$$

which implies that there exists a constant $\alpha := C(\rho, \sigma, r) > 0$ such that

$$\mathfrak{a}(u,u) \geq \alpha ||u||^2_{H^1(\mathscr{O},\mathfrak{w})}, \text{ for all } u \in H^1_0(\mathscr{O} \cup \Gamma_0)$$

That is, $\mathfrak{a}$ is coercive. □

Remark 2.1.20 (Extension to non-constant coefficients)**.** *Proposition 2.1.19 can be extended easily to non-constant coeficientes of the Heston PDE as long as parameters a_1 and b_1 of Equation* (2.7) *are both functions of y only. Also r must be bounded below by a positive constant.*

In a variational problem it is highly desirable to have coercivity of the associated bilinear form, hence it is of interest to explore under which conditions the proof of Proposition 2.1.19 can be extended to Heston-like operators. Consider a differential operator of the form

$$Bu := -\frac{y}{2}\left(u_{xx} + 2\rho\sigma u_{xy} + \sigma^2 u_{yy}\right) + c_1(x,y)u_x - \kappa(\theta - y)u_y + c_2(x,y)u, \tag{2.14}$$

prescribed on an open subset $\mathscr{O}$, as in Definition 2.1.1 and bounded in the x- direction, where ρ and σ still satisfy Assumption 2.1.2, and c_1 and c_2 are measurable functions on which we will impose conditions shortly.

Motivated by the Heston operator, and given we are keeping the leading second-order coefficient $-y/2$, we continue to use $H^1(\mathscr{O}, \mathfrak{w})$ and $H^2(\mathscr{O}, \mathfrak{w})$ as the underlying Hilbert spaces where to define a bilinear form associated to this slightly more general operator.

We define a new bilinear form $\mathfrak{a}_B$ by,

$$\begin{aligned}\mathfrak{a}_B(u,v) := &\frac{1}{2}\int_{\mathscr{O}} \left(u_x v_x + \rho\sigma u_y v_x + \rho\sigma u_x v_y + \sigma^2 u_y v_y\right) y\, \mathfrak{w}\, dx\, dy \\ &+ \int_{\mathscr{O}} \left(\rho\sigma\mu\frac{\theta - y}{2} + c_1(x,y)\right) u_x v\, \mathfrak{w}\, dx\, dy \\ &+ \int_{\mathscr{O}} c_2(x,y) uv\, \mathfrak{w}\, dx\, dy, \quad \forall\, u, v \in H^1(\mathscr{O}, \mathfrak{w}),\end{aligned} \tag{2.15}$$

Let us prove that an integration-by-parts formula like the one in Lemma 2.1.12 still holds.

Lemma 2.1.21. *Let $u \in H^2(\mathscr{O}, \mathfrak{w})$ and $v \in H^1(\mathscr{O}, \mathfrak{w})$. Then $Bu \in L^2(\mathscr{O}, \mathfrak{w})$ and*

$$(Bu, v)_{L^2(\mathscr{O},\mathfrak{w})} = \mathfrak{a}_B(u, v) - \frac{1}{2}\int_{\Gamma_1} \left(n^x(u_x + \rho\sigma u_y) + n^y(\rho\sigma u_x + \sigma^2 u_y)\right) vy\mathfrak{w}\, dS, \quad (2.16)$$

where $\mathbf{n} := (n^x, n^y)$ is the outward-pointing unit normal vector field along Γ_1, dS is the curve measure on Γ_1 induced by Lebesgue measure on $\mathbb{R}^2$, and the integrand on Γ_1 is defined in the trace sense.

Proof. As in Feehan and Daskalopoulos, by density arguments [12, Corollary A.14] and [12, Lemma A.26], we can assume that $u \in C^2(\overline{\mathscr{O}})$ and $v \in C^1(\overline{\mathscr{O}})$.

From its definition in (2.14), we observe that the expression Bu in $\mathscr{O}$ can be written conveniently as

$$\begin{aligned} Bu = -\frac{1}{2}y^{1-\beta}\Big(\left(y^\beta u_x\right)_x + \rho\sigma\left(y^\beta u_x\right)_y + \rho\sigma\left(y^\beta u_y\right)_x + \sigma^2\left(y^\beta u_y\right)_y\Big) \\ + \frac{\rho\sigma}{2}\beta u_x + \frac{\sigma^2}{2}\beta u_y + c_1(x,y)u_x - \kappa(\theta - y)u_y + c_2(x,y)u \quad \text{on } \mathscr{O}. \end{aligned}$$

Thus, using $\beta = 2\kappa\theta/\sigma^2$ and $\mu = 2\kappa/\sigma^2$, the preceding expression simplifies to

$$\begin{aligned} Bu = -\frac{1}{2}y^{1-\beta}\Big(\left(y^\beta u_x + \rho\sigma y^\beta u_y\right)_x + \left(\rho\sigma y^\beta u_x + y^\beta\sigma^2 u_y\right)_y\Big) \\ + \left(\frac{\rho\sigma\mu\theta}{2} + c_1(x,y)\right)u_x + \kappa y u_y + c_2(x,y)u \quad \text{on } \mathscr{O}. \end{aligned} \quad (2.17)$$

Recall that for bounded open subsets in the x-direction, we can take $\mathfrak{w}(y) = y^{\beta-1}e^{-\mu y}$. Multiplying both sides of (2.17) by $v\,\mathfrak{w}$ and integrating over $\mathscr{O}$, gives

$$\begin{aligned} \int_{\mathscr{O}} (Bu)v\,\mathfrak{w}\,dx\,dy = -\frac{1}{2}\int_{\mathscr{O}}\Big(\left(y^\beta u_x + \rho\sigma y^\beta u_y\right)_x + \left(\rho\sigma y^\beta u_x + y^\beta\sigma^2 u_y\right)_y\Big)ve^{-\mu y}\,dx\,dy \\ + \int_{\mathscr{O}}\left(\left(\frac{\rho\sigma\mu\theta}{2} + c_1(x,y)\right)u_x + \kappa y u_y + c_2(x,y)u\right)v\,\mathfrak{w}\,dx\,dy. \end{aligned}$$

Integrating by parts, using $(e^{-\mu y})_y = -\mu e^{-\mu y}$ and denoting by $d\mathfrak{w} = \mathfrak{w}\,dx\,dy$, gives

$$\begin{aligned} (Bu, v)_{L^2(\mathscr{O},\mathfrak{w})} = &\frac{1}{2}\int_{\mathscr{O}} y\left(u_x v_x + \rho\sigma u_x v_y + \rho\sigma u_y v_x + \sigma^2 u_y v_y\right)\,d\mathfrak{w} \\ &- \frac{1}{2}\int_{\mathscr{O}} y\mu\left(\rho\sigma u_x + \sigma^2 u_y\right)v\,d\mathfrak{w} \\ &+ \int_{\mathscr{O}}\left(\left(\frac{\rho\sigma\mu\theta}{2} + c_1(x,y)\right)u_x + \kappa y u_y + c_2(x,y)u\right)v\,d\mathfrak{w} \\ &- \frac{1}{2}\int_{\partial\mathscr{O}}\left(n^x\left(y^\beta u_x + \rho\sigma y^\beta u_y\right) + n^y\left(\rho\sigma y^\beta u_x + y^\beta\sigma^2 u_y\right)\right)ve^{-\mu y}\,dS \end{aligned}$$

After gathering terms, the preceding expression becomes

$$(Bu,v)_{L^2(\mathscr{O},\mathfrak{w})} = \mathfrak{a}_B(u,v) - \frac{1}{2}\int_{\Gamma_1}\left(n^x\left(u_x+\rho\sigma u_y\right)+n^y\left(\rho\sigma u_x+\sigma^2 u_y\right)\right)vy\mathfrak{w}\,dS \\ -\frac{1}{2}\int_{\Gamma_0} n^y\left(\rho\sigma u_x+\sigma^2 u_y\right)vy\mathfrak{w}\,dx,$$

where $\mathfrak{a}_B(u,v)$ is defined by (2.15). But

$$\int_{\Gamma_0} n^y\left(\rho\sigma u_x+\sigma^2 u_y\right)vy\mathfrak{w}\,dx = -\int_{\Gamma_0}\left(\rho\sigma u_x+\sigma^2 u_y\right)vy^\beta e^{-\mu y}\,dx.$$

Now, since $u_x, u_y, v \in C(\overline{\mathscr{O}})$ and $\beta > 0$, then

$$\int_{\Gamma_0} n^y\left(\rho\sigma u_x+\sigma^2 u_y\right)vy\mathfrak{w}\,dx = 0, \tag{2.18}$$

because the integrand is identically zero along Γ_0. This yields (2.16) for $u \in C^2(\overline{\mathscr{O}})$ and $v \in C^1(\overline{\mathscr{O}})$, and this completes the proof. □

This formula, as in the case of the Heston operator A, motivates the definition of weak solutions for, say, an associated boundary value problem. We are now ready to state a more general coercivity result.

Proposition 2.1.22 (Coercivity of $\mathfrak{a}_B$)**.** *Let $c_1 \in L^1_{\mathrm{loc}}(\mathscr{O})$ such that $\frac{\partial c_1}{\partial x} \in L^1_{\mathrm{loc}}(\mathscr{O})$, and c_2 be a measurable function such that,*

$$c_2(x,y) - \frac{\partial c_1}{\partial x}(x,y) \geq C > 0$$

almost everywhere for some positive constant C. Then $\mathfrak{a}_B$ is coercive on $H^1_0(\mathscr{O}\cup\Gamma_0,\mathfrak{w})$.

Proof. Let $u \in H^1_0(\mathscr{O}\cup\Gamma_0,\mathfrak{w})$. From Equation (2.15) it follows that,

$$\mathfrak{a}_B(u,u) = \frac{1}{2}\int_{\mathscr{O}}\left(u_x^2+2\rho\sigma u_x u_y+\sigma^2 u_y^2\right)y\,\mathfrak{w}\,dx\,dy \\ +\int_{\mathscr{O}}\left(\rho\sigma\mu\frac{(\theta-y)}{2}+c_1(x,y)\right)u_x u\,\mathfrak{w}\,dx\,dy + \int_{\mathscr{O}} c_2(x,y)u^2\,\mathfrak{w}\,dx\,dy. \tag{2.19}$$

We bound from below the first integral exactly as we did it in Proposition 2.1.19. Let us rewrite the second integral:

$$\int_{\mathscr{O}}\left(\rho\sigma\mu\frac{(\theta-y)}{2}+c_1(x,y)\right)u_x u\,\mathfrak{w}\,dx\,dy = \frac{1}{2}\int_{\mathscr{O}}\left(\rho\sigma\mu\frac{(\theta-y)}{2}+c_1(x,y)\right)\left(u^2\right)_x\,\mathfrak{w}\,dx\,dy$$

Integrating by parts with respect to the x-variable, we then get,

$$\int_{\mathscr{O}} \left(\rho\sigma\mu\frac{(\theta-y)}{2} + c_1(x,y)\right) u_x u\, \mathfrak{w}\, dx\, dy = \frac{1}{2}\int_{\partial\mathscr{O}} \left(\rho\sigma\mu\frac{(\theta-y)}{2} + c_1(x,y)\right) u^2 n^x \mathfrak{w}\, dS - \frac{1}{2}\int_{\mathscr{O}} u^2 \frac{\partial c_1}{\partial x}(x,y)\, \mathfrak{w}\, dx\, dy.$$

Notice that along $\partial\mathscr{O}$, either $n^x = 0$ (along Γ_0) or $u = 0$ (along Γ_1). Thus,

$$\int_{\mathscr{O}} \left(\rho\sigma\mu\frac{(\theta-y)}{2} + c_1(x,y)\right) u_x u\, \mathfrak{w}\, dx\, dy = -\frac{1}{2}\int_{\mathscr{O}} u^2 \frac{\partial c_1}{\partial x}(x,y)\, \mathfrak{w}\, dx\, dy.$$

and we have then,

$$\mathfrak{a}_B(u,u) \geq \frac{1}{2}\int_{\mathscr{O}} (1-|\rho|)\min\{1,\sigma^2\}(u_x^2+u_y^2) y\, d\,\mathfrak{w} + \int_{\mathscr{O}} \left(c_2(x,y) - \frac{\partial c_1}{\partial x}(x,y)\right) u^2\, d\,\mathfrak{w}.$$

Since $c_2(x,y) - \partial c_1/\partial x \geq C > 0$, it follows that $\mathfrak{a}_B(u,u)$ is bounded from below by a positive multiple of $\|u\|^2_{H^1(\mathscr{O},\mathfrak{w})}$, and therefore that $\mathfrak{a}_B$ is coercive on $H^1_0(\mathscr{O}\cup\Gamma_0,\mathfrak{w})$. □

2.2 Elliptic Heston obstacle problem

As with the boundary value problem for the Heston partial differential operators, we can consider questions of existence, uniqueness, and regularity of solutions, $u : \mathscr{O} \to \mathbb{R}$, to the obstacle problem

$$\min\{Au - f, u - \psi\} = 0 \quad \text{a.e. in } \mathscr{O}, \quad u = g \quad \text{on } \Gamma_1, \tag{2.20}$$

where $\mathscr{O} \subset \mathbb{H}$ is a possibly unbounded open subsets of the open upper half-plane $\mathbb{H} := \mathbb{R}\times(0,\infty)$, $\Gamma_1 = \partial\mathscr{O}\cap\mathbb{H}$ is the portion of the boundary $\partial\mathscr{O}$ of $\mathscr{O}$ which lies in $\mathbb{H}$, $f : \mathscr{O} \to \mathbb{R}$ is a source function, the function $g : \mathscr{O}\cup\Gamma_1 \to \mathbb{R}$ prescribes a Dirichlet boundary condition along Γ_1, and $\psi : \mathscr{O}\cup\Gamma_1 \to \mathbb{R}$ is an obstacle function which is compatible with g in the sense that $\psi \leq g$ on Γ_1, and $-A$ is the elliptic differential operator defined by Equation (2.2), that is, the generator of the two-dimensional Heston stochastic volatility process with killing [27].

As in the boundary value problem, no boundary condition is prescribed along Γ_0. Feehan and Daskalopoulos [12] proved that Problem 2.20 is well-posed when one seeks for solutions in the weighted Sobolev spaces already introduced for the equation. All notation and assumptions made in Section 2.1 will be used and assumed in this section as well.

2.2.1 Variational formulation

Consider Problem 2.20. By considering a sufficiently regular extension of the function g, and by making a change of variable in u, we can focus on solving the homogeneous obstacle problem,

$$\min\{Au - f, u - \psi\} = 0 \quad \text{a.e. in } \mathscr{O}, \quad u = 0 \quad \text{on } \Gamma_1, \tag{2.21}$$

where $\psi : \mathscr{O} \cup \Gamma_1 \to \mathbb{R}$ satisfies the compatibility condition, $\psi \leq 0$ on Γ_1, and f is a source function. We will refer to Problem 2.21 as *the obstacle problem for the Heston operator with homogeneous Dirichlet condition along* Γ_1, and we will restrict our analysis to it.

The bilinear form $\mathfrak{a}$ of Definition 2.1.11 is still well-defined on $H^1(\mathscr{O}, \mathfrak{w})$; continuous on $H_0^1(\mathscr{O}, \mathfrak{w})$, when assuming that $b_1 = 0$, as stated in Proposition 2.1.17; coercive on $H_0^1(\mathscr{O}, \mathfrak{w})$, when $\gamma = 0$, as proved in Proposition 2.1.19; and the integration-by-parts formula of Lemma 2.1.12 still holds. Therefore, in the setting of the obstacle problem, we can once again consider weak solutions as we did for the equation.

2.2.2 Classical, strong, and weak solutions

The integration-by-parts formula of Lemman 2.1.12 motivates analogous definitions for classical, strong and weak solutions for the Heston obstacle problem.

Definition 2.2.1 (Classical solution)**.** *Given functions $f \in C^\alpha(\mathscr{O})$, for some $0 < \alpha < 1$, $g \in C^{2,\alpha}(\mathscr{O}) \cap C_{\text{loc}}(\mathscr{O} \cup \Gamma_1)$, and $\psi \in C_{\text{loc}}(\mathscr{O} \cup \Gamma_1)$ with*

$$\psi \leq g \quad \text{on } \Gamma_1, \tag{2.22}$$

we call $u \in C^{1,1}(\mathscr{O}) \cap C_{\text{loc}}(\mathscr{O} \cup \Gamma_1)$ a classical solution *to an obstacle problem for the elliptic Heston operator with inhomogeneous Dirichlet condition along Γ_1 if*

$$\min\{Au - f, u - \psi\} = 0 \quad \text{on } \mathscr{O}, \tag{2.23}$$

$$u = g \quad \text{on } \Gamma_1, \tag{2.24}$$

$$\lim_{y \downarrow 0} y^\beta (\rho u_x + \sigma u_y) = 0 \quad \text{on } \Gamma_0. \tag{2.25}$$

Definition 2.2.2 (Strong solution)**.** *Given functions $f \in L^2(\mathscr{O}, \mathfrak{w})$, $g \in H^2(\mathscr{O}, \mathfrak{w})$, and $\psi \in H^2(\mathscr{O}, \mathfrak{w})$ obeying* (2.22)*, we call $u \in H^2(\mathscr{O}, \mathfrak{w})$ a* strong solution *to an obstacle problem for the elliptic Heston operator with inhomogeneous Dirichlet boundary condition along Γ_1 if u obeys* (2.23) *(a.e. on $\mathscr{O}$) and* (2.24)*.*

Definition 2.2.3 (Weak solution for the case of non-homogeneous Dirichlet boundary condition)**.** *Given functions $f \in L^2(\mathscr{O}, \mathfrak{w})$, $g \in H^1(\mathscr{O}, \mathfrak{w})$, and $\psi \in H^1(\mathscr{O}, \mathfrak{w})$ obeying* (2.22) *in the sense that*

$$(\psi - g)^+ \in H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w}),$$

we call $u \in H^1(\mathscr{O}, \mathfrak{w})$ a solution to the variational inequality for the Heston operator with inhomogeneous Dirichlet boundary condition along Γ_1 if

$$\begin{gathered} \mathfrak{a}(u, v-u) \geq (f, v-u)_{L^2(\mathscr{O}, \mathfrak{w})}, \\ u \geq \psi \text{ a.e. on } \mathscr{O} \text{ and } u - g \in H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w}), \\ \forall v \in H^1(\mathscr{O}, \mathfrak{w}) \text{ with } v \geq \psi \text{ a.e. on } \mathscr{O} \text{ and } v - g \in H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w}). \end{gathered} \tag{2.26}$$

A reduction to a variational inequality with homogeneous Dirichlet boundary condition can be done just as in the case of the equation. Therefore, for the remainder of this dissertation we may consider without loss of generality variational inequalities and obstacle problems with homogeneous Dirichlet boundary condition on Γ_1.

Definition 2.2.4 (Weak solution for the case of homogeneous Dirichlet boundary condition)**.** *Given functions $f \in L^2(\mathscr{O}, \mathfrak{w})$ and $\psi \in H^1(\mathscr{O}, \mathfrak{w})$ such that $\psi \leq 0$ on Γ_1, in the sense that $\psi^+ \in H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$, define $\mathbb{K} = \{v \in H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w}) | v \geq \psi \text{ a.e. on } \mathscr{O}\}$. We call $u \in H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$ a solution to the variational inequality for the Heston operator with homogeneous Dirichlet boundary condition along Γ_1, if for every $v \in \mathbb{K}$ we have*

$$\begin{gathered} \mathfrak{a}(u, v-u) \geq (f, v-u)_{L^2(\mathscr{O}, \mathfrak{w})}, \\ u \geq \psi \text{ a.e. on } \mathscr{O}. \end{gathered} \tag{2.27}$$

Feehan and Daskalopoulos [12] proved the well-posedness of Problem 2.2.4, however, they do not achieve this by setting up the usual framework since they didn't have coercivity of the bilinear form $\mathfrak{a}$. They proved continuity of $\mathfrak{a}$, but together with our proposition 2.1.19, where we proof coercivity of $\mathfrak{a}$ in the case that coefficient $\gamma = 0$, we

automatically get the existence and uniqueness of a solution via the Lions-Stampacchia Theorem [25, Theorem 3.1]. This provided us with an appropriate theoretical framework to implement the finite-element method to solve their variational formulations approximately.

Feehan and Daskalopoulos [12] also proved the well-posedness of Problem 2.2.2, and they showed that if $u \in H^2(\mathscr{O}, \mathfrak{w})$, then u is a strong solution if and only if u is a weak solution.

Chapter 3

Finite-element method for the elliptic boundary value problem

We follow [25, Appendix I] to give first a brief introduction to a family of linear variational problems, their internal approximations, and the Galerkin method to solve them for a general basis of the underlying Hilbert space. Then we specify a basis, obtain the linear system corresponding to the finite-dimensional approximating problem, and provide a rate of convergence for the approximating solutions. Finally, we illustrate our method with numerical results.

3.1 Elliptic linear variational problems and the finite-element method

We consider,

i) A real Hilbert space V with scalar product $(\cdot,\cdot)$ and associated norm $\|\cdot\|$.

ii) V^*, the topological dual space of V.

iii) A bilinear form, $a : V \times V \longrightarrow \mathbb{R}$, continuous (that is, there exists a constant $C > 0$ such that $\mathfrak{a}(u,v) \leq C\|u\|\|v\|$ for all $u, v \in V$) and coercive (that is, there exists a constant $\alpha > 0$ such that $\mathfrak{a}(v,v) \leq \alpha\|v\|^2$ for all $v \in V$; $\mathfrak{a}$ is possibly non symmetric).

iv) A continuous linear functional, $L \in V^*$.

The fundamental *linear* variational problem under consideration reads as follows:

Definition 3.1.1 (The fundamental linear variational problem)**.** *Find $u \in V$ such that*

$$\mathfrak{a}(u,v) = L(v), \quad \forall\, v \in V \tag{P}$$

We recall the Lax-Milgram Theorem:

Theorem 3.1.2. [25, App. I, Thm 2.1] *Under the above hypothesis i)-iv), Problem P has a unique solution.*

3.1.1 Internal approximations

We suppose that we are given a small parameter h and a family $\{V_h\}_{h>0}$ of closed subspaces of V. We suppose that $\{V_h\}_{h>0}$ satisfies the following *internal approximation* condition:

$$\exists \mathcal{V} \subset V \text{ s.t. } \overline{\mathcal{V}} = V \text{ and } r_h : \mathcal{V} \longrightarrow V_h \text{ s.t. } \lim_{h\to 0} \|r_h v - v\| = 0, \quad \forall\, v \in \mathcal{V}. \tag{3.1}$$

In practice, h is given by a sequence and the subspaces, $\{V_h\}_{h>0}$, are finite-dimensional. We approximate problem (P) by the following problem:

Definition 3.1.3 (The internal approximation to the linear variational problem)**.** *Find $u_h \in V_h$ such that*

$$\mathfrak{a}(u_h, v_h) = L(v_h), \quad \forall\, v_h \in V_h \tag{P_h}$$

We expect (P_h) to be much easier to solve than (P). From the Lax-Milgram Theorem it follows that (P_h) also has a unique solution. These solutions, $\{u_h\}_{h>0}$, under the hypotheses on V, $\mathfrak{a}$ and L above, and Condition (3.1), converge to the unique solution u of Problem P. This follows from the Cea's Lemma [25, Appendix I Lemma 3.1] below and a simple application of it, which we state here as Theorem 3.1.5.

Lemma 3.1.4 (Cea's Lemma)**.** *Let u be the solution to Problem P, and for every $h > 0$, let u_h be the solution to Problem P_h. We then have,*

$$\|u - u_h\| \leq \frac{C}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|$$

Theorem 3.1.5. [25, App. I Theorem 3.2] *Suppose that $\{V_h\}_{h>0}$ obeys the internal approximation condition* (3.1)*. Let u be the solution to Problem P, and for every $h > 0$ let u_h be the solution to Problem P_h. We then have,*

$$\lim_{h\to 0} \|u - u_h\| = 0.$$

3.1.2 The Galerkin method

In this section we suppose that V is a separable real Hilbert space in the sense that there exists a countable subset $\mathcal{B} = \{w_j\}_{j=1}^{+\infty}$ of V, linearly independent, such that the subspace $\mathcal{V}$ of V generated by $\mathcal{B}$ is dense in V. For any integer $m \geq 1$ we define $\mathcal{B}_m = \{w_j\}_{j=1}^{m}$ and V_m as the subspace of V generated by $\mathcal{B}_m$.

Let us denote by π_m the projection operator from V to V_m. Then, since $\mathcal{B}$ is a countable basis of V it follows that $\lim_{m\to\infty} \|v - \pi_m v\| = 0$ for all $v \in V$. That is, Condition (3.1) is satisfied (with $\mathcal{V} = V$).

The Galerkin approximation of Problem P is then defined as follows:

Definition 3.1.6 (The Galerkin approximation to the linear variational problem)**.** *Find $u_m \in V_m$ such that*

$$\mathfrak{a}(u_m, v_m) = L(v_m), \quad \forall\, v_m \in V_m \tag{P$_m$}$$

By Theorem 3.1.5, with $\mathcal{V} = V$, $h = 1/m$, $V_h = V_m$, and $r_h = \pi_m$, it follows that,

$$\lim_{m\to\infty} \|u - u_m\| = 0$$

3.1.3 Reducing the Galerkin approximation problem to a linear system

Let N_m be the dimension of V_m. Problem P$_m$ is clearly equivalent to

Definition 3.1.7 (The Galerkin approximation for a specified basis)**.** *Find $u_m \in V_m$ such that*

$$\mathfrak{a}(u_m, w_i) = L(w_i), \quad \forall\, i = 1, \ldots, N_m \tag{3.2}$$

Since $u_m \in V_m$, there exists a unique vector, $\Lambda_m = (\lambda_1, \ldots, \lambda_{N_m}) \in \mathbb{R}^{N_m}$, such that $u_m = \sum_{j=1}^{N_m} \lambda_j w_j$. Thus, we find that u_m is obtained through the solution of the linear system,

$$\sum_{j=1}^{N_m} \mathfrak{a}(w_j, w_i)\lambda_j = L(w_i), \quad \forall\, i = 1, \ldots, N_m, \tag{3.3}$$

whose unknowns are the coefficients λ_j, for $j = 1, \ldots, N_m$.

The linear system (3.3) can be written as follows:

$$A_m \Lambda_m = F_m, \tag{3.4}$$

where $F_m = (L(w_1), \ldots, L(w_{N_m}))$ and the matrix A_m is defined by $A_m = (\mathfrak{a}(w_j, w_i))_{1 \le i,j \le N_m}$.

3.2 A particular basis

Let $X_0, X_1 \in \mathbb{R}$ with $X_0 < X_1$, and $\mathscr{O}$ be of the form $(X_0, X_1) \times (0, \infty)$. We now focus on the Hilbert space $V = H^1_0(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$, and consider $\mathfrak{a}$, the Heston bilinear form of Definition 2.1.11 with $\gamma = 0$. We already know that $\mathfrak{a}$ is continuous, by Proposition 2.1.17, and coercive, by Proposition 2.1.19.

Let h_x and h_y be positive numbers, and consider the partitions,

$$\begin{aligned} \mathcal{P}_x &: \{X_0 = x_1 < x_2 < x_3 < \cdots < x_M = X_1\} \\ \mathcal{P}_y &: \{0 = y_1 < y_2 < y_3 < \cdots < y_N < \cdots\}, \end{aligned} \tag{3.5}$$

where $x_{i+1} = x_i + h_x$ and $y_{j+1} = y_j + h_y$. We will approximate $\mathscr{O}$ internally by the sequence of domains $\{\mathscr{O}_{M,N}\}_{M \ge 3, N \ge 3}$, where $\mathscr{O}_{M,N} := (x_1, x_M) \times (y_1, y_N)$. Clearly, $\Gamma_0(\mathscr{O}_{M,N}) = (X_0, X_1) \times \{0\} = \Gamma_0(\mathscr{O}) =: \Gamma_0$ and $\Gamma_1(\mathscr{O}_{M,N}) = \partial \mathscr{O}_{M,N} \cap \mathbb{H}$.

Once we chose a basis, Problem P reduces to solving the linear system of Equation (3.3). It will be important to have a basis such that the matrix that defines the linear system is easy to calculate, and the linear system itself it easy to solve. Given that $\mathscr{O}_{M,N}$ is a rectangle, we will consider a basis of tensor products of linear B-splines, "hat functions", in both directions x and y.

Recall the definition of the hat function

$$\text{hat}(x) = \begin{cases} x + 1, & -1 < x \le 0, \\ -x + 1, & 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Define $\varphi_i(x) := \text{hat}\left(\frac{x - x_i}{h_x}\right)$ and $\psi_j(y) := \text{hat}\left(\frac{y - y_j}{h_y}\right)$, and their tensor product functions $\phi_{i,j}(x, y) := \varphi_i(x)\psi_j(y)$. Notice that for each point (x_i, y_j) in the mesh we have

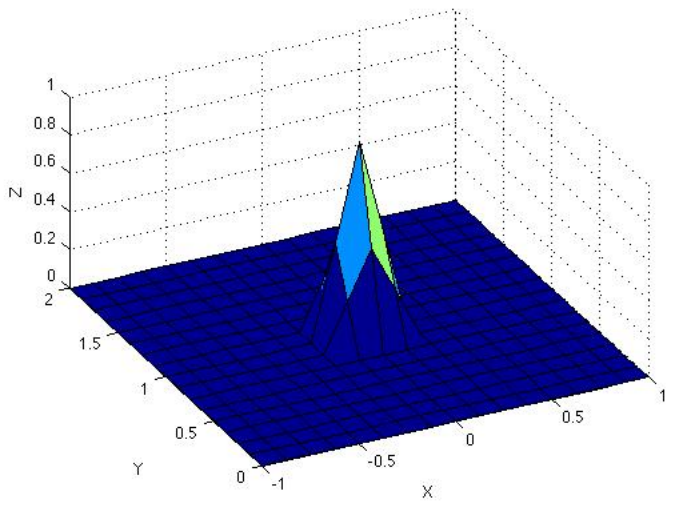


Figure 3.1: Basis function corresponding to an interior point.

$\phi_{i,j}(x_i, y_j) = 1$ and $0 < \phi_{i,j}(x, y) \leq 1$ on $(x_{i-1}, x_{i+1}) \times (y_{j-1}, y_{j+1})$, and $\phi_{i,j}(x, y) = 0$ for all other (x, y). A couple of typical $\phi_{i,j}$'s are in Figure 3.1 and Figure 3.2.

Let $\mathcal{B}_M$ be the family of functions $\{\phi_{i,j}\}_{1<i<M, j<N}$ and $\mathcal{B}$ be their union, $\mathcal{B} = \bigcup_M \mathcal{B}_M$. Notice we have excluded the indices corresponding to Γ_1. We will prove that $\mathcal{B}$ is a basis for $H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$, but first we state a preliminary lemma.

Lemma 3.2.1. *Let $u \in H^1((a, b))$ and $v \in H^1((0, \infty), \mathfrak{w}(y))$. Set $w(x, y) := u(x)v(y)$ and $\mathcal{R}_{a,b} := (a, b) \times (0, \infty)$. Then $w \in H^1(\mathcal{R}_{a,b}, \mathfrak{w})$ and*

$$||w||^2_{H^1(\mathcal{R}_{a,b}, \mathfrak{w})} \leq 2||u||^2_{W^{1,2}(a,b)}||v||^2_{H^1((0,\infty), \mathfrak{w})}$$

Proof. We only need to write out the definition of the H^1-norm again, for the case of

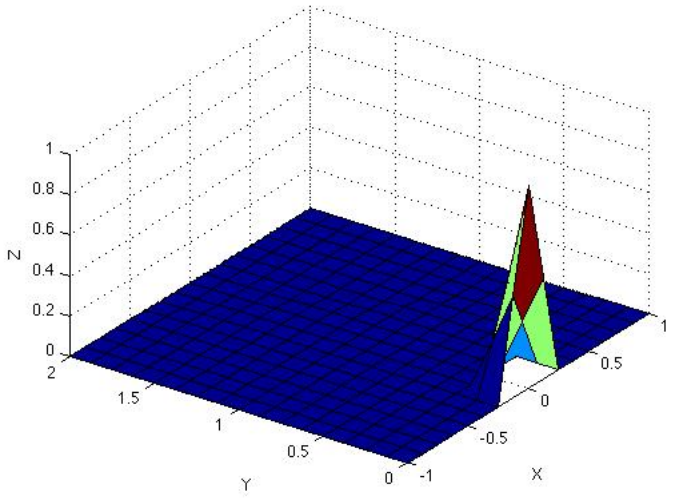


Figure 3.2: Basis function corresponding to a point along $y = 0$.

$\gamma = 0$, and make some simple estimates:

$$\begin{aligned}
||w||^2_{H^1(\mathcal{R}_{a,b},\mathfrak{w})} &= \int_{\mathcal{R}_{a,b}} (y|Dw|^2 + (1+y)w^2)\,\mathfrak{w}\,dx\,dy \\
&= \int_{\mathcal{R}_{a,b}} (y((u_x v)^2 + (uv_y)^2) + (1+y)u^2v^2)\,\mathfrak{w}\,dx\,dy \\
&\leq \int_a^b u_x^2\,dx \int_0^\infty yv^2\mathfrak{w}dy + \int_a^b u^2\,dx \int_0^\infty yv_y^2\mathfrak{w}dy + \int_a^b u^2\,dx \int_0^\infty (1+y)v^2\mathfrak{w}dy \\
&\leq ||u||^2_{H^1((a,b))} \left(\int_0^\infty yv^2\mathfrak{w}dy + \int_0^\infty yv_y^2\mathfrak{w}dy + \int_0^\infty (1+y)v^2\mathfrak{w}dy \right) \\
&= 2||u||^2_{H^1((a,b))}||v||^2_{H^1((0,\infty),\mathfrak{w})}
\end{aligned}$$

This completes the proof. □

The standard Sobolev space $H_0^1(\mathscr{O})$ is a separable Hilbert space as one can select a countable collection of linearly independent piecewise linear continuous functions as a basis. As expected, we can prove something similar for $H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$,

Proposition 3.2.2. *$\mathcal{B}$ is a basis for $H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$.*

Proof. We want to prove that $\overline{\text{span}\{\mathcal{B}\}} = H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$, where the closure is being taken with respect to H^1. We will prove both inclusions:

i) $\overline{\text{span}\{\mathcal{B}\}} \subseteq H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$: It is enough to prove that each $\phi_{i,j}$ is in $H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$. That is, each $\phi_{i,j}$ is the $H^1(\mathscr{O}, \mathfrak{w})$-limit of a sequence of smooth functions with support contained in $\mathscr{O} \cup \Gamma_0$. Given $\phi_{i,j}(x,y) \equiv \varphi_i(x)\psi_j(y)$ we will show that $\phi_{i,j}$ is the $H^1(\mathscr{O}, \mathfrak{w})$-limit of a sequence of smooth functions $\left\{\overline{\phi}_m\right\}_m$ with compact support contained in $\mathscr{O} \cup \Gamma_0$, where $\overline{\phi}_m(x,y) = \overline{\varphi}_m(x)\overline{\psi}_m(y)$, and $\{\overline{\varphi}_m(x)\}_m$ and $\{\overline{\psi}_m(y)\}_m$ have the properties that $\overline{\varphi}_m$ converges to φ_i on $H^1((X_0, X_1))$, and ψ_m converges to ψ_j on $H^1((0,\infty), \mathfrak{w}(y))$.

The existence of the sequence $\{\overline{\varphi}_m\}_m$ follows from standard Sobolev Spaces theory [14, Theorem 2, Section 5.3]. As for the sequence $\left\{\overline{\psi}_m\right\}_m$, it suffices to choose a sequence of smooth functions such that they are identical to ψ_j outside of (y_{j-1}, y_{j+1}), converging pointwise to ψ_j on (y_{j-1}, y_{j+1}), and their derivatives converging pointwise to the derivative of ψ_j, ψ_j'. Furthermore, require this sequence and their derivatives to be uniformly bounded. With $\left\{\overline{\psi}_m\right\}_m$ chosen this way, one can see that $\left\{\overline{\psi}_m\right\}_m$ actually converges to ψ_j on $H^1((0,\infty), \mathfrak{w})$. Indeed, it follows from

$$\left\|\psi_j - \overline{\psi}_m\right\|^2_{H^1((0,\infty),\mathfrak{w})} = \int_0^\infty \left(y \left| D\psi_j - D\overline{\psi}_m \right|^2 + (1+y)\left(\psi_j - \overline{\psi}_m\right)^2 \right) \mathfrak{w} dy,$$

by the Lebesgue Dominated Convergence Theorem given the bounded pointwise convergences noted above, and the fact that both $y^\beta e^{-\mu y}$ and $y^{\beta-1}e^{-\mu y}$ are in $L^1(0,\infty)$ since $\beta > 0$ and $\mu > 0$.

Notice that $\overline{\varphi}_m\overline{\psi}_m$ has compact support contained in $\mathscr{O} \cup \Gamma_0$. Notice also that none of the basis functions with peak on a node on the Γ_1-boundary were included in $\mathcal{B}$. Thus, by Lemma 3.2.1, it follows that,

$$\begin{aligned}
\left\|\phi_{i,j} - \overline{\varphi}_m\overline{\psi}_m\right\|_{H^1(\mathscr{O},\mathfrak{w})} &\leq ||\varphi_i\psi_j - \overline{\varphi}_m\psi_j||_{H^1(\mathscr{O},\mathfrak{w})} + \left\|\overline{\varphi}_m\psi_j - \overline{\varphi}_m\overline{\psi}_m\right\|_{H^1(\mathscr{O},\mathfrak{w})} \\
&= ||(\varphi_i - \overline{\varphi}_m)\psi_j||_{H^1(\mathscr{O},\mathfrak{w})} + ||\overline{\varphi}_m(\psi_j - \overline{\psi}_m)||_{H^1(\mathscr{O},\mathfrak{w})} \\
&\leq \sqrt{2}||\varphi_i - \overline{\varphi}_m||_{H^1((X_0,X_1))}||\psi_j||_{H^1((0,\infty),\mathfrak{w})} \\
&\quad + \sqrt{2}||\overline{\varphi}_m||_{H^1((X_0,X_1))}\left\|\psi_j - \overline{\psi}_m\right\|_{H^1((0,\infty),\mathfrak{w})}
\end{aligned}$$

The main assertion follows from this.

ii) $H_0^1(\mathscr{O}\cup\Gamma_0, \mathfrak{w}) \subseteq \overline{\text{span}\{\mathcal{B}\}}$: It is sufficient to prove that any smooth function with compact support contained in $\mathscr{O} \cup \Gamma_0$ is the $H^1(\mathscr{O}, \mathfrak{w})$- limit of a sequence of functions

in $span\{\mathcal{B}\}$. Let $w \in C_0^\infty(\mathscr{O} \cup \Gamma_0)$. Consider a mesh for $\mathscr{O} = (X_0, X_1) \times (0, \infty)$ defined by Equation (3.5). To simplify notation, without loss of generality, we may assume that $h_x = h_y = h$. Let

$$K_h := \{(i,j) | (x_i, y_j) \in \operatorname{supp} w\}$$

be the set of pair of indices corresponding to mesh points in the support of w. Clearly K_h is finite. For each $(i,j) \in K_h$, consider the basis function $\phi_{i,j}$ and define

$$\overline{w}_h(x,y) := \sum_{(i,j)\in K_h} w(x_i, y_j)\phi_{i,j}(x,y), \quad \forall\, (x,y) \in \mathscr{O} \tag{3.6}$$

Just as in part i), by the Lebesgue Dominated Convergence Theorem it is enough to show that the sequence $\{\overline{w}_h\}_h$ is uniformly bounded and converges pointwise to w, and that the same is true for their derivatives, while converging to Dw. Clearly $\overline{w}_h$ is uniformly bounded, since w is continuous and has compact support, and all of the $\phi_{i,j}$ are bounded between 0 and 1. We only need to prove that $\{D\overline{w}_h\}_{h>0}$ is uniformly bounded and converges pointwise to Dw as $h \to 0$.

Let us first focus on the derivative $(\overline{w}_h)_x$. Notice that the sum in Equation (3.6) is actually of at most four terms. Let $(x,y) \in \operatorname{supp} w$ be a point not on the mesh. Then (x,y) is in the interior of a rectangle of the form $(x_{i_1}, y_{j_1}) \times (x_{i_2}, y_{j_2}) \subset \mathscr{O} \cup \Gamma_0$, where $x_{i_2} = x_{i_1} + h$ and $y_{j_2} = y_{j_1} + h$. Then,

$$(\overline{w}_h)_x(x,y) = \sum_{i_1,i_2} \left(\varphi_i'(x) \sum_{j_1,j_2} w(x_i, y_j)\psi_j(y) \right).$$

Now, notice that for such (x,y), $\varphi_{i_1}'(x) = -1/h$ and $\varphi_{i_2}'(x) = 1/h$. Thus, expanding the sums we obtain,

$$(\overline{w}_h)_x(x,y) = \sum_{j_1,j_2} \left(\frac{w(x_{i_2}, y_j) - w(x_{i_1}, y_j)}{h} \right) \psi_j(y).$$

By the Mean Value Theorem, there exists a point $x_{h,j} \in (x_{i_1}, x_{i_2})$, that depends on h and y_j, such that,

$$(\overline{w}_h)_x(x,y) = \sum_{j_1,j_2} w_x(x_{h,j}, y_j)\psi_j(y).$$

But $\psi_{j_1}(y) + \psi_{j_2}(y) = 1$, hence $(\overline{w}_h)_x(x,y)$ is the weighted average of two numbers that

converge to $w_x(x,y)$ as the mesh gets finer. From this it follows that $\{(\overline{w}_h)_x\}_h$ is uniformly bounded, and it converges pointwise to w_x. Similarly for $\{(\overline{w}_h)_y\}_h$. Therefore, $\{\overline{w}_h\}_h$ converges to w on $H^1(\mathscr{O}, \mathfrak{w})$.

This completes the proof. □

For any integers $M, N \geq 3$, we define $\mathcal{B}_{M,N} := \{\phi_{i,j}\}_{i=2,j=1}^{M-1,N-1}$, and $V_{M,N}$ as the finite-dimensional subspace of $H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$ generated by $\mathcal{B}_{M,N}$. Notice the basis functions with $i = 1$ or M, or $j = N$, were not included. Since $\mathcal{B}$ is a basis for $H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$, it follows that the family of closed subspaces, $\{V_{M,N}\}_{M\geq 3, N\geq 3}$, provides and internal approximation of $H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$, as it satisfies Condition (3.1). The Galerkin approximation of Problem 2.12, as stated in Section 3.1.3 is then given by,

Find $u \in V_{M,N}$ such that

$$\mathfrak{a}(u, v) = (f, v)_{L^2(\mathscr{O}_{M,N}, \mathfrak{w})} \quad \text{for all } v \in V_{M,N} \tag{3.7}$$

Problem 3.7, the Galerkin approximation, for the specified basis, can be formulated as,

Find $u \in V_{M,N}$ such that

$$a\left(u, \phi_{k,l}\right) = \left(f, \phi_{k,l}\right)_{L^2(\mathscr{O}, \mathfrak{w})}, \quad \text{for all } k = 2, ..., M-1 \text{ and } l = 1, ..., N-1. \tag{3.8}$$

For $u \in V_{M,N}$, there exists a unique vector $\Lambda_{M,N} = (\lambda_{i,j})_{i=1,j=1}^{M,N} \in \mathbb{R}^{MN}$, such that

$$u^{M,N} = \sum_{i,j}^{MN} \lambda_{i,j} \phi_{i,j}. \tag{3.9}$$

By combining Equations (3.8) and (3.9), we find that $u^{M,N}$ is obtained from the solution to the linear system,

$$\sum_{i,j} \mathfrak{a}(\phi_{i,j}, \phi_{k,l}) \lambda_{i,j} = (f, \phi_{k,l})_{L^2(\Omega, \mathfrak{w})}, \tag{3.10}$$

whose unknown is the vector $\Lambda_{M,N}$. Let us enumerate the nodes (i,j) with a mapping $s = I(i,j)$, to be specified later, and write Equation (3.10) as

$$\sum_{s=1}^{(M-2)(N-1)} \mathfrak{a}(\phi_s, \phi_t) \lambda_s = (f, \phi_t)_{L^2(\Omega, \mathfrak{w})}, \tag{3.11}$$

where $\phi_s = \phi_{I(i,j)} = \phi_{i,j}$ and $\phi_t = \phi_{I(k,l)} = \phi_{k,l}$. Suitable mappings I make the linear system sparser. For our implementation we use the mapping illustrated in Figure 3.3.

We enumerate the nodes horizontally from left to right, and from bottom to top, but every other row. To solve this problem we need to calculate the $(M-2)(N-1)$ by $(M-2)(N-1)$ matrix $A_{M,N} = \{\mathfrak{a}(\phi_s, \phi_t)\}_{s,t}$.

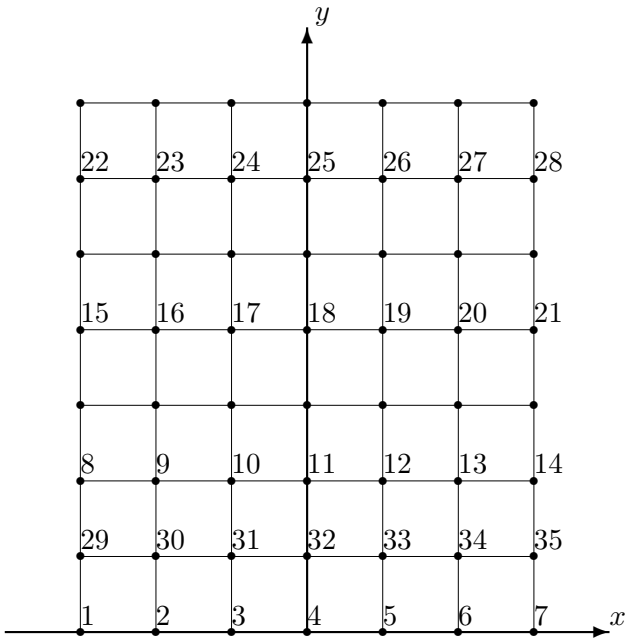


Figure 3.3: Example of a mesh for the finite-element method with $M = 7$ and $N = 8$.

3.3 Calculation of the matrix

We are interested in calculating $\mathfrak{a}(\phi_s, \phi_t)$ for $\gamma = 0$, since our integrals are now restricted to a domain bounded in the x-direction. From Definition 2.1.11, we have:

$$\begin{aligned}\mathfrak{a}(u,v) := &\frac{1}{2}\int_{\mathscr{O}} u_x v_x y \,\mathfrak{w}\, dx\, dy + \frac{1}{2}\rho\sigma \int_{\mathscr{O}} u_y v_x y \,\mathfrak{w}\, dx\, dy + \frac{1}{2}\rho\sigma \int_{\mathscr{O}} u_x v_y y \,\mathfrak{w}\, dx\, dy \\ &+ \frac{1}{2}\sigma^2 \int_{\mathscr{O}} u_y v_y y \,\mathfrak{w}\, dx\, dy - a_1 \int_{\mathscr{O}} u_x v y \,\mathfrak{w}\, dx\, dy - b_1 \int_{\mathscr{O}} u_x v \,\mathfrak{w}\, dx\, dy \\ &+ r \int_{\mathscr{O}} uv \,\mathfrak{w}\, dx\, dy.\end{aligned} \tag{3.12}$$

We will compute formulas for each of the integrals above. Let

$$u(x,y) = \phi_s(x,y) = \phi_{i,j}(x,y) = \varphi_i(x)\psi_j(y),$$

and

$$v(x,y) = \phi_t(x,y) = \phi_{k,l}(x,y) = \varphi_k(x)\psi_l(y).$$

Hence,

$$\begin{aligned}
u_x(x,y) &= =: \frac{1}{h_x}\mathcal{I}_i(x)\psi_j(y) = \frac{1}{h_x}\psi_j(y) \times \begin{cases} 1, & x_{i-1} < x < x_i, \\ -1, & x_i < x < x_{i+1}, \\ 0, & \text{otherwise}, \end{cases} \\
v_x(x,y) &= \frac{1}{h_x}\mathcal{I}_k(x)\psi_l(y), \\
u_y(x,y) &= =: \frac{1}{h_y}\varphi_i(x)\mathcal{J}_j(y) = \frac{1}{h_y}\varphi_i(x) \times \begin{cases} 1, & y_{j-1} < y < y_j, \\ -1, & y_j < y < y_{j+1}, \\ 0, & \text{otherwise}, \end{cases} \\
v_y(x,y) &= \frac{1}{h_y}\varphi_k(x,y)\mathcal{J}_l(y).
\end{aligned}$$

The integrals present in Equation (3.12) can then be written as,

$$\int_{\mathscr{O}} u_x v_x y\, \mathfrak{w}\, dx\, dy = \frac{1}{h_x^2}\int \mathcal{I}_i\mathcal{I}_k\, dx \cdot \int \psi_j\psi_l y^\beta e^{-\mu y} dy =: X_1(i,k)Y_1(j,l), \tag{3.13}$$

$$\int_{\mathscr{O}} u_y v_x y\, \mathfrak{w}\, dx\, dy = \frac{1}{h_x}\int \varphi_i\mathcal{I}_k\, dx \cdot \frac{1}{h_y}\int \mathcal{J}_j\psi_l y^\beta e^{-\mu y} dy =: X_2(i,k)Y_2(j,l), \tag{3.14}$$

$$\int_{\mathscr{O}} u_x v_y y\, \mathfrak{w}\, dx\, dy = \frac{1}{h_x}\int \mathcal{I}_i\varphi_k\, dx \cdot \frac{1}{h_y}\int \psi_j\mathcal{J}_l y^\beta e^{-\mu y} dy =: X_3(i,k)Y_3(j,l), \tag{3.15}$$

$$\int_{\mathscr{O}} u_y v_y y\, \mathfrak{w}\, dx\, dy = \int \varphi_i\varphi_k\, dx \cdot \frac{1}{h_y^2}\int \mathcal{J}_j\mathcal{J}_l y^\beta e^{-\mu y} dy =: X_4(i,k)Y_4(j,l), \tag{3.16}$$

$$\int_{\mathscr{O}} u_x v y\, \mathfrak{w}\, dx\, dy = \frac{1}{h_x}\int \mathcal{I}_i\varphi_k\, dx \cdot \int \psi_j\psi_l y^\beta e^{-\mu y} dy =: X_5(i,k)Y_5(j,l), \tag{3.17}$$

$$\int_{\mathscr{O}} u_x v\, \mathfrak{w}\, dx\, dy = \frac{1}{h_x}\int \mathcal{I}_i\varphi_k\, dx \cdot \int \psi_j\psi_l y^{\beta-1} e^{-\mu y} dy =: X_6(i,k)Y_6(j,l), \tag{3.18}$$

$$\int_{\mathscr{O}} uv\, \mathfrak{w}\, dx\, dy = \int \varphi_i\varphi_k\, dx \cdot \int \psi_j\psi_l y^{\beta-1} e^{-\mu y} dy =: X_7(i,k)Y_7(j,l). \tag{3.19}$$

We will obtain closed formulas for all of the $X(i,k)$-integrals in terms of the x-coordinates of the nodes, and write the $Y(j,l)$-integrals in terms of some fundamental integrals that will depend only on the y-coordinates of the nodes and the constants β and μ. Let us

start with the $X(i,k)$- integrals:

$$X_1(i,k) = \frac{1}{h_x^2}\int \mathcal{I}_i\mathcal{I}_k\,dx = \begin{cases} -1/h_x, & |k-i|=1, \\ 2/h_x, & i=k, \\ 0, & \text{otherwise}, \end{cases} \tag{3.20}$$

$$X_2(i,k) = \frac{1}{h_x}\int \varphi_i\mathcal{I}_k\,dx = \begin{cases} 1/2, & k=i+1, \\ -1/2, & i=k+1, \\ 0, & \text{otherwise}, \end{cases} \tag{3.21}$$

$$X_3(i,k) = \frac{1}{h_x}\int \mathcal{I}_i\varphi_k\,dx = X_2(k,i), \tag{3.22}$$

$$X_4(i,k) = \int \varphi_i\varphi_k\,dx = \begin{cases} h_x/6, & |k-i|=1, \\ 2h_x/3, & i=k, \\ 0, & \text{otherwise}, \end{cases} \tag{3.23}$$

$$X_5(i,k) = \frac{1}{h_x}\int \mathcal{I}_i\varphi_k\,dx = X_3(i,k), \tag{3.24}$$

$$X_6(i,k) = \frac{1}{h_x}\int \mathcal{I}_i\varphi_k\,dx = X_3(i,k), \tag{3.25}$$

$$X_7(i,k) = \int \varphi_i\varphi_k\,dx = X_4(i,k). \tag{3.26}$$

Now let us proceed with the $Y(j,l)$-integrals. Clearly if $|l-j|>1$ all of these integrals are identically zero, so let us suppose $|l-j|\le 1$. If $l=j+1$, then

$$\begin{aligned} Y_1(j,l) = \int \psi_j\psi_l y^\beta e^{-\mu y}dy = &\frac{y_j+y_l}{h_y^2}\int_{y_j}^{y_l} y^{\beta+1}e^{-\mu y}dy - \frac{y_jy_l}{h_y^2}\int_{y_j}^{y_l} y^\beta e^{-\mu y}dy \\ &- \frac{1}{h_y^2}\int_{y_j}^{y_l} y^{\beta+2}e^{-\mu y}dy. \end{aligned} \tag{3.27}$$

If $j=l+1$, then

$$\begin{aligned} Y_1(j,l) = \int \psi_j\psi_l y^\beta e^{-\mu y}dy = &\frac{y_j+y_l}{h_y^2}\int_{y_l}^{y_j} y^{\beta+1}e^{-\mu y}dy - \frac{y_jy_l}{h_y^2}\int_{y_l}^{y_j} y^\beta e^{-\mu y}dy \\ &- \frac{1}{h_y^2}\int_{y_l}^{y_j} y^{\beta+2}e^{-\mu y}dy. \end{aligned} \tag{3.28}$$

Last, if $j = l$, we need to consider two cases. If $j = l \neq 1$ (recall $j < N$), then

$$\begin{aligned} Y_1(j,j) = \int \psi_j^2 y^\beta e^{-\mu y} dy =& \frac{1}{h_y^2} \int_{y_{j-1}}^{y_j} y^{\beta+2} e^{-\mu y} dy - \frac{2y_{j-1}}{h_y^2} \int_{y_{j-1}}^{y_j} y^{\beta+1} e^{-\mu y} dy \\ &+ \frac{y_{j-1}^2}{h_y^2} \int_{y_{j-1}}^{y_j} y^\beta e^{-\mu y} dy + \frac{1}{h_y^2} \int_{y_j}^{y_{j+1}} y^{\beta+2} e^{-\mu y} dy \\ &- \frac{2y_{j+1}}{h_y^2} \int_{y_j}^{y_{j+1}} y^{\beta+1} e^{-\mu y} dy + \frac{y_{j+1}^2}{h_y^2} \int_{y_j}^{y_{j+1}} y^\beta e^{-\mu y} dy. \end{aligned} \tag{3.29}$$

If $j = l = 1$, then

$$\begin{aligned} Y_1(1,1) = Y_1(j,j) =& \frac{1}{h_y^2} \int_{y_j}^{y_{j+1}} y^{\beta+2} e^{-\mu y} dy - \frac{2y_{j+1}}{h_y^2} \int_{y_j}^{y_{j+1}} y^{\beta+1} e^{-\mu y} dy \\ &+ \frac{y_{j+1}^2}{h_y^2} \int_{y_j}^{y_{j+1}} y^\beta e^{-\mu y} dy. \end{aligned} \tag{3.30}$$

Notice that all integrals on the right-hand side of Equations (3.27-3.30) are of the form,

$$F_j(c) := \int_{y_j}^{y_{j+1}} y^c e^{-\mu y} dy, \tag{3.31}$$

for some $j \in \{1, ..., N-1\}$ and some $c \in \{\beta, \beta+1, \beta+2\}$, in this case. In terms of the fundamental integrals, F_j, we can write $Y_1(j,l)$ as follows:

$$Y_1(j,l) = \begin{cases} \frac{y_j + y_l}{h_y^2} F_j(\beta+1) - \frac{y_j y_l}{h_y^2} F_j(\beta) - \frac{1}{h_y^2} F_j(\beta+2), & \text{for} \quad l = j+1, \\ \frac{y_j + y_l}{h_y^2} F_l(\beta+1) - \frac{y_j y_l}{h_y^2} F_l(\beta) - \frac{1}{h_y^2} F_l(\beta+2), & \text{for} \quad j = l+1, \\ \frac{1}{h_y^2} F_{j-1}(\beta+2) - \frac{2y_{j-1}}{h_y^2} F_{j-1}(\beta+1) + \frac{y_{j-1}^2}{h_y^2} F_{j-1}(\beta) \\ + \frac{1}{h_y^2} F_j(\beta+2) - \frac{2y_{j+1}}{h_y^2} F_j(\beta+1) + \frac{y_{j+1}^2}{h_y^2} F_j(\beta), & \text{for} \quad j = l \neq 1, \\ \frac{1}{h_y^2} F_j(\beta+2) - \frac{2y_{j+1}}{h_y^2} F_j(\beta+1) + \frac{y_{j+1}^2}{h_y^2} F_j(\beta), & \text{for} \quad j = l = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3.32}$$

Similarly, for the other $Y(j,l)$-integrals we have,

$$Y_2(j,l) = \begin{cases} \frac{1}{h_y^2}(y_j F_j(\beta) - F_j(\beta+1)), & \text{for} \quad l = j+1, \\ \frac{1}{h_y^2}(y_j F_l(\beta) - F_l(\beta+1)), & \text{for} \quad j = l+1, \\ -\frac{1}{h_y^2}(y_{j-1} F_{j-1}(\beta) - F_{j-1}(\beta+1)) & \\ -\frac{1}{h_y^2}(y_{j+1} F_j(\beta) - F_j(\beta+1)), & \text{for} \quad j = l \neq 1, \\ -\frac{1}{h_y^2}(y_{j+1} F_j(\beta) - F_j(\beta+1)), & \text{for} \quad j = l = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{3.33}$$

$$Y_3(j,l) = Y_2(l,j), \tag{3.34}$$

$$Y_4(j,l) = \begin{cases} -\frac{1}{h_y^2} F_j(\beta), & \text{for} \quad l = j+1, \\ -\frac{1}{h_y^2} F_l(\beta), & \text{for} \quad j = l+1, \\ \frac{1}{h_y^2}(F_{j-1}(\beta) + F_j(\beta)), & \text{for} \quad j = l \neq 1, \\ \frac{1}{h_y^2} F_j(\beta), & \text{for} \quad j = l = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{3.35}$$

$$Y_5(j,l) = Y_1(l,j), \tag{3.36}$$

$$Y_6(j,l) = Y_1(j,l) \restriction_{\beta \to \beta - 1}, \tag{3.37}$$

$$Y_7(j,l) = Y_6(j,l). \tag{3.38}$$

3.4 Convergence of finite-element scheme

In the case of strictly elliptic differential operators one can show that an internal approximation scheme, like a finite-element scheme, converges to the unique solution of the associated variational problem, in the norm of the underlying Hilbert space (see [25, Appendix I] for a good brief introduction to this subject). Coerciveness of the bilinear form is a fundamental hypothesis that the Heston bilinear form may not satisfy in general, but as we proved in Proposition 2.1.19, if the domain $\mathscr{O}$ is bounded in the x-direction, and satisfies Hypotheses 2.1.3 and 2.1.5, then the Heston bilinear form is actually coercive on $H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$. Since $\mathfrak{a}$ is continuous by Proposition 2.1.17 and the

$\{V_{M,N}\}_{M\geq 3,N\geq 3}$ provide an internal approximation of $H_0^1(\mathscr{O}\cup\Gamma_0,\mathfrak{w})$, then by Theorem 3.1.5 (via Cea's Lemma) it follows that the sequence of functions $\{u_{M,N}\}_{M\geq 3,N\geq 3}$ converges strongly to the solution u of Problem 2.12.

Cea's Lemma is quite useful as well for obtaining error estimates. In our case it tells us that the approximation error is actually of order $O\left(1/M+1/N\right)$.

Hypothesis 3.4.1 ($H^2(\mathscr{O})$ regularity hypothesis on u)**.** *Let u be the solution to the Heston linear variational problem 2.12 on $\mathscr{O}$. We assume that*

$$u\in H^2(\mathscr{O}).$$

Notice this is the non-weighted Sobolev space.

This hypothesis is key for our rate of convergence Lemma below.

Lemma 3.4.2 (Rate of convergence)**.** *Let $\{u_{M,N}\}_{M\geq 3,N\geq 3}$ be the sequence of functions obtained by the finite-element method, that converges to $u\in H_0^1(\mathscr{O}\cup\Gamma_0,\mathfrak{w})$ as M,N become large. Assume $\beta>1$. If $u\in H^2(\mathscr{O})$, then the rate of convergence at which the $u_{M,N}$'s converge to u is of order 1. That is,*

$$||u-u_{M,N}||_{H^1(\mathscr{O},\mathfrak{w})}=O\left(\frac{1}{M}+\frac{1}{N}\right) \tag{3.39}$$

To prove this Lemma we will use Cea's lemma combined with some error estimates of how well linear combinations of tensor products of B-splines approximate $H_0^1(\mathscr{O}\cup\Gamma_0,\mathfrak{w})$ functions in the $H^1(\mathscr{O},\mathfrak{w})$-norm. Let us first rewrite Cea's lemma for our context:

Lemma 3.4.3. *Let u be the solution to the variational problem* (2.12) *and $u_{M,N}$ be the solution to the finite-dimensional approximation Problem 3.7. We then have*

$$||u-u_{M,N}||_{H^1(\mathscr{O},\mathfrak{w})}\leq\frac{C}{\alpha}\inf_{v_{M,N}\in V_{M,N}}||u-v_{M,N}||_{H^1(\mathscr{O},\mathfrak{w})}, \tag{3.40}$$

where C is the constant of the continuity estimate in Proposition 2.1.17 and α is the coercivity constant of $\mathfrak{a}$ in Proposition 2.1.19.

In view of Equation (3.40), to get a rate of convergence for the $u_{M,N}$'s, we need to know how well functions in $H_0^1(\mathscr{O}\cup\Gamma_0,\mathfrak{w})$ can be approximated by functions in the

finite-dimensional space $V_{M,N}$. For that, we will state and use in the next section some error estimates for how well $H^1(\mathscr{O})$ (the non-weighted Sobolev space) functions can be approximated by tensor products of linear B-splines, as shown in Larry L. Schumaker's book Spline Functions: Basic Theory [35, Chapter 12].

3.4.1 Error estimate for $H^1(\mathscr{O}, \mathfrak{w})$ functions with respect to tensor products of linear B-splines

The basis $\{\phi_{i,j}\}_{i,j}$ of $V_{M,N}$ is the tensor product basis of bases $\{\varphi_i\}_i$ and $\{\psi_j\}_j$, which are B-Splines (of order 1) as defined in [35, Chapter 12]. They are called tensor product B-splines (of degree 1).

Definition 3.4.4 (Regular set of multi-indices)**.** *Let $d \in \mathbb{Z}$ be positive. Denote by e_i the unit vector in the i-th direction, $(0, ..., 1, ..., 0) \in \mathbb{Z}^d$. Let $I \subset \mathbb{Z}^d_+$ be a set of multi-indices. Then we say that I is regular provided:*

a) For some nonnegative integers $r_1, r_2, ...r_d$, we have $r_i e_i \in I$ for every i;

b) If $\alpha \in I$, then there is no $\beta \in I$ such that $\alpha < \beta$.

Definition 3.4.5 (Generalized Sobolev spaces)**.** *Let Ω be a bounded open subset in $\mathbb{R}^d$ and let $I \subset \mathbb{Z}^d_+$ be a regular set of multi-indices. The vector space,*

$$L^I_p(\Omega) := \left\{ f \in L^p(\Omega) : ||f||_{L^I_p(\Omega)} < \infty \right\},$$

is a Banach space, where $||f||_{L^I_p(\Omega)} := ||f||_{L^p(\Omega)} + \sum_{\alpha \in I} ||D^\alpha f||_{L^p(\Omega)}$ and $1 \leq p \leq \infty$.

The classic Sobolev spaces, $W^{k,p}(\Omega)$, are obtained from Definition 3.4.5 by considering the set of multi-indices $I = \{\alpha \in \mathbb{Z}^d_+ : 0 < |\alpha| = k\}$ [2, Corollary 4.16], but this generalization allows Schumaker to introduce other Banach spaces which are better suited for the derivation of error estimates for tensor products of linear B-splines.

Definition 3.4.6 (Tensor-product Sobolev spaces)**.** *Let $r_1, ..., r_d$ be positive integers and consider $I = \{r_1 e_1, ..., r_d e_d\}$, where e_i denotes the unit vector in the i-th direction, so I is a regular set of multi-indices. For $r := (r_1, ..., r_d)$, we call $L^r_p(\Omega) := L^I_p(\Omega)$ a* tensor Sobolev space.

In [35, Theorem 12.7] it is proved that functions in the tensor-product Sobolev space, $L_p^r(R)$, with R a rectangle in $\mathbb{R}^d$, can be approximated by taking linear combinations of tensor products of B-splines (of generalized order $r = (r_1, ..., r_d)$) and an error estimate is provided. Hence, in the context of our dissertation, by [35, Theorem 12.7], it follows that for each $f \in L_p^{(1,1)}(\mathscr{O})$, there exists a constant $\mathcal{C}$ such that

$$\|f - Q(f)\|_{L^p(\mathscr{O})} \leq \mathcal{C}\left(h_x \|D_x f\|_{L^p(\mathscr{O})} + h_y \|D_y f\|_{L^p(\mathscr{O})}\right), \tag{3.41}$$

where $Q(\cdot)$ is an operator defined by a linear combination of B-splines of order 1 with coefficients given by evaluating the function f itself on an underlying mesh defined on $\mathscr{O}$ by h_x and h_y. We notice here that Equation (3.41) holds for $1 \leq p \leq \infty$, particularly when $p = 2$ which is the choice we will make.

The function $Q(f)$ is known as the *quasi-interpolant of* f. Much can be said about the operator, Q, and we refer the reader to Schumaker's book. Here we point out that $Q(f)$ is just a linear combination of our basic B-spline functions, $\phi_{i,j}$, introduced earlier.

The proof of [35, Theorem 12.7] relies on the fact that any smooth function f can be approximated quite well by a (tensor) Taylor expansion [35, Theorem 13.8]. The derivatives of such a Taylor polynomial are also good of approximations to the derivatives of f [35, Theorem 13.20]. Hence, we also have that if $f \in L_p^{(2,2)}(\mathscr{O})$, then

$$\|D(f - Q(f))\|_{L^p(\mathscr{O})} \leq \mathcal{C}\left(h_x \|D_{xx} f\|_{L^p(\mathscr{O})} + h_y \|D_{yy} f\|_{L^p(\mathscr{O})}\right), \tag{3.42}$$

for the same constants as in Equation (3.41).

3.4.2 Proof of convergence

We are ready to use the estimates in Equations (3.41 - 3.42), for $p = 2$, to prove Lemma 3.4.2.

Proof of Lemma 3.4.2. Without loss of generality we will assume that $h_x = h_y = h$ and denote $u_{M,N}$ by u_h. We want to estimate $||u - u_h||^2_{H^1(\mathscr{O},\mathfrak{w})}$. For that purpose, to apply Cea's Lemma, we will first estimate $||u - v_h||^2_{H^1(\mathscr{O},\mathfrak{w})}$ when v_h is the linear combination of tensor product B-splines, $\phi_{i,j}$, with coefficients $\lambda_{i,j} = u(x_i, y_j)$, as in Equation (3.9).

By Definition 2.1.9,

$$\begin{aligned} ||u-v_h||^2_{H^1(\mathscr{O},\mathfrak{w})} &= \int_{\mathscr{O}} y^\beta (u-v_h)^2 e^{-\mu y}\,dx\,dy + \int_{\mathscr{O}} y^{\beta-1}(u-v_h)^2 e^{-\mu y}\,dx\,dy \\ &\quad + \int_{\mathscr{O}} y^\beta |D(u-v_h)|^2 e^{-\mu y}\,dx\,dy \\ &\leq ||u-v_h||^2_{L^2(\mathscr{O})} \left\| \left(y^\beta + y^{\beta-1}\right) e^{-\mu y} \right\|_{L^\infty(\mathscr{O})} \\ &\quad + ||D(u-v_h)||^2_{L^2(\mathscr{O})} \left\| y^\beta e^{-\mu y} \right\|_{L^\infty(\mathscr{O})}. \end{aligned}$$

Both expressions above are finite for $\beta > 1$. Thus, from Equations (3.41) and (3.42), since $u \in H^2(\mathscr{O})$, there exists a constant $\mathcal{C} > 0$, depending only on the domain $\mathscr{O}$ and parameters β and μ, such that

$$\begin{aligned} ||u-v_h||^2_{H^1(\mathscr{O},\mathfrak{w})} &\leq \mathcal{C}h^2 \left(||u_x||^2_{L^2(\mathscr{O})} + ||u_y||^2_{L^2(\mathscr{O})} \right) + \\ &\quad \mathcal{C}h^2 \left(||u_{xx}||^2_{L^2(\mathscr{O})} + ||u_{yy}||^2_{L^2(\mathscr{O})} \right) \\ &\leq \mathcal{C}h^2 \left(||u_x||^2_{L^2(\mathscr{O})} + ||u_y||^2_{L^2(\mathscr{O})} + ||u_{xx}||^2_{L^2(\mathscr{O})} + ||u_{yy}||^2_{L^2(\mathscr{O})} \right). \end{aligned} \tag{3.43}$$

Thus, by Cea's Lemma [25, Appendix I Lemma 3.1],

$$||u-u_h||_{H^1(\mathscr{O},\mathfrak{w})} \leq Ch, \tag{3.44}$$

for some constant $C(u)$ that depends on u, the parameters β and μ, and the diameter of the domain $\mathscr{O}$. That is, this finite-element method has order 1. □

Remark 3.4.7. *If we had $u \in W^{2,\infty}(\mathscr{O})$, then the same rate of convergence could be obtained for all $\beta > 0$, but clearly this is a stronger hypothesis. The proof would be identical to that of Lemma 3.4.2, except that we would use Equations* (3.41) *and* (3.42) *with $p = \infty$ instead of $p = 2$. In fact, we can weaken the hypothesis on u to be $u \in H^2(\mathscr{O},\mathfrak{w})$, for any $\beta > 0$, by verifying that Schumaker's results on $L_2^{(2,2)}(\mathscr{O})$-convergence [35, Theorem 13.8], are also valid for the analogous tensor-product Sobolev spaces $L_2^{(2,2)}(\mathscr{O},\mathfrak{w})$.*

3.5 Numerical results for the solution of the elliptic Heston boundary value problem via the finite-element method

Consider Problem 2.1,

$$\begin{cases} Au = f & \text{a.e. in } \mathscr{O} \\ u = g & \text{on } \Gamma_1 \end{cases} \tag{3.45}$$

for the Heston operator A of Equation (2.2) with coefficients given by,

$$\theta = 0.1, \quad \sigma = 0.4, \quad \kappa = 1.0, \quad r = 0.05, \quad q = 0.01, \quad \rho = -0.7.$$

Hence, $\beta = 2\kappa\theta/\sigma^2 = 1.25$, $\mu = 2\kappa/\sigma^2 = 12.5$, $a_1 = \kappa\rho/\sigma - 1/2 = -2.25$, and $b_1 = r - q - \kappa\theta\rho/\sigma = 0.215$. All of these coefficients satisfy our strict ellipticity Condition 2.1.2.

To illustrate our numerical results, we focus our attention on open subsets of the form $\mathscr{O} = (-L, L) \times (0, V)$ for $V < \infty$, and choose $L = 1$ and $V = 2$ without loss of generality. We also choose the source function to be $f = -1/2$ and choose the boundary condition $g : \Gamma_1 \longrightarrow \mathbb{R}$ to be the restriction to Γ_1 of the function,

$$\tilde{g}(x, y) = \left(1 - \frac{y}{V}\right)\left(ax^2 + bx + c\right) + \frac{y}{V}e^{-x/2}, \tag{3.46}$$

where $a = e^{L/2}/2, b = -\left(e^{L/2} - e^{-L/2}\right)/2L$, and $c = e^{L/2} + bL - aL^2$. The source function and boundary condition can be anything admissible for the framework outlined in previous sections, but we have chosen these functions f and $\tilde{g}$, in particular, to obtain graphs that illustrate our results well. A graph of the function $\tilde{g}$ is shown in Figure 3.4.

The elliptic Heston boundary value problem with homogeneous Dirichlet condition, Problem 2.5, in this case is,

$$\begin{cases} Au = -\dfrac{1}{2} - A\tilde{g} & \text{a.e. in } \mathscr{O}, \\ u = 0 & \text{on } \Gamma_1, \end{cases} \tag{3.47}$$

Solving this problem by the finite-element method outlined throughout this Chapter, gives us the (approximate) solution illustrated in Figure 3.5.

The graph in Figure 3.5 was obtained by considering partitions with 64 subintervals in both x and y coordinate directions. The solution, u, is identically zero along Γ_1 and

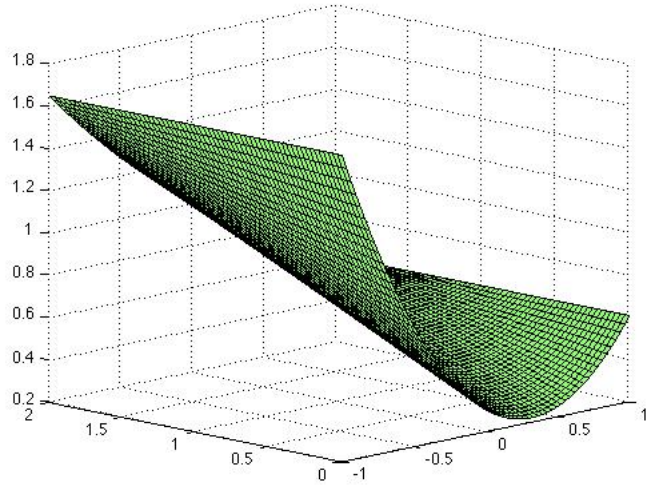


Figure 3.4: An extension, $\bar{g}$, of the boundary condition g in Equation (3.46)

takes a shape along $\Gamma_0 = (-L, L) \times \{y = 0\}$ implied by the partial differential equation in Problem 3.47, and not by the prescription of any boundary condition function. In Figure 3.6 we have the (approximate) solution to Problem 3.45 for the non-homogeneous boundary condition.

We now present numerical evidence that the rate of convergence at which the finite-element solutions converge to the solution of Problem 3.47, is at least of order one, the order proved in Lemma 3.4.2. We include graphs of the approximating finite-element solutions in Figure 3.7 for increasingly finer meshes with a summary of results compiled in Table 3.1.

Our MATLAB current code implementation takes a very long time to solve the underlying finite-dimensional problem when the mesh size becomes finer. This time is spent not only in solving the linear system in Equation (3.10), but also in generating its coefficients as we have to recalculate the fundamental integrals whenever the mesh size changes (as outlined in Section 3.3). There is definitely plenty of room to improve the performance of our algorithm if we were to implement it in a more efficient way, however that computational efficiency was not really the objective of our dissertation. We did find evidence that our theoretical results for order of convergence were verified numerically.

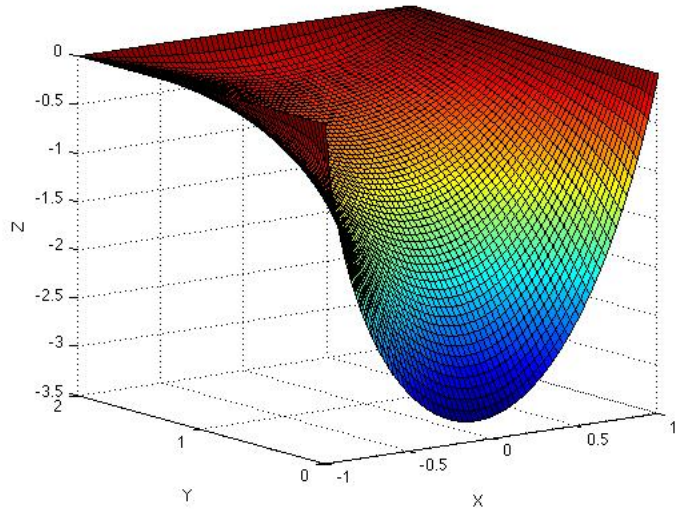


Figure 3.5: Approximate solution to the homogeneous problem, Problem 3.47

The quantity e_i we kept track of in Table 3.1 is the L^∞ norm of the difference between the i-th finite-element solution and the $(i-1)$-th one. That is,

$$e_i = \|u_i - u_{i-1}\|_{L^\infty(\mathscr{O})}. \tag{3.48}$$

In Remark 3.4.7, we noted that under a certain regularity hypothesis, namely, the true solution to the variational formulation of Problem 3.45 being in $W^{2,\infty}(\mathscr{O})$, we can use the $L^\infty(\mathscr{O})$ norm to dominate the $H^1(\mathscr{O}, \mathfrak{w})$ norm. Hence, under this hypothesis,

$$\tilde{e}_i := \|u_i - u_{i-1}\|_{H^1(\mathscr{O},\mathfrak{w})} \leq C\|u_i - u_{i-1}\|_{L^\infty(\mathscr{O})},$$

for some constant $C(u)$ that depends on the solution u, the coefficients of the Heston differential operator, and the domain $\mathscr{O}$. This implies that if the e_i had a certain order of convergence (to zero), then the increments, $\tilde{e}_i$, would have at least the same order of convergence (to zero).

By Lemma 3.4.2, we expect the $H^1(\mathscr{O}, \mathfrak{w})$-error between the finite-element approximation and the true solution to be of order 1 in $1/N$, where the finite-element approximation was calculated on a $N \times N$ mesh, thus, we expect the increments, $\tilde{e}_i$, to have at least that order of convergence as well. This is what we found empirically as noted in Table 3.1.

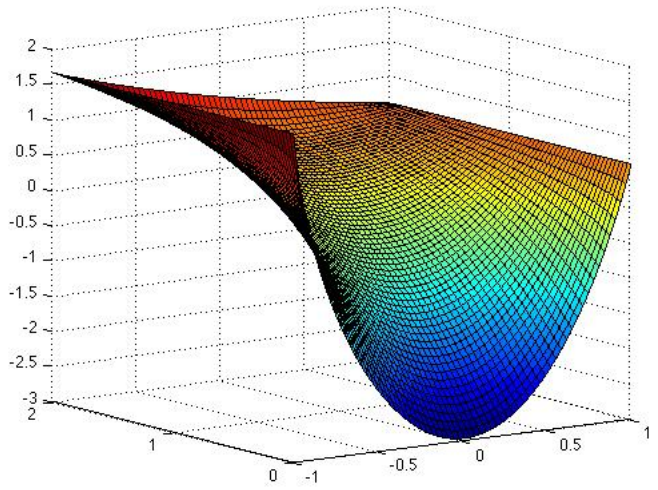


Figure 3.6: Approximate solution to the non-homogeneous problem, Problem (3.45)

The way we estimated the order of convergence of the e_i is as follows. If we assume that the L^∞-error has order of convergence α in $1/N$, then

$$\|u - u_i\|_{L^\infty(\mathscr{O})} \sim C \left(\frac{1}{N_i}\right)^\alpha.$$

Thus,

$$\begin{aligned} \tilde{e}_i = \|u_i - u_{i-1}\|_{L^\infty(\mathscr{O})} &\leq C\left(\left(\frac{1}{N_i}\right)^\alpha + \left(\frac{1}{N_{i-1}}\right)^\alpha\right) \\ &\leq 2C\left(\frac{1}{N_{i-1}}\right)^\alpha. \end{aligned}$$

Assume the increments $\tilde{e}_i$'s are given by

$$\tilde{e}_i = 2C\left(\frac{1}{N_{i-1}}\right)^\alpha$$

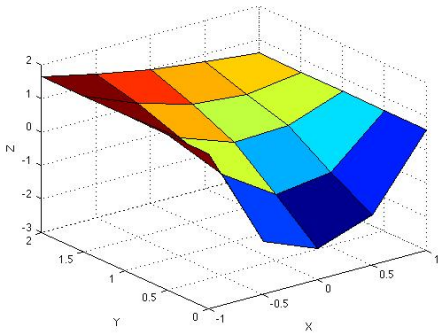
for some positive constant C. We get an estimating formula for α:

$$\alpha = \alpha_{i+1} := \frac{\log\left(\tilde{e}_{i+1}/\tilde{e}_i\right)}{\log\left(N_{i-1}/N_i\right)}$$

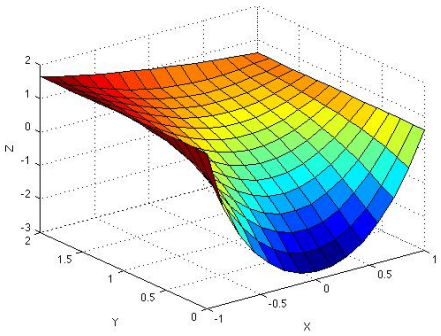
The different values for α are in Table 3.1 and they support our theoretical findings. Furthermore they hint at a better order of convergence than the one proved in Lemma 3.4.2.

Iteration (i)	Mesh ($N_i \times N_i$)	Time [seconds]	Increment (e_i)	Order (α_i)
1	4×4	1.22	———	———
2	14×14	18.80	0.861914	———
3	24×24	($\sim$ 1m) 57.25	0.237952	1.0274
4	34×34	117.41	0.118124	1.2993
5	44×44	198.79	0.071738	1.4318
6	54×54	($\sim$ 5m) 302.18	0.048634	1.5076
7	64×64	434.35	0.035355	1.5572
8	74×74	($\sim$ 10m) 574.63	0.026975	1.5923
9	84×84	745.88	0.021325	1.6186
10	94×94	949.98	0.017325	1.6391
11	104×104	($\sim$ 20m) 1189.74	0.014381	1.6555

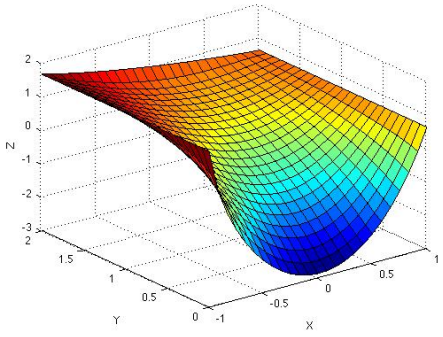
Table 3.1: Numerical results for the finite-element solution of the elliptic Heston boundary value problem



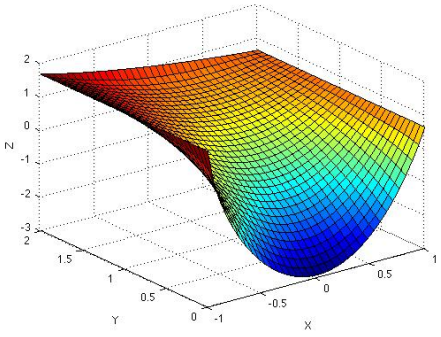
(a) 4×4 subintervals



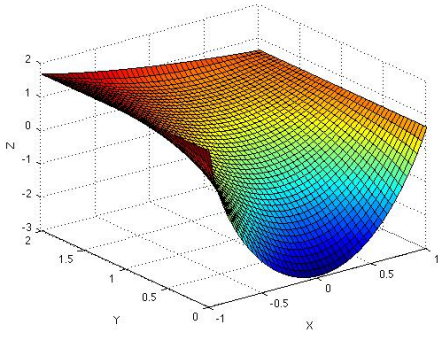
(b) 14×14 subintervals



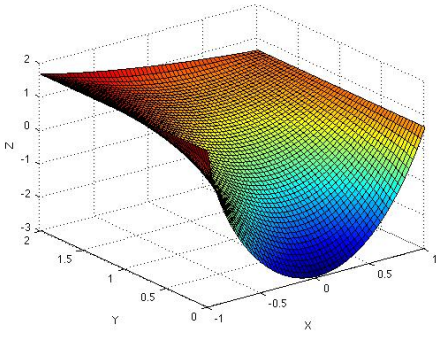
(c) 24×24 subintervals



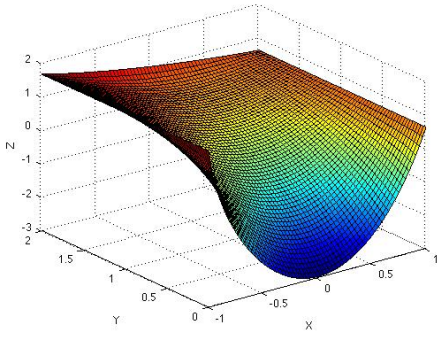
(d) 34×34 subintervals



(e) 44×44 subintervals



(f) 54×54 subintervals



(g) 64×64 subintervals

Figure 3.7: Finite-element solutions

Chapter 4

Finite-element method for the obstacle problem for the elliptic Heston operator

As with the case of an equation, we follow [25, Chapter I and II] to give a brief introduction to variational inequalities, their internal approximations, and the Galerkin method for a general basis for the underlying Hilbert space. Then we specify a basis, reduce the problem to finite-dimensional linear complementarity problems, and provide a rate of convergence of their solutions to the solution of the original problem. We do not include numerical results in this case as our MATLAB code implementation of the finite-element method takes a very long time for even a coarse mesh, but we will include numerical results when we solve the problem by the much faster finite-difference method.

4.1 Elliptic variational inequalities and their approximation

We adopt the same notation as in Section 3.1. Let V be a real Hilbert space with scalar product $(\cdot,\cdot)$ and associated norm $\|\cdot\|$, let V^* denote its topological dual space, $\mathfrak{a}$ a continuous and coercive bilinear form defined on $V \times V$, and a linear continuous functional $L \in V^*$. We consider furthermore,

i) K, a closed convex nonempty subset of V.

ii) $j : V \longrightarrow \overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$, a convex lower semicontinuous (l.s.c) and proper functional. Recall that j is proper if $j(v) > -\infty$ for all v and $j \not\equiv +\infty$.

The elliptic variational inequalities of the first and second kind read as follows:

Definition 4.1.1 (The elliptic variational inequality of the first kind)**.** *Find* $u \in K$

such that

$$\mathfrak{a}(u, v-u) \geq L(v-u), \text{ for all } v \in K \tag{P$_1$}$$

Definition 4.1.2 (The elliptic variational inequality of the second kind)**.** *Find $u \in V$ such that*

$$\mathfrak{a}(u, v-u) + j(v) - j(u) \geq L(v-u), \text{ for all } v \in K \tag{P$_2$}$$

The distinction between (P$_1$) and (P$_2$) is artificial, since (P$_1$) is a particular case of (P$_2$) by considering the following indicator functional $j = I_K(\cdot)$ of K defined by

$$I_K(v) = \begin{cases} 0, \text{ if } v \in K, \\ +\infty, \text{ if } v \notin K. \end{cases}$$

The following results is a generalization of the Lax-Milgram Theorem.

Theorem 4.1.3 (Lions-Stampacchia Theorem)**.** [25, Theorem I.3.1 and Theorem I.4.1] *The problems* (P$_1$) *and* (P$_2$) *have an unique solution.*

4.2 Internal approximation of the elliptic variational inequality of the first kind

We suppose we are given a small parameter $h > 0$, and a collection $\{V_h\}_{h>0}$ of closed subspaces of V. We are also given a family $\{K_h\}_{h>0}$ of closed convex nonempty subsets of V with $K_h \subset V_h$, for all h, such that $\{K_h\}_{h>0}$ satisfies the following internal approximation conditions:

a) If $\{v_h\}_{h>0}$ is such that $v_h \in K_h$ for all h, and $\{v_h\}_{h>0}$ is bounded in V, then the weak cluster points of $\{v_h\}_{h>0}$ belong to K.

b) There exists $\mathcal{K} \subset V$ with $\overline{\mathcal{K}} = K$, and $r_h : \mathcal{K} \longrightarrow K_h$ such that $\lim_{h\to 0} r_h v = v$ for all $v \in \mathcal{K}$.

In practice, the family $h \in (0, 1]$ is given by a sequence and the vector spaces V_h are finite-dimensional. We approximate Problem P$_1$ by the following problem:

Definition 4.2.1 (The internal approximation to the variational inequality)**.** *Find* $u_h \in K_h$ *such that*

$$\mathfrak{a}(u_h, v_h - u_h) \geq L(v_h - u_h), \text{ for all } v_h \in K_h \qquad (\mathrm{P}_{1,h})$$

Once again, we expect $(\mathrm{P}_{1,h})$ to be much easier to solve than (P_1). From the Lions-Stampacchia Theorem (Theorem 4.1.3), it follows that $(\mathrm{P}_{1,h})$ has an unique solution. These approximating solutions $\{u_h\}_h$, under the hypothesis on V, $\mathfrak{a}$ and L above, plus the internal approximation conditions a) and b), converge to the unique solution u of Problem P_1 by [25, Theorem I.5.2] as $h \to 0$.

Theorem 4.2.2. *With the above assumptions on* K *and* $\{K_h\}_{h>0}$*, we have*

$$\lim_{h \to 0} \|u - u_h\| = 0,$$

if u_h *denotes the solution of* $(\mathrm{P}_{1,h})$ *and* u *denotes the solution of* (P_1)*.*

Unlike the case of equality, the proof of convergence presented in [25, Theorem I.5.2] does not provide us with a method (a Cea's Lemma) to find its rate of convergence, however Falk proved a generalization of Cea's Lemma in his Ph.D. thesis [15], which we will now introduce, and use later in this chapter to find the rate of convergence for a particular internal approximation.

For this purpose, we need to introduce some more notation. Suppose that W is a Hilbert space which is dense in V^* and that the injection of W into V^* is continuous. Hence, we know there exists a continuous injection i of V into W^* such that $i(V)$ is dense in W^* and

$$\langle i(v), w \rangle_{W,W^*} = \langle v, w \rangle_{V,V^*} \quad \text{for all} \quad v \in V, w \in W.$$

We will henceforth identify V with a subspace of W^*, which is dense in W, through the continuous injection map. We now state Falk's general error estimate.

Theorem 4.2.3 (Falk's error estimate for the solution to a variational inequality)**.** *Let* u *and* u_h *be the solutions of* (P_1) *and* $(\mathrm{P}_{1,h})$*, respectively. Let* $A : V \longrightarrow V^*$ *denote*

the linear continuous map defined, for $\tilde{v} \in V$*, by* $\mathfrak{a}(\tilde{v}, v) = \langle A\tilde{v}, v\rangle_{V^*,V}$ *for all* $v \in V$*. Finally, suppose that* $L - Au \in W$*. Then,*

$$\|u - u_h\|^2 \leq \frac{C^2}{\alpha^2}\|u - v_h\|^2 + \frac{2}{\alpha}\|L - Au\|_W \left(\|u - v_h\|_{W^*} + \|u_h - v\|_{W^*}\right),$$

for all $v \in K$ *and all* $v_h \in K_h$*.*

4.3 A particular basis

We are interested in the approximation of Problem 2.27 for $\mathscr{O} = (X_0, X_1) \times (0, \infty)$, where X_0, X_1 are finite and $X_0 < X_1$. That is, find $u \in \mathbb{K}$ such that

$$\mathfrak{a}(u, v - u) \geq (f, v - u)_{L^2(\mathscr{O},\mathfrak{w})}, \text{ for all } v \in \mathbb{K}, \tag{4.1}$$

where $\mathbb{K} = \{v \in H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w}) | v \geq \psi \text{ a.e in } \mathscr{O}\}$, and $\psi \in H^1(\mathscr{O}, \mathfrak{w})$ with $\psi \leq 0$ on Γ_1. For the case that $\mathscr{O}$ is bounded in the x-direction, we have that $\mathfrak{a}$ is continuous and coercive, thus by the Lions-Stampacchia Theorem (Theorem 4.1.3), it follows that Problem 4.1 has a unique solution.

Let $\Sigma_{M,N}$ be the set of points in $\mathscr{O}$ defined by the partitions introduced in Equation (3.5). Consider $\mathcal{B}$, the basis of $H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$ defined in Section 3.2, and the family, $\{V_{M_N}\}_{M\geq 3, N\geq 3}$, of finite-dimensional subspaces of $H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$. For each $M, N \geq 3$, define $\mathbb{K}_{M,N} := \{v \in V_{M,N} | v \geq \psi \text{ on } \Sigma_{M,N}\}$. Clearly, the $\{\mathbb{K}_{M,N}\}_{M\geq 3,N\geq 3}$ is a family of closed convex nonempty subsets of $H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$.

To use the classical convergence results for elliptic variational inequalities we will need to verify the two properties a) and b) of Section 4.2 in the case of the family $\{\mathbb{K}_{M,N}\}_{M,N}$. Without loss of generality and to simplify notation, we assume $M = N$. Let us rewrite these two conditions again:

i) If $(v_n)_n$ is a sequence of functions such that $v_n \in \mathbb{K}_n := \mathbb{K}_{n,n}$ for each n, and the $(v_n)_n$ converges weakly to $v \in H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$ as n increases, then $v \in \mathbb{K}$.

ii) There exists $\mathcal{K} \subset H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$ and $\rho_n : \mathcal{K} \longrightarrow \mathbb{K}_n$ such that $\overline{\mathcal{K}} = \mathbb{K}$ and $\lim_{n\to\infty} \rho_n(v) = v$ in $H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$ for every $v \in \mathcal{K}$.

Since $\mathcal{B}$ is a basis of $H^1_0(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$, ii) follows by just taking $\mathcal{K}$ to be $\mathbb{K}$ itself and ρ_n to be the linear combination of basis functions in $\mathcal{B}_n$ such that $\rho_n(v)(x,y) = v(x,y)$ for every $v \in \mathbb{K}$ and $(x,y) \in \Sigma_n := \Sigma_{n,n}$. We prove i) below.

Lemma 4.3.1. *Set $\mathbb{K}_n = \mathbb{K}_{n,n} = \{v \in V_{n,n} | v \geq \psi \text{ on } \Sigma_{n,n}\}$ for $n \geq 3$. Let $\{v_n\}_{n\geq 3}$ be a sequence of functions in $H^1_0(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$ such that $v_n \in \mathbb{K}_n$ for every $n \geq 3$, and such that $\{v_n\}_{n\geq 3}$ converges weakly to $v \in H^1_0(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$. Then $v \in \mathbb{K}$.*

Proof. Recall that $\psi \in H^1(\mathscr{O}, \mathfrak{w})$. By the proof of Proposition 3.2.2, that $\mathcal{B}$ is a basis of $H^1_0(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$, there exists a sequence of *continuous* functions $\{\psi_n\}_n$ on $\overline{\mathscr{O}}$ such that ψ_n converges pointwise to ψ in $H^1(\mathscr{O}, \mathfrak{w})$. This follows by just considering linear combinations of functions in $\mathcal{B}_n$ using as coefficients the values of ψ itself on Σ_n. Thus, $\psi_n \in V_n$ and $\psi \equiv \psi_n$ on Σ_n, and ψ_n converges to ψ in $L^\infty(\mathscr{O})$ as $n \to \infty$.

Since $v_n \geq \psi$ and $\psi = \psi_n$ on Σ_n, then $v_n - \psi_n \geq 0$ on Σ_n, but that implies that $v_n - \psi_n \geq 0$ on all of $\mathscr{O}$. In particular,

$$\int_{\mathscr{O}} (v_n - \psi_n)\phi \, d\mathfrak{w} \geq 0,$$

where $d\mathfrak{w} = \mathfrak{w}\, dx\, dy$, for all non-negative smooth functions, ϕ, with compact support in $\mathscr{O}$. Taking the limit as $n \to \infty$, given that v_n converges weakly to v and ψ_n converges to ψ in $L^\infty(\mathscr{O})$ as $n \to \infty$, we have

$$\int_{\mathscr{O}} (v - \psi)\phi \, d\mathfrak{w} \geq 0, \text{ for all } \phi \in C_0(\mathscr{O}).$$

This implies that $v \geq \psi$ a.e. in $\mathscr{O}$. That is, $v \in \mathbb{K}$. □

Hence, it makes sense to consider the solutions to the approximate problems ($\mathrm{P}_{1,h}$).

4.4 Galerkin method for the chosen basis

Let $\mathfrak{a}$ be the same bilinear form as in the case of equality. The Galerkin approximation of Problem 4.1 is defined as follows:

Definition 4.4.1 (Galerkin approximation of the variational inequality)**.** *Find $u \in K_n$ such that*

$$\mathfrak{a}(u, v-u) \geq (f, v-u)_{L^2(\mathscr{O}_n, \mathfrak{w})} \text{ for all } v \in K_n \qquad (\mathrm{P}_n)$$

Now since $\mathfrak{a}$ is continuous and coercive, then Problem P_n has a unique solution, and since the family $\{\mathbb{K}_n\}_{n\geq 3}$ satisfies conditions i) and ii) of Section 4.3, then by Theorem 4.2.2,

$$\lim_{n\to\infty} \|u - u_n\|_{H^1(\mathscr{O},\mathfrak{w})} = 0, \tag{4.2}$$

where u is the unique solution to Problem 4.1 and each u_n is the unique solution to each Problem P_n for $n \geq 3$. Thus, we focus on solving Problem P_n and what we can expect for the order of convergence of Equation (4.2).

For each n, Problem P_n is a variational inequality on a finite-dimensional space. This linear complementarity problem can be solved by several methods, including the penalty methods [25, Section I.7] combined with the gradient method or Newton's method, or iterative methods like iterative relaxation methods [33, Chapter 7].

We implemented finite-element methods to solve numerically the obstacle problem, by reusing our MATLAB code for the case of equality combined with the penalty method and the relaxation method with projection, however the computational time to solve the obstacle problem for one mesh size was way too long, given that through these methods we have to iterate on the case of equality and that case was already taking a long time, as per our numerical results of Section 3.5. We are going to give an exposition of the penalty method and the relaxation method with projection for the case of the finite-difference method rather than for the finite-element method.

4.5 Rate of convergence

Let u_n be the unique solution to Problem P_n. To find the order at which the solution, u_n, converges to u (the unique solution of Problem 4.1), we will again assume that $u \in H^2(\mathscr{O})$ and use Falk's general error estimate (Theorem 4.2.3) together with approximation estimates with respect to the basis $\mathcal{B}$ that we already used in the case of equality.

First, let us restate his general error estimate in our context. We set $W = L^2(\mathscr{O},\mathfrak{w})$ in Theorem 4.2.3, and notice that for weighted Sobolev spaces it also holds that $L^2(\mathscr{O},\mathfrak{w})$ is dense in $H^{-1}(\mathscr{O},\mathfrak{w}) = \left(H_0^1(\mathscr{O}\cup\Gamma_0,\mathfrak{w})\right)^*$ and the injection of $L^2(\mathscr{O},\mathfrak{w})$

into $H^{-1}(\mathscr{O}, \mathfrak{w})$ is continuous.

Theorem 4.5.1 (Falk's error estimate for a solution to a variational inequality)**.** *Let u and u_n be the solutions to problems* (4.1) *and* (P_n)*, respectively. Denote by $A : H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w}) \longrightarrow H^{-1}(\mathscr{O}, \mathfrak{w})$ the continuous linear operator defined for $\tilde{v} \in H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$, such that*

$$\langle A\tilde{v}, v\rangle_{H^{-1}(\mathscr{O},\mathfrak{w}),H_0^1(\mathscr{O}\cup\Gamma_0,\mathfrak{w})} = \mathfrak{a}(\tilde{v}, v) \quad \textit{for all} \quad v \in H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w}).$$

Then,

$$\begin{aligned} \|u - u_n\|^2_{H^1(\mathscr{O},\mathfrak{w})} \quad \leq \quad & \frac{C^2}{\alpha^2}\|u - v_n\|^2_{H^1(\mathscr{O},\mathfrak{w})} \\ & + \frac{2}{\alpha}\|f - Au\|_{L^2(\mathscr{O},\mathfrak{w})} \left(\|u - v_n\|_{L^2(\mathscr{O},\mathfrak{w})} + \|u_n - v\|_{L^2(\mathscr{O},\mathfrak{w})}\right), \end{aligned}$$

for all $v \in \mathbb{K}$ and all $v_n \in \mathbb{K}_n$.

Lemma 4.5.2 (Rate of convergence for the obstacle problem)**.** *For each n let u_n be the solution of Problem P_n. If $u, \psi \in H^2(\mathscr{O})$ and $\beta > 1$, then the rate of convergence at which the u_n converge to u is of order 1. That is,*

$$||u - u_n||_{H^1(\mathscr{O},\mathfrak{w})} = O\left(\frac{1}{n}\right) \tag{4.3}$$

Proof. It is sufficient to estimate $\|u - v_n\|_{H^1(\mathscr{O},\mathfrak{w})}$ and $\|u - v_n\|_{L^2(\mathscr{O},\mathfrak{w})}$ for a particular v_n, and $\|u_n - v\|_{L^2(\mathscr{O},\mathfrak{w})}$ for a particular v.

i) By the proof of Lemma 3.4.2, we already have an estimate for $\|u - v_n\|_{H^1(\mathscr{O},\mathfrak{w})}$ when v_n is the linear combination of tensor product B-splines with coefficients given by the values of u on the nodes. Given that $u \in H^2(\mathscr{O})$, we have from Equation (3.43) that there exists a constant $\mathcal{C}(u)$ that depends on u, the parameters β and μ, and the diameter of the domain $\mathscr{O}$, such that,

$$\|u - v_n\|_{H^1(\mathscr{O},\mathfrak{w})} \leq \mathcal{C}(u)\frac{1}{n}.$$

We notice that v_n clearly belongs to $\mathbb{K}_n$.

ii) $\|u - v_n\|_{L^2(\mathscr{O},\mathfrak{w})}$: We will use the same estimate as employed in the proof of Lemma 3.4.2. Since $\beta > 1$, by [35, Theorem 12.7] there exists a constant $\mathcal{C}$, that depends on

β, μ and the diameter of $\mathscr{O}$, such that,

$$\|u - v_n\|_{L^2(\mathscr{O},\mathfrak{w})} \leq \mathcal{C}\|u - v_n\|_{L^2(\mathscr{O})} \leq \mathcal{C}\frac{1}{n^2}\left(\|D_{xx}u\|_{L^2(\mathscr{O})} + \|D_{yy}u\|_{L^2(\mathscr{O})}\right).$$

Thus, since we are assuming that $u \in H^2(\mathscr{O})$, there exists a constant $\mathcal{C}(u)$, that depends on u, β, μ and the diameter of $\mathscr{O}$, such that,

$$\|u - v_n\|_{L^2(\mathscr{O},\mathfrak{w})} \leq \mathcal{C}(u)\frac{1}{n^2}$$

iii) $\|u_n - v\|_{L^2(\mathscr{O},\mathfrak{w})}$: For this estimate we will use the assumption that $\psi \in H^2(\mathscr{O})$. As in Falk [15], we will proceed by considering one particular v. Let $v := \sup\{u_n, \psi\}$. Since $u_n \in H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$ and $\psi \in H^1(\mathscr{O}, \mathfrak{w})$ with $\psi \leq 0$ on Γ_1, then $v \in \mathbb{K}$. For each n, define $\psi_n \in V_n$ to be the linear combination of tensor product B-splines with coefficients the values of ψ on the nodes Σ_n. That is, $\psi_n \in \mathbb{K}_n$. First we notice that $u_n \geq \psi_n$ in $\mathscr{O}$. Indeed, this follows since $u_n \geq \psi = \psi_n$ on Σ_n, and all of our tensor product B-splines are non-negative and equal to 1 on their defining node. Let us estimate $\|u_n - v\|_{L^2(\mathscr{O},\mathfrak{w})}$: If $u_n \geq \psi$, then $|u_n - v| = 0 \leq |\psi_n - \psi|$. If $u_n < \psi$, then

$$\psi_n - \psi = \psi_n - v \leq u_n - v < 0,$$

which implies that $|u_n - v| \leq |\psi_n - \psi|$. In either case, $|u_n - v| \leq |\psi_n - \psi|$, and then by the same argument applied to ii), but for ψ (not for u), there exists a constant $\mathcal{C}$ that depends on β, μ and the diameter of $\mathscr{O}$, such that,

$$\|u_n - v\|_{L^2(\mathscr{O},\mathfrak{w})} \leq \|\psi_n - \psi\|_{L^2(\mathscr{O},\mathfrak{w})} \leq \mathcal{C}\|\psi_n - \psi\|_{L^2(\mathscr{O})} \leq \mathcal{C}\frac{1}{n^2}\left(\|D_{xx}\psi\|_{L^2(\mathscr{O})} + \|D_{yy}\psi\|_{L^2(\mathscr{O})}\right)$$

Therefore, since $\psi \in H^2(\mathscr{O})$, there exists a constant $\mathcal{C}(\psi)$ that depends on β, μ, ψ and the diameter of $\mathscr{O}$ such that,

$$\|u_n - v\|_{L^2(\mathscr{O},\mathfrak{w})} \leq \mathcal{C}(\psi)\frac{1}{n^2}$$

Combining i), ii), and iii), together with Falk's general error estimate, Theorem 4.5.1, then the main assertion follows. □

Remark 4.5.3. *As in the case of equality, if we had that $u, \psi \in W^{2,\infty}(\mathscr{O})$, or $u, \psi \in H^2(\mathscr{O}, \mathfrak{w})$, then the same rate of convergence can be obtained for all $\beta > 0$.*

Chapter 5

Review of existence and uniqueness results for elliptic Heston boundary value and obstacle problems in weighted Hölder spaces

We summarize higher regularity results for solutions to both the boundary value problem and the obstacle problem for the elliptic Heston operator. These results will allow us in the next chapter to use finite-differences to construct approximate numerical solutions.

5.1 Notation

We review the notation of the Daskalopoulos-Hamilton-Koch families of Hölder Banach spaces and their correspondent norms [13].

Let $\alpha \in (0,1)$ and let u be a function defined on an open subset $\mathscr{O} \subset \mathbb{H}$. Let us recall that the standard Hölder semi-norm of u is defined by

$$[u]_{C^{\alpha}(\overline{\mathscr{O}})} = \sup_{\substack{(x_1,y_1),(x_2,y_2)\in\mathscr{O} \\ (x_1,y_1)\neq(x_2,y_2)}} \frac{|u(x_2,y_2)-u(x_1,y_1)|^{\alpha}}{|(x_2,y_2)-(x_1,y_1)|^{\alpha}},$$

and that the Daskalopoulos-Hamilton-Koch Hölder semi-norm of u is

$$[u]_{C_s^{\alpha}(\overline{\mathscr{O}})} = \sup_{\substack{(x_1,y_1),(x_2,y_2)\in\mathscr{O} \\ (x_1,y_1)\neq(x_2,y_2)}} \frac{|u(x_2,y_2)-u(x_1,y_1)|^{\alpha}}{s((x_1,y_1),(x_2,y_2))^{\alpha}}, \tag{5.1}$$

where the usual Euclidean distance between points,

$$|(x_2,y_2)-(x_1,y_1)| = (|x_2-x_1|^2+|y_2-y_1|^2)^{1/2},$$

for $(x_1,y_1),(x_2,y_2) \in \overline{\mathbb{H}}$, is replaced by the distance function $s((x_1,y_1),(x_2,y_2))$,

$$s((x_1,y_1),(x_2,y_2)) = \frac{|(x_2,y_2)-(x_1,y_1)|}{\sqrt{y_1+y_2+|(x_2,y_2)-(x_1,y_1)|}}.$$

Notice that $s((x_1,y_1),(x_2,y_2)) \leq |(x_2,y_2)-(x_1,y_1)|^{1/2}$ and if $\mathscr{O}$ is bounded, then there exists a constant $K(\mathscr{O})$, that depends on the height and diameter of $\mathscr{O}$, such that $|(x_2,y_2)-(x_1,y_1)| \leq K(\mathscr{O})s((x_1,y_1),(x_2,y_2))$.

Daskalopoulos and Hamilton provide the following definition,

Definition 5.1.1 (C_s^α norm and Banach space)**.** [13, p. 901] *Given $\alpha \in (0,1)$ and an open set $\mathscr{O} \subset \mathbb{H}$, we say that $u \in C_s^\alpha(\overline{\mathscr{O}})$ if $u \in C(\overline{\mathscr{O}})$ and*

$$\|u\|_{C_s(\overline{\mathscr{O}})} < \infty,$$

where

$$\|u\|_{C_s(\overline{\mathscr{O}})} := [u]_{C_s(\overline{\mathscr{O}})} + \|u\|_{C(\overline{\mathscr{O}})}. \tag{5.2}$$

We say that $u \in C_s^\alpha(\mathscr{O} \cup \Gamma_0)$ if $u \in C_s^\alpha(\overline{V})$ for all precompact open subsets $V \Subset \mathscr{O} \cup \Gamma_0$.

It is known that $C_s^\alpha(\overline{\mathscr{O}})$ is a Banach space [13, Section I.1] with respect to the norm defined in Equation (5.2). Let us now recall the definition of higher-order $C_s^{k,\alpha}$ Hölder Banach spaces and their correspondent norms.

Definition 5.1.2 ($C_s^{k,\alpha}$ norm and Banach space)**.** [13, p. 902] *Given an integer $k \geq 0$, $\alpha \in (0,1)$, and an open subset $\mathscr{O} \subset \mathbb{H}$, we say that $u \in C_s^{k,\alpha}(\overline{\mathscr{O}})$ if $u \in C^k(\overline{\mathscr{O}})$ and*

$$\|u\|_{C_s^{k,\alpha}(\overline{\mathscr{O}})} < \infty,$$

where

$$\|u\|_{C_s^{k,\alpha}(\overline{\mathscr{O}})} := \sum_{|\beta| \leq k} \|D^\beta u\|_{C_s^\alpha(\overline{\mathscr{O}})},$$

with $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{N}^d$, $|\beta| := \beta_1 + \cdots + \beta_d$, and

$$D^\beta u := \frac{\partial^{|\beta|} u}{\partial_{x_1}^{\beta_1} \cdots \partial_{x_d}^{\beta_d}}.$$

If $k = 0$, we denote $C_s^{k,\alpha}(\overline{\mathscr{O}}) = C_s^{0,\alpha}(\overline{\mathscr{O}})$ by $C_s^\alpha(\overline{\mathscr{O}})$.

We are only interested in the case $d = 2$. Finally, we recall the definition of the higher-order $C_s^{k,2+\alpha}$ Hölder Banach spaces and their correspondent norms.

Definition 5.1.3 ($C_s^{k,2+\alpha}$ norm and Banach space)**.** [13, pp. 901-902] *Given an integer $k \geq 0$, $\alpha \in (0,1)$, and an open subset $\mathscr{O} \subset \mathbb{H}$, we say that $u \in C_s^{k,2+\alpha}(\overline{\mathscr{O}})$ if $u \in C_s^{k+1,\alpha}(\overline{\mathscr{O}})$, the derivatives, $D^\beta u$, $\beta \in \mathbb{N}^d$ with $|\beta| = k+2$ are continuous on U, and the functions $yD^\beta u$, $\beta \in \mathbb{N}^d$ with $|\beta| = k+2$, extend continuously up to the boundary, $\partial\mathscr{O}$, and those extensions belong to $C_s^\alpha(\overline{\mathscr{O}})$. We define,*

$$\|u\|_{C_s^{k,2+\alpha}(\overline{\mathscr{O}})} := \|u\|_{C_s^{k+1,\alpha}(\overline{\mathscr{O}})} + \sum_{|\beta|=k+2} \|yD^\beta u\|_{C_s^\alpha(\overline{\mathscr{O}})}.$$

We say that $u \in C_s^{k,2+\alpha}(\mathscr{O} \cup \Gamma_0)$ if $u \in C_s^{k,2+\alpha}(\overline{V})$ for all precompact open subsets $V \Subset \mathscr{O} \cup \Gamma_0$. When $k = 0$, we denote $C_s^{k,2+\alpha}(\overline{\mathscr{O}}) = C_s^{0,2+\alpha}(\overline{\mathscr{O}})$ by $C_s^{2+\alpha}(\overline{\mathscr{O}})$.

The following Lemma will be useful when we solve the boundary value problem and obstacle problems by the finite-difference method.

Lemma 5.1.4 (Boundary properties of functions in weighted Holder spaces)**.** [18] *If $u \in C_s^{2+\alpha}(\overline{\mathbb{H}})$, then for all $(x_0, y_0) \in \partial\mathbb{H}$,*

$$\lim_{\substack{(x,y)\to(x_0,y_0)\\(x,y)\in\mathbb{H}}} yD^2u(x,y) = 0.$$

Remark 5.1.5. *A consequence of Lemma 5.1.4 is that for $u \in C_s^{2+\alpha}(\overline{\mathscr{O}})$ we have that*

$$\begin{aligned} y|u_{xx}(x,y)| &\leq \|u\|_{C_s^{2+\alpha}(\overline{\mathscr{O}})} y^{\alpha/2} \\ y|u_{xy}(x,y)| &\leq \|u\|_{C_s^{2+\alpha}(\overline{\mathscr{O}})} y^{\alpha/2} \\ y|u_{yy}(x,y)| &\leq \|u\|_{C_s^{2+\alpha}(\overline{\mathscr{O}})} y^{\alpha/2} \end{aligned} \tag{5.3}$$

for all $(x,y) \in \Gamma_0$.

As reviewed in Chapter 2, by generalizing the methods of Koch [29] and focusing only on the Heston operator in two dimensions, Daskalopoulos, Feehan, and Pop successfully proved existence, uniqueness, and regularity of solutions to the elliptic boundary value problem and obstacle problem by solving the associated variational equation and inequality for solutions, u, in weighted Sobolev spaces. They also achieved higher regularity results in their joint work. Feehan proved uniqueness [17]; Feehan and Pop proved that u is continuous up to the boundary [19] using a Moser iteration technique, proved Schauder regularity when $f \in C_0^\infty(\mathscr{O} \cup \Gamma_0) \cap C_b(\mathscr{O})$ using

a variational method [20], and provided the expected Schauder regularity result of $u \in C_s^{2+\alpha}(\mathscr{O} \cup \Gamma_0) \cap C_b(\mathscr{O} \cup \Gamma_1)$ for a solution to the boundary value problem, using elliptic a priori interior Schauder estimates and regularity results [21]; also, Daskalopoulos and Feehan proved that $u \in C_s^{1,1}(\mathscr{O} \cup \Gamma_0) \cap C_b(\mathscr{O} \cup \Gamma_1)$ for a solution to the obstacle problem by adapting arguments of Caffarelli [8] in [11].

However, Feehan in [16] used Perron methods to provide a more direct approach to prove all of the results above, except the continuity of the solution at the corner points, although such continuity properties are proved by Pop and Feehan in [19] for the case of the elliptic Heston operator. To state Feehan results [16] we need the following definition first:

Definition 5.1.6 (Regular boundary point)**.** *If $u \in C_s^{2+\alpha}(\mathscr{O} \cup \Gamma_0)$ is a bounded solution to the boundary value problem for the elliptic Heston operator* (2.1)*, we say that a point $(x_0, y_0) \in \Gamma_1$ is* regular with respect to f, and g *if the point (x_0, y_0) admits a local barrier in the sense of [24, p. 105]; if (x_0, y_0) is a regular point, then [24, Lemma 6.12] implies that*

$$\lim_{\substack{x \to x_0 \\ x \in \mathscr{O}}} u(x) = g(x_0).$$

A point $x_0 \in \Gamma_1$ will be regular, for instance, if $\mathscr{O}$ obeys an exterior condition at x_0 [24, p. 106], or an exterior cone condition at x_0 [24, Problem 6.3].

5.2 Elliptic Heston boundary value problem

The Perron methods developed by Feehan in [16] are analogues of their classical counterpart in [24, Chapter 2 and 6] for the existence of smooth solutions to a Dirichlet problem for a linear, second-order, strictly elliptic operator. Feehan proved [16] the following:

Theorem 5.2.1 (Existence of a smooth solution to the boundary value problem for the Heston operator)**.** [16, Theorem 1.4] *Let $\mathscr{O} \subset \mathbb{H}$ be a bounded domain and $\alpha \in (0, 1)$. Let A be the elliptic Heston operator as in Equation* (2.2)*. If $f \in C_s^{\alpha}(\mathscr{O} \cup \Gamma_0) \cap C_b(\mathscr{O})$ and $g \in C_b(\Gamma_1)$, and each point of Γ_1 is regular with respect to A, f and g in the sense*

of Definition 5.1.6, then there is a unique solution,

$$u \in C_s^{2+\alpha}(\mathscr{O} \cup \Gamma_0) \cap C_b(\mathscr{O} \cup \Gamma_1),$$

to the boundary value problem for the elliptic Heston operator (2.1).

When $f \in C_s^{\alpha}(\mathscr{O} \cup \Gamma_0 \cup \Gamma_1)$, and $g \in C_s^{2+\alpha}(\mathscr{O} \cup \Gamma_0 \cup \Gamma_1)$, and Γ_1 is of class $C^{2,\alpha}$, then the standard regularity theory for boundary value problems for strictly elliptic operators [24, Lemma 6.18] implies that the solution to Problem 2.1 belongs to $C_s^{2+\alpha}(\mathscr{O} \cup \Gamma_0 \cup \Gamma_1)$.

Remark 5.2.2. *Feehan [16] remarks that given a solution $u \in C_s^{2+\alpha}(\mathscr{O} \cup \Gamma_0)$ to* (2.1)*, continuity up to Γ_1 is assured by the existence of a local barrier at each point of Γ_1 [24, pp. 104-106], however, because A is degenerate when $y = 0$, it is unclear how to construct a local barrier at the corner points, $\overline{\Gamma_0} \cap \overline{\Gamma_1}$.*

5.3 Elliptic Heston obstacle problem

As for the equality case, in this section we summarize higher regularity results, proved by Feehan [16], for solutions to the variational inequality for the elliptic Heston operator. He proved the following analogue of [23, Theorems 1.3.2, 1.3.4, 1.4.1, and 1.4.3].

Theorem 5.3.1 (Existence of a smooth solution to the obstacle problem)**.** [16, Theorem 1.7]. *Let $\mathscr{O} \subset \mathbb{H}$ be a bounded domain, $2 < p < \infty$, and $\alpha \in (0,1)$. Assume the hypothesis for f and g in Theorem 5.2.1. If $\psi \in C^2(\mathscr{O} \cup \Gamma_0) \cap C(\mathscr{O} \cup \Gamma_1)$ obeys the compatibility condition, $\psi \geq g$ on Γ_1, and each point of Γ_1 is regular with respect to A, f, and g in the sense of Definition 5.1.6, then there is a unique solution,*

$$u\ C_s^{2+\alpha}(\Omega \cup \Gamma_0(\Omega)) \cap W^{2,p}_{\mathrm{loc}}(\mathscr{O}) \cap C_s^{1,\alpha}(\mathscr{O} \cup \Gamma_0) \cap C_b(\mathscr{O} \cup \Gamma_1),$$

to the obstacle problem (2.20)*, where $\Omega = \{(x,y) \in \mathscr{O} : u(x,y) > \psi(x,y)\}$.*

When $f \in C_s^{\alpha}(\mathscr{O} \cup \Gamma_0 \cup \Gamma_1)$, and $g \in C_s^{2+\alpha}(\mathscr{O} \cup \Gamma_0 \cup \Gamma_1)$, and $\psi \in C^2(\mathscr{O} \cup \Gamma_0 \cup \Gamma_1)$, and Γ_1 is of class $C^{2,\alpha}$, then again standard regularity theory for obstacle problems for strictly elliptic operators [23, Theorem 1.3.2 or 1.3.5] implies that the solution u belongs to $W^{2,p}_{\mathrm{loc}}(\mathscr{O} \cup \Gamma_1)$. Feehan [16] points out that we actually have $u \in W^{2,\infty}_{\mathrm{loc}}(\mathscr{O} \cup \Gamma_1)$ by [23, Theorems 1.4.1 and 1.4.3].

Remark 5.3.2 (Optimal regularity of a solution to the obstacle problem up to Γ_0)**.** *Optimal regularity up to the degenerate boundary,* Γ_0*, that is,* $u \in C_1^{1,1}(\mathscr{O} \cup \Gamma_0)$ *in the sense of [11, Definition 2.2], for a solution* u *to* (2.20) *is proved by Daskalopoulos and Feehan in [11].*

Theorem 5.2.1 and Theorem (5.3.1), plus a reasonable assumption about regularity of the solution to the boundary value problem (2.1) and the obstacle problem (2.20) on the corner points, will allow us to talk about approximations of its first-order and second-order derivatives when we use finite differences to solve both the equation and the inequality approximately. This will also allows us, in the case of the boundary value problem, to prove convergence of the finite-difference solutions and to provide a rate of convergence.

Chapter 6

Finite-difference method for the boundary value problem for the elliptic Heston operator

For suitable functions f and g, and bounded domain $\mathscr{O}$, Theorem 5.2.1 provides regularity for the solution to the boundary value problem (2.1), u, up to $\overline{\mathscr{O}}$ except at the"corner points", $\overline{\Gamma_0} \cap \overline{\Gamma_1}$. While this is still an active research item, towards asserting that $u \in C_s^{2+\alpha}(\overline{\mathscr{O}})$, that is including the regularity everywhere including the corner points, we consider it as a hypothesis for the analysis of the finite-difference method for this problem exposed in this chapter. That is, throughout this chapter, we assume,

$$u \in C_s^{2+\alpha}(\overline{\mathscr{O}}).$$

6.1 Introduction to finite-difference methods for the elliptic Heston PDE

Recall the boundary value Problem 2.5 with homogeneous boundary condition,

$$\begin{cases} Au = f \quad \text{a.e. in } \mathscr{O}, \\ u = 0 \quad \text{on } \Gamma_1, \end{cases} \tag{6.1}$$

where the operator A is given by,

$$Au = -\frac{y}{2}\left(u_{xx} + 2\rho\sigma u_{xy} + \sigma^2 u_{yy}\right) - (r - q - y/2)u_x - \kappa(\theta - y)u_y + ru. \tag{6.2}$$

Consider again a mesh like the one considered in the case of the finite-element method in Section 3.2, but on a bounded domain, and let us further assume it is uniform in each direction. Thus, given $L > 0$ and $V > 0$ we choose

$$\mathcal{P}_x : \{X_0 \leq x_1 < x_2 < x_3 < \cdots < x_M \leq X_1\},$$

$$\mathcal{P}_y : \{Y_0 \le y_1 < y_2 < y_3 < \cdots < y_N \le Y_1\},$$

where $X_0 = -L$, $X_1 = L$, $Y_0 = 0$, $Y_1 = V$, and $x_{i+1} = x_i + h_x$, $y_{j+1} = y_j + h_y$ for $h_x = (X_1 - X_0)/M$ and $h_y = (Y_1 - Y_0)/N$. We will approximate the derivatives of u at each point (x_i, y_j) in $\mathscr{O}$ by using central differences in both directions, except along the line $\{y = 0\}$, where we will use central differences for the x-derivatives and a forward difference for the y-derivative. Our finite-difference approximations are then

$$u_x(x_i, y_j) = \widehat{u_x^{i,j}} + O(h_x^2) := \frac{1}{2h_x}\left(u_{i+1,j} - u_{i-1,j}\right) + O(h_x^2), \tag{6.3}$$

$$u_{xx}(x_i, y_j) = \widehat{u_{xx}^{i,j}} + O(h_x^2) := \frac{1}{h_x^2}\left(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}\right) + O(h_x^2), \tag{6.4}$$

$$\begin{aligned} u_y(x_i, y_j) &= \widehat{u_y^{i,j}} + \begin{cases} O(h_y^2), & \text{if } j > 1, \\ O(h_y), & \text{if } j = 1, \end{cases} \\ &:= \begin{cases} \dfrac{1}{2h_y}\left(u_{i,j+1} - u_{i,j-1}\right) + O(h_y^2) & \text{if } j > 1, \\ \dfrac{1}{h_y}\left(u_{i,j+1} - u_{i,j}\right) + O(h_y) & \text{if } j = 1, \end{cases} \end{aligned} \tag{6.5}$$

$$u_{yy}(x_i, y_j) = \widehat{u_{yy}^{i,j}} + O(h_y^2) := \frac{1}{h_y^2}\left(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}\right) + O(h_y^2), \tag{6.6}$$

$$\begin{aligned} u_{xy}(x_i, y_j) &= \widehat{u_{xy}^{i,j}} + O(h_x h_y) \\ &:= \frac{1}{4h_x h_y}\left(u_{i+1,j+1} - u_{i+1,j-1} - u_{i-1,j+1} + u_{i-1,j-1}\right) + O(h_x^2) + O(h_y^2), \end{aligned} \tag{6.7}$$

where $u_{i,j} = u(x_i, y_j)$. Notice that if we assume that the second-order derivatives exist along $\{y = 0\}$, that is for $j = 1$, then our operator A does not involve any second-order derivatives when evaluating Au along any point on that line because of Lemma 5.1.4. We will not impose a boundary condition along $\{y = 0\}$ but allow the function values $u_{i,1} = u(x_i, y_1)$ to be unknowns, by analogy with the boundary value problem for the Cox-Ingersoll-Ross ordinary differential equation (see Appendix A).

The discretized equation evaluated at (x_i, y_j), for $1 < i < M$ and $1 \le j < N$, is

$$\sum_{k=i-1}^{k=i+1} \sum_{l=j-1}^{l=j+1} d_{k,l} u_{k,l} = f_{i,j}, \tag{6.8}$$

and its normalized version is then,

$$\sum_{k=i-1}^{k=i+1} \sum_{l=j-1}^{l=j+1} m_{k,l} u_{k,l} = 2h_x^2 h_y^2 f_{i,j}, \tag{6.9}$$

where $f_{i,j} = f(x_i, y_j)$, and all of the $m_{k,l}$ are guaranteed to be zero except $m_{i-1,j-1}, ..., m_{i+1,j+1}$. For $j > 1$, the $m_{k,l}$ are

$$\begin{aligned}
m_{i-1,j-1} &= -\frac{1}{2}\rho\sigma y_j h_x h_y, \\
m_{i-1,j} &= (r - q - y_j/2)h_x h_y^2 - y_j h_y^2, \\
m_{i-1,j+1} &= \frac{1}{2}\rho\sigma y_j h_x h_y, \\
m_{i,j-1} &= \kappa(\theta - y_j)h_x^2 h_y - \sigma^2 y_j h_x^2, \\
m_{i,j} &= 2y_j h_y^2 + 2\sigma^2 y_j h_x^2 + 2r h_x^2 h_y^2, \\
m_{i,j+1} &= -\kappa(\theta - y_j)h_x^2 h_y - \sigma^2 y_j h_x^2, \\
m_{i+1,j-1} &= \frac{1}{2}\rho\sigma y_j h_x h_y, \\
m_{i+1,j} &= -(r - q - y_j/2)h_x h_y^2 - y_j h_y^2, \\
m_{i+1,j+1} &= -\frac{1}{2}\rho\sigma y_j h_x h_y.
\end{aligned} \tag{6.10}$$

For $j = 1$, that is when $y = 0$, all of the formulas above hold except for the formulas for $m_{i,j-1}, m_{i,j}$, and $m_{i,j+1}$. The set of formulas for $j = 1$ is

$$\begin{aligned}
m_{i-1,0} &= -\frac{1}{2}\rho\sigma y_1 h_x h_y, \\
m_{i-1,1} &= (r - q - y_1/2)h_x h_y^2 - y_1 h_y^2, \\
m_{i-1,2} &= \frac{1}{2}\rho\sigma y_1 h_x h_y, \\
m_{i,0} &= 0, \\
m_{i,1} &= 2\kappa\theta h_x^2 h_y + 2h_x^2 h_y^2 r, \\
m_{i,2} &= -2\kappa\theta h_x^2 h_y, \\
m_{i+1,0} &= \frac{1}{2}\rho\sigma y_1 h_x h_y, \\
m_{i+1,1} &= -(r - q - y_1/2)h_x h_y^2 - y_1 h_y^2, \\
m_{i+1,2} &= -\frac{1}{2}\rho\sigma y_1 h_x h_y.
\end{aligned} \tag{6.11}$$

The nodes will be numbered from bottom to top, and then from left to right, as the example in Figure 6.1 indicates.

This enumeration defines a mapping $(i, j) \to I(i, j)$. We can write Equation (6.9) as,

$$\sum_s \widehat{m}_{t,s} u_s = 2h_x^2 h_y^2 f_t, \tag{6.12}$$

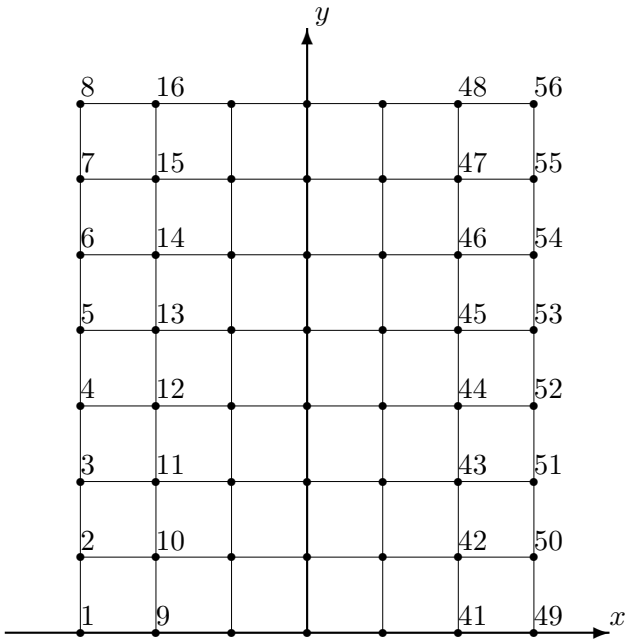


Figure 6.1: Example of a mesh for $[-3,3] \times [0,7]$ and enumeration of its nodes

for every $t = I(i,j)$ with (x_i, y_j) not on Γ_1. Notice $\widehat{m}_{t,s} = 0$ unless $s = I(k,l)$ corresponds to (k,l) being a neighbor of (i,j). That is, if $|k-i| \leq 1$ and $|l-j| \leq 1$, like in Figure 6.2.

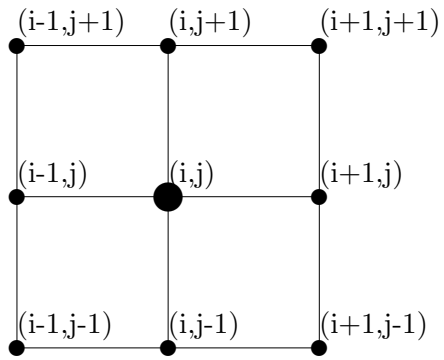


Figure 6.2: Example of a node and its neighbors

The linear system defined by Equation (6.12) on the unknowns $\{u_s\}_{s\geq 0}$ can be written as

$$\widehat{M}\vec{u} = 2h_x^2 h_y^2 \vec{f}, \tag{6.13}$$

where $\widehat{M} = \{\widehat{m}_{t,s}\}_{t,s}$ is a sparse matrix, and $\vec{u}, \vec{f} \in \mathbb{R}^{(M-2)(N-1)}$.

6.2 A discrete maximum principle

The finite-difference scheme proposed in the previous section satisfies a discrete maximum principle when the parameters appearing in the operator A, as in Equation (6.2), satisfy certain conditions. From Equation (6.8), define $\mathcal{L}$ as,

$$\mathcal{L}u_{i,j} := \sum_{k=i-1}^{k=i+1} \sum_{l=j-1}^{l=j+1} d_{k,l} u_{k,l}, \tag{6.14}$$

where $d_{k,l} = m_{k,l}/2h_x^2 h_y^2$ and $m_{k,l}$ are the coefficients in the previous section, and u is any mesh function.

As in Ciarlet [9], we define the following analogous property for finite-difference operators like the operator $\mathcal{L}$:

Definition 6.2.1 (Discrete maximum principle for the Heston operator)**.** *A finite-difference operator $\mathcal{L}$ satisfies the discrete maximum principle for the Heston operator if and only if whenever $\mathcal{L}u_{i,j} \leq 0$ for all mesh points $(x_i, y_j) \in \mathscr{O} \cup \Gamma_0$, then*

$$\max\{u_{i,j} | (x_i, y_j) \in \mathscr{O} \cup \Gamma_0\} \leq \max\{0, \max\{u_{i,j} | (x_i, y_j) \in \Gamma_1\}\} \tag{6.15}$$

Notice that the definition of this property states a non-negative maximum is attained on Γ_1 rather than on $\partial\mathscr{O}$. We believe the operator $\mathcal{L}$, or a modification of it, satisfies this discrete maximum principle for all possible values of all parameters appearing in the operator A (as in Equation (6.2)), however we are able to prove this only when these parameters obey certain conditions, namely,

Lemma 6.2.2 (A first discrete maximum principle)**.** *If $\rho = 0$, $r = q > 0$, $\kappa\theta \leq \sigma^2$ and the mesh size is sufficiently small, then $\mathcal{L}$ satisfies the discrete maximum principle.*

Proof. First, let us write formulas for the operator $\mathcal{L}$ evaluated on u at all internal points for the values specified for the different parameters. For $j > 1$,

$$\begin{aligned}
\mathcal{L}u_{i,j} \quad = \quad & \left(-\frac{y_j}{2h_x^2} - \frac{y_j}{4h_x}\right) u_{i-1,j} + \left(\frac{\kappa(\theta - y_j)}{2h_y} - \frac{\sigma^2 y_j}{2h_y^2}\right) u_{i,j-1} \\
& + \left(\frac{y_j}{h_x^2} + \frac{\sigma^2 y_j}{h_y^2} + r\right) u_{i,j} + \left(-\frac{\kappa(\theta - y_j)}{2h_y} - \frac{\sigma^2 y_j}{2h_y^2}\right) u_{i,j+1} \\
& + \left(-\frac{y_j}{2h_x^2} + \frac{y_j}{4h_x}\right) u_{i+1,j},
\end{aligned}$$

and for $j = 1$,

$$\mathcal{L}u_{i,1} = \left(\frac{\kappa\theta}{h_y} + r\right) u_{i,j} + \left(-\frac{\kappa\theta}{h_y}\right) u_{i,j+1}.$$

Suppose that $\mathcal{L}u_{k,l} \leq 0$ for all k, l such that $(x_k, y_l) \in \mathscr{O} \cup \Gamma_0$, and let us assume by contradiction that at an "internal" point $(x_i, y_j) \in \mathscr{O} \cup \Gamma_0$ a positive maximum is attained. That is, assume that

$$M := u_{i,j} = \max\{u_{k,l}; (x_k, y_l) \in \mathscr{O} \cup \Gamma_0\} > \max\{0, \max\{u_{k,l}; (x_k, y_l) \in \Gamma_1\}\}.$$

In particular, $M > 0$, $\mathcal{L}u_{i,j} \leq 0$, and for every $(x_k, y_l) \in \Gamma_1$ we have that $u_{k,l} < M$. Let us consider first the case $j > 1$. Thus,

$$\begin{aligned}\left(\frac{y_j}{h_x^2} + \frac{\sigma^2 y_j}{h_y^2} + r\right) u_{i,j} \leq & \left(\frac{y_j}{2h_x^2} + \frac{y_j}{4h_x}\right) u_{i-1,j} + \left(-\frac{\kappa(\theta - y_j)}{2h_y} + \frac{\sigma^2 y_j}{2h_y^2}\right) u_{i,j-1} \\ & + \left(\frac{\kappa(\theta - y_j)}{2h_y} + \frac{\sigma^2 y_j}{2h_y^2}\right) u_{i,j+1} + \left(\frac{y_j}{2h_x^2} - \frac{y_j}{4h_x}\right) u_{i+1,j},\end{aligned}$$

and notice that all coefficients are non-negative given that $\kappa\theta \leq \sigma^2$ and the mesh size is sufficiently small ($h_x \leq 2$ and $h_y \leq \frac{\sigma^2}{\kappa}$ suffices). Therefore, by estimating from above all values of u appearing in the right hand side of the inequality,

$$\begin{aligned}\left(\frac{y_j}{h_x^2} + \frac{\sigma^2 y_j}{h_y^2} + r\right) M \leq & \left(\frac{y_j}{2h_x^2} + \frac{y_j}{4h_x}\right) M + \left(-\frac{\kappa(\theta - y_j)}{2h_y} + \frac{\sigma^2 y_j}{2h_y^2}\right) M \\ & + \left(\frac{\kappa(\theta - y_j)}{2h_y} + \frac{\sigma^2 y_j}{2h_y^2}\right) M + \left(\frac{y_j}{2h_x^2} - \frac{y_j}{4h_x}\right) M,\end{aligned}$$

implying that,

$$\left(\frac{y_j}{h_x^2} + \frac{\sigma^2 y_j}{h_y^2} + r\right) M \leq \left(\frac{y_j}{h_x^2} + \frac{\sigma^2 y_j}{h_y^2}\right) M,$$

which is a contradiction since $r > 0$ and $M > 0$. Now let's consider the case $j = 1$. From $\mathcal{L}u_{i,1} \leq 0$, it follows that

$$\left(\frac{\kappa\theta}{h_y} + r\right) M = \left(\frac{\kappa\theta}{h_y} + r\right) u_{i,1} \leq \left(\frac{\kappa\theta}{h_y}\right) u_{i,2} < \left(\frac{\kappa\theta}{h_y}\right) M,$$

which is a contradiction again for the same reason. Hence $\mathcal{L}$ satisfies the discrete maximum principle for the Heston operator. □

We can improve our first discrete maximum principle to include all values of ρ if we change the way we approximate the mixed partial derivative u_{xy} of Equation (6.8). If

we estimate it instead by

$$u_{xy}(x_i, y_j) \approx \frac{1}{2h_x h_y}(u_{i+1,j+1} - u_{i+1,j} - u_{i,j+1} + 2u_{i,j} - u_{i-1,j} - u_{i,j-1} + u_{i-1,j-1}),$$

and recalculate Equation (6.8), we obtain a new finite-difference operator $\widetilde{\mathcal{L}}$. The new operator's formula, for $j > 1$, is

$$\begin{aligned}
\widetilde{\mathcal{L}}u_{i,j} = &\left(-\frac{\rho\sigma}{2h_x h_y}y_j\right)u_{i-1,j-1} \\
&+\left(-y_j\left(\frac{1}{2h_x^2}+\frac{1}{4h_x}-\frac{\rho\sigma}{2h_x h_y}\right)+\frac{r-q}{2h_x}\right)u_{i-1,j} \\
&+\left(\frac{\kappa(\theta-y_j)}{2h_y}-\frac{\sigma^2}{2h_y^2}y_j+\frac{\rho\sigma}{2h_x h_y}y_j\right)u_{i,j-1} \\
&+\left(\frac{y_j}{h_x^2}+\frac{\sigma^2}{h_y^2}y_j+r-\frac{\rho\sigma}{h_x h_y}y_j\right)u_{i,j} \\
&+\left(-\frac{\kappa(\theta-y_j)}{2h_y}-\frac{\sigma^2}{2h_y^2}y_j+\frac{\rho\sigma}{2h_x h_y}y_j\right)u_{i,j+1} \\
&+\left(-y_j\left(\frac{1}{2h_x^2}-\frac{1}{4h_x}-\frac{\rho\sigma}{2h_x h_y}\right)-\frac{r-q}{2h_x}\right)u_{i+1,j} \\
&+\left(-\frac{\rho\sigma}{2h_x h_y}y_j\right)u_{i+1,j+1},
\end{aligned} \tag{6.16}$$

and for $j = 1$, it remains unchanged since there is no u_{xy} term in A along $\{y = 0\}$. For $\widetilde{\mathcal{L}}$ we have the following discrete maximum principle.

Theorem 6.2.3 (A second discrete maximum principle)**.** *If $r = q$, and $\kappa\theta \le \sigma^2(1-\rho^2)$, and the mesh is sufficiently fine such that*

$$\rho\sigma < \frac{h_y}{h_x} \le \frac{\sigma^2 - \kappa\theta}{\rho\sigma},$$

then $\widetilde{\mathcal{L}}$ satisfies the discrete maximum principle.

Proof. The proof goes very much along the same lines as Lemma 6.2.2. Since $r = q$, from the definition of $\widetilde{\mathcal{L}}$, Equation (6.16), we have

$$\begin{aligned}
\widetilde{\mathcal{L}}u_{i,j} \;=\; &\left(-\frac{\rho\sigma}{2h_x h_y}y_j\right)u_{i-1,j-1} - \frac{y_j}{2h_x}\left(\frac{1}{h_x}+\frac{1}{2}-\frac{\rho\sigma}{h_y}\right)u_{i-1,j} \\
&+\left(\frac{\kappa\theta}{2h_y}-\frac{y_j}{2h_y}\left(\kappa+\frac{\sigma^2}{h_y}-\frac{\rho\sigma}{h_x}\right)\right)u_{i,j-1} + \left(\frac{y_j}{h_x^2}+\frac{\sigma^2}{h_y^2}y_j+r-\frac{\rho\sigma}{h_x h_y}y_j\right)u_{i,j} \\
&+\left(-\frac{\kappa\theta}{2h_y}+\frac{y_j}{2h_y}\left(\kappa-\frac{\sigma^2}{h_y}+\frac{\rho\sigma}{h_x}\right)\right)u_{i,j+1} - \frac{y_j}{2h_x}\left(\frac{1}{h_x}-\frac{1}{2}-\frac{\rho\sigma}{h_y}\right)u_{i+1,j} \\
&+\left(-\frac{\rho\sigma}{2h_x h_y}y_j\right)u_{i+1,j+1}, \text{ for } j > 1,
\end{aligned}$$

and for $j = 1$, once again,

$$\mathcal{L}u_{i,1} = \left(\frac{\kappa\theta}{h_y} + r\right) u_{i,j} + \left(-\frac{\kappa\theta}{h_y}\right) u_{i,j+1}.$$

Let us focus on $j > 1$. First, we point out that from the definition of the Heston process, Equation (1.1), we can modify either ρ or σ so that $\rho\sigma \geq 0$ and so the coefficients of $u_{i-1,j-1}$ and $u_{i+1,j+1}$ are non-positive.

Denote by Δ the ratio between h_y and h_x, and assume it to be fixed. That is, $\Delta := h_y/h_x$. Then it follows that the coefficients of $u_{i-1,j}$ and $u_{i+1,j}$ are both non-positive if h_y is small enough, namely, if $h_y \leq 2(\Delta - \rho\sigma)$.

The coefficient of $u_{i,j-1}$ is non-positive if and only if $y_j \left(\kappa + \sigma^2/h_y - \rho\sigma/h_x\right) \geq \kappa\theta$, but since $j > 1$ and $\kappa\theta > 0$, that is equivalent to $\kappa h_y + \sigma^2 \geq \Delta\rho\sigma + \kappa\theta$, which happens for all $h_y > 0$ given our hypothesis $\sigma^2 \geq \Delta\rho\sigma + \kappa\theta$.

As for the coefficient of $u_{i,j+1}$, it is non-positive if and only if

$$y_j \left(\kappa - \sigma^2/h_y + \rho\sigma/h_x\right) \leq \kappa\theta.$$

For that to hold for any $j > 1$, it is sufficient to have $\kappa - \sigma^2/h_y + \rho\sigma/h_x \leq 0$, but this is equivalent to $\kappa h_y \leq \sigma^2 - \Delta\rho\sigma$, which is guaranteed when h_y is small enough, given that $\sigma^2 - \Delta\rho\sigma \geq \kappa\theta > 0$.

Therefore, for $j > 1$, all coefficients $u_{k,l}$ in $\widetilde{\mathcal{L}}u_{i,j}$ are non-positive, except $u_{i,j}$, which is positive itself. It is trivial to see that we have the same situation for $\widetilde{\mathcal{L}}u_{i,j}$ when $j = 1$. Hence, proceeding exactly as we did in the proof of Theorem 6.2.2, by contradiction, we conclude that such an operator $\widetilde{\mathcal{L}}$ satisfies the discrete maximum principle for the Heston operator. □

6.3 Convergence of the finite-difference approximation

Following the exposition in Krylov [30, Sections 6.6-6.7], we will consider from now on the two finite-difference operators, $\mathcal{L}$ and $\widetilde{\mathcal{L}}$, in Equations (6.14) and (6.16), respectively, which satisfy the discrete maximum principle for the Heston operator.

The first observation we make is that for an operator $\mathcal{L}$ that satisfies the discrete maximum principle, its defining linear system of Equation (6.13) has a unique solution,

hence it makes sense to ask a computational algorithm to find approximations to the unique solution. Indeed, since the linear system is represented by a square matrix, all we need to prove is the uniqueness of the solution. For that, suppose we have two solutions $\vec{u}, \vec{v} \in \mathbb{R}^{MN}$ to the linear system in Equation (6.13) such that $u_s = v_s = 0$ whenever s is an index corresponding to a point on Γ_1. From the definition of $\mathcal{L}$ in Equation (6.14), we have $(\mathcal{L}(u-v))_{i,j} = 0$ for all $(x_i, y_j) \in \mathscr{O} \cup \Gamma_0$, and by the discrete maximum principle,

$$\max\{u_{i,j} - v_{i,j} | (x_i, y_j) \in \mathscr{O} \cup \Gamma_0\} \leq \max\{0, \max\{u_{i,j} - v_{i,j} | (x_i, y_j) \in \Gamma_1\}\} = 0.$$

Similarly, $\max\{v_{i,j} - u_{i,j} | (x_i, y_j) \in \mathscr{O} \cup \Gamma_0\} \leq 0$, and thus $\vec{u} = \vec{v}$.

Also, operators like $\mathcal{L}$ and $\widetilde{\mathcal{L}}$ approximate the Heston operator A in the following sense,

Lemma 6.3.1 (The operators $\mathcal{L}$ and $\widetilde{\mathcal{L}}$ are consistent with the continuous operator A)**.** *Assume the finite-difference partitions are such that $h_y/h_x \leq \Lambda$, for some $\Lambda > 0$, for all $h_x > 0, h_y > 0$. For any $u \in C_s^{2+\alpha}(\overline{\mathscr{O}})$ and any $(x_i, y_j) \in \mathscr{O} \cup \Gamma_0$ we have*

$$|Au(x_i, y_j) - \mathcal{L}u_{i,j}| \leq K \max\{h_x, h_y\}^{\alpha/2} |u|_{C_s^{2+\alpha}(\overline{\mathscr{O}})}, \tag{6.17}$$

$$\left|Au(x_i, y_j) - \widetilde{\mathcal{L}}u_{i,j}\right| \leq K \max\{h_x, h_y\}^{\alpha/2} \|u\|_{C_s^{2+\alpha}(\overline{\mathscr{O}})}, \tag{6.18}$$

for some constant K that depends on α, the height of $\mathscr{O}$, and Λ.

Proof. We first derive the estimate (6.17). By definition of $\mathcal{L}$, we just need to check how well each finite-difference term approximates the corresponding derivative term in the operator A. These approximations are given in Equations (6.3)-(6.8).

All these estimates will follow from simple applications of the Mean Value Theorem and the Mean Value Theorem for Sums. Suppose $(x_i, y_j) \in \mathscr{O} \cup \Gamma_0$.

i) $\widehat{u_x^{i,j}} = \frac{1}{2h_x}(u_{i+1,j} - u_{i-1,j})$: By the Mean Value Theorem, there exist x_i^1 and $x_i^2 \in (x_i, x_{i+1})$ such that,

$$u_{i+1,j} = u_{i,j} + h_x u_x(x_i^1, y_j),$$

$$u_{i-1,j} = u_{i,j} - h_x u_x(x_i^2 y_j),$$

Thus, by the Mean Value Theorem for Sums, there exists $x_i^3 \in [x_{i-1}, x_{i+1}]$ such that

$$\widehat{u_x^{i,j}} = \frac{1}{2}(u_x(x_i^1, y_j) + u_x(x_i^2, y_j)) = u_x(x_i^3, y_j)$$

Therefore,

$$\begin{aligned}\left|\widehat{u_x^{i,j}} - u_x(x_i, y_j)\right| &= \frac{\left|u_x(x_i^3, y_j) - u_x(x_i, y_j)\right|}{s((x_i^3, y_j), (x_i, y_j))^\alpha} s((x_i^3, y_j), (x_i, y_j))^\alpha \\ &\leq \|u\|_{C_s^{2+\alpha}(\overline{\mathscr{O}})} h_x^{\alpha/2}.\end{aligned}$$

This completes the analysis for this term.

ii) $\widehat{u_{xx}^{i,j}} = \frac{1}{h_x^2}(u_{i+1,j} - 2u_{i,j} + u_{i-1,j})$: Here $j > 1$ since there is no u_{xx} term along $\{y = 0\}$. By the Mean Value Theorem, there exist $x_i^1 \in (x_i, x_{i+1})$ and $x_i^2 \in (x_{i-1}, x_i)$ such that

$$\begin{aligned}u_{i+1,j} &= u_{i,j} + h_x u_x(x_i, y_j) + \frac{h_x^2}{2} u_{xx}(x_i^1, y_j), \\ u_{i-1,j} &= u_{i,j} - h_x u_x(x_i, y_j) + \frac{h_x^2}{2} u_{xx}(x_i^2, y_j),\end{aligned}$$

and thus, by the Mean Value Theorem for Sums, there exists $x_i^3 \in [x_{i-1}, x_{i+1}]$ such that

$$\widehat{u_{xx}^{i,j}} = \frac{u_{xx}(x_i^1, y_j) + u_{xx}(x_i^2, y_j)}{2} = u_{xx}(x_i^3, y_j),$$

This implies that

$$\begin{aligned}y_j \left|\widehat{u_{xx}^{i,j}} - u_{xx}(x_i, y_j)\right| &\leq \frac{\left|y_j u_{xx}(x_i^3, y_j) - y_j u_{xx}(x_i, y_j)\right|}{s((x_i^3, y_j), (x_i, y_j))^\alpha} s((x_i^3, y_j), (x_i, y_j))^\alpha \\ &\leq \|u\|_{C_s^{2+\alpha}(\overline{\mathscr{O}})} h_x^{\alpha/2}.\end{aligned}$$

The analysis of this approximate derivative is finished.

iii) $\widehat{u_y^{i,j}} = \frac{1}{2h_y}(u_{i,j+1} - u_{i,j-1})$ for $j > 1$ and $\widehat{u_y^{i,j}} = \frac{1}{h_y}(u_{i,j+1} - u_{i,j})$ for $j = 1$: The analysis of the case $j > 1$ is similar to the analysis we gave for when $j > 1$ for $\widehat{u_x^{i,j}}$. For $j = 1$, there exists $y_j^1 \in (y_j, y_{j+1})$ such that

$$\widehat{u_y^{i,j}} = u_y(x_i, y_j^1),$$

which implies that,

$$\left|\widehat{u_y^{i,j}} - u_y(x_i, y_j)\right| \leq \|u\|_{C_s^{2+\alpha}(\overline{\mathscr{O}})} h_y^{\alpha/2}.$$

This concludes the analysis of this finite-difference.

iv) $\widehat{u^{i,j}_{yy}} = \frac{1}{h_y^2}(u_{i,j+1} - 2u_{i,j} + u_{i,j-1})$: Once again we only need to consider this case when $j > 1$. By the Mean Value Theorem, there exist $y_j^1 \in (y_j, y_{j+1})$ and $y_j^2 \in (y_{j-1}, y_j)$ such that

$$u_{i,j+1} = u_{i,j} + h_y u_y(x_i, y_j) + \frac{h_y^2}{2} u_{yy}(x_i, y_j^1),$$
$$u_{i,j-1} = u_{i,j} - h_y u_y(x_i, y_j) + \frac{h_y^2}{2} u_{yy}(x_i, y_j^2),$$

and thus, by the Mean Value Theorem for Sums, there exists $y_j^3 \in [y_{j-1}, y_{j+1}]$ such that

$$\widehat{u^{i,j}_{yy}} = \frac{u_{yy}(x_i, y_j^1) + u_{yy}(x_i, y_j^2)}{2} = u_{yy}(x_i, y_j^3),$$

This implies that

$$\begin{aligned} y_j \left| \widehat{u^{i,j}_{yy}} - u_{yy}(x_i, y_j) \right| &\leq |y_j u_{yy}(x_i, y_j^3) - y_j u_{yy}(x_i, y_j)| \\ &\leq |y_j^3 u_{yy}(x_i, y_j^3) - y_j u_{yy}(x_i, y_j)| + |(y_j - y_j^3) u_{yy}(x_i, y_j^3)| \\ &\leq |y_j^3 u_{yy}(x_i, y_j^3) - y_j u_{yy}(x_i, y_j)| + h_y |u_{yy}(x_i, y_j^3)|. \end{aligned}$$

By Equation 5.3, in Remark 5.1.5,

$$|u_{yy}(x_i, y_j^3)| \leq \|u\|_{C_s^{2+\alpha}(\overline{\mathscr{O}})} (y_j^3)^{\alpha/2-1},$$

and since $y_j^3 \geq y_{j-1}$ and $j > 1$, then $y_j^3 \geq h_y$, and so,

$$\begin{aligned} y_j \left| \widehat{u^{i,j}_{yy}} - u_{yy}(x_i, y_j) \right| &\leq \frac{|y_j u_{yy}(x_i, y_j^3) - y_j u_{yy}(x_i, y_j)|}{s((x_i, y_j^3), (x_i, y_j))^\alpha} s((x_i, y_j^3), (x_i, y_j))^\alpha \\ &\quad + h_y \|u\|_{C_s^{2+\alpha}(\overline{\mathscr{O}})} (y_j^3)^{\alpha/2-1} \\ &\leq \|u\|_{C_s^{2+\alpha}(\overline{\mathscr{O}})} h_y^{\alpha/2} + \frac{h_y}{h_y^{1-\alpha/2}} \|u\|_{C_s^{2+\alpha}(\overline{\mathscr{O}})} \\ &\leq 2\|u\|_{C_s^{2+\alpha}(\overline{\mathscr{O}})} h_y^{\alpha/2}. \end{aligned}$$

The discussion regarding this approximate derivative is finished. We have only one more term to estimate.

v) $\widehat{u^{i,j}_{xy}} = \frac{1}{4h_x h_y}(u_{i+1,j+1} - u_{i+1,j-1} - u_{i-1,j+1} + u_{i-1,j-1})$: Notice that $j > 1$ since there is no u_{xy} term in A along $\{y = 0\}$. This term is the central difference in the x-direction of the central difference in the y-direction, however we need to be a bit

careful in this case with the estimate as we do not know regularity past order 2. Let us re-write $\widehat{u_{xy}^{i,j}}$ as

$$\widehat{u_{xy}^{i,j}} = \frac{1}{2h_x}\left(\frac{1}{2h_y}(u_{i+1,j+1} - u_{i+1,j-1}) - \frac{1}{2h_y}(u_{i-1,j+1} - u_{i-1,j-1})\right) \tag{6.19}$$

First, by the Mean Value theorem there exist $y_j^1 \in (y_j, y_{j+1})$ and $y_j^2 \in (y_{j-1}, y_j)$ such that

$$\begin{aligned} u_{i+1,j+1} &= u_{i+1,j} + h_y u_y(x_{i+1}, y_j) + \frac{h_y^2}{2} u_{yy}(x_{i+1}, y_j^1), \\ u_{i+1,j-1} &= u_{i+1,j} - h_y u_y(x_{i+1}, y_j) + \frac{h_y^2}{2} u_{yy}(x_{i+1}, y_j^2), \end{aligned}$$

and thus,

$$\frac{u_{i+1,j+1} - u_{i+1,j-1}}{2h_y} = u_y(x_{i+1}, y_j) + \frac{h_y}{4}(u_{yy}(x_{i+1}, y_j^1) - u_{yy}(x_{i+1}, y_j^2)).$$

Similary, there exist $y_j^3 \in (y_j, y_{j+1})$ and $y_j^4 \in (y_{j-1}, y_j)$ such that

$$\frac{u_{i-1,j+1} - u_{i-1,j-1}}{2h_y} = u_y(x_{i-1}, y_j) + \frac{h_y}{4}(u_{yy}(x_{i-1}, y_j^3) - u_{yy}(x_{i-1}, y_j^4)).$$

Hence, from Equation (6.19) it follows that

$$\begin{aligned} \widehat{u_{xy}^{i,j}} =& \frac{1}{2h_x}(u_y(x_{i+1}, y_j) - u_y(x_{i-1}, y_j)) \\ &+ \frac{h_y}{8h_x}(u_{yy}(x_{i+1}, y_j^1) - u_{yy}(x_{i+1}, y_j^2) - (u_{yy}(x_{i-1}, y_j^3) - u_{yy}(x_{i-1}, y_j^4))). \end{aligned} \tag{6.20}$$

Let us focus on the first term of Equation (6.20). By the Mean Value theorem we have that there exist $x_i^1 \in (x_i, x_{i+1})$ and $x_i^2 \in (x_{i-1}, x_i)$ such that

$$\begin{aligned} u_y(x_{i+1}, y_j) &= u_y(x_i, y_j) + h_x u_{xy}(x_i^1, y_j), \\ u_y(x_{i-1}, y_j) &= u_y(x_i, y_j) - h_x u_{xy}(x_i^2, y_j), \end{aligned}$$

and so by the Mean Value theorem for Sums there exist $x_i^3 \in (x_{i-1}, x_{i+1})$ such that

$$\frac{1}{2h_x}(u_y(x_{i+1}, y_j) - u_y(x_{i-1}, y_j)) = u_{xy}(x_i^3, y_j).$$

Hence, from (6.20) it follows that

$$\begin{aligned} \widehat{u_{xy}^{i,j}} =& u_{xy}(x_i^3, y_j) \\ &+ \frac{h_y}{8h_x}(u_{yy}(x_{i+1}, y_j^1) - u_{yy}(x_{i+1}, y_j^2) - (u_{yy}(x_{i-1}, y_j^3) - u_{yy}(x_{i-1}, y_j^4))), \end{aligned}$$

which, arguing as we did for the case of $\widehat{u_{yy}^{i,j}}$, given that again $j > 1$, implies that

$$y_j \left| \widehat{u_x^{i,j}} - u_{xy}(x_i, y_j) \right| \leq \frac{h_y}{8h_x} \|u\|_{C_s^{2+\alpha}(\overline{\mathscr{O}})} |y_j^2 - y_j^1|^{\alpha/2} + \frac{h_y}{8h_x} \|u\|_{C_s^{2+\alpha}(\overline{\mathscr{O}})} |y_j^4 - y_j^3|^{\alpha/2}.$$

Now, since $|y_j^2 - y_j^1| \leq 2h_y$ and $|y_j^2 - y_j^1| \leq 2h_y$,

$$y_j \left| \widehat{u_x^{i,j}} - u_{xy}(x_i, y_j) \right| \leq 2^{\alpha/2-2} \frac{h_y}{h_x} \|u\|_{C_s^{2+\alpha}(\overline{\mathscr{O}})} h_y^{\alpha/2},$$

and given that by hypothesis $h_y/h_x \leq \Lambda$, we conclude that

$$\left| \widehat{u_{xy}^{i,j}} - u_{xy}(x_i, y_j) \right| \leq 2^{\alpha/2-2} \Lambda \|u\|_{C_s^{2+\alpha}(\overline{\mathscr{O}})} h_y^{\alpha/2}.$$

Putting together the estimates for i), ii), iii), iv) and v), then Equation (6.17) follows. The estimates for Equation (6.18) follow similarly without any additional difficulty. □

We continue the exposition along the lines of Krylov's book [30, Sections 6.6-6.7] by now providing a lemma equivalent to [30, Lemma 6.7.1].

Lemma 6.3.2. *Let $\Lambda > 0$. There are positive constants $h_{x,0}$, $h_{y,0}$ depending on $\mathscr{O}, \alpha$, and the parameters appearing in the definition of A, such that for all $h_x \leq h_{x,0}$ and $h_y \leq h_{y,0}$, with $h_y/h_x \leq \Lambda$, and for any bounded functions $f, g \in C\left(\overline{\mathscr{O}}\right)$, the system of linear equations*

$$\begin{cases} \mathcal{L}^h u_{i,j} = f(x_i, y_j) \text{ for } (x_i, y_j) \text{ in } \mathscr{O} \cup \Gamma_0, \\ u(x_i, y_j) = g(x_i, y_j) \text{ for } (x_i, y_j) \text{ on } \Gamma_1 \end{cases} \tag{6.21}$$

has a unique solution $u_h(x_i, y_j)$, where $h = (h_x, h_y)$ and $(x_i, y_j) \in \mathscr{O} \cup \Gamma_0$. In addition,

$$\max_{(x_i,y_j)\in\overline{\mathscr{O}}} |u_h(x_i, y_j)| \leq K \max_{(x_i,y_j)\in\mathscr{O}\cup\Gamma_0} |f(x_i, y_j)| + \max_{(x_i,y_j)\in\Gamma_1} |g(x_i, y_j)| \tag{6.22}$$

for some constant K that depends on the height of $\mathscr{O}$, r, κ and θ. The same assertion holds for $\widetilde{\mathcal{L}}$.

Proof. The existence and uniqueness of a solution to the linear system (6.21) follow from the fact that $\mathcal{L}$ and $\widetilde{\mathcal{L}}$ satisfy the discrete maximum principle. As far as Inequality (6.22) is concerned, it suffices to prove that

$$\max_{(x_i,y_j)\in\overline{\mathscr{O}}} (u_h(x_i, y_j))_+ \leq K \max_{(x_i,y_j)\in\mathscr{O}\cup\Gamma_0} (f(x_i, y_j))_+ + \max_{(x_i,y_j)\in\Gamma_1} (g(x_i, y_j))_+ \tag{6.23}$$

Let $v_0(x,y) = y + c$, with c large enough so that $rc - \kappa\theta > 1$. Observe that

$$\begin{aligned} Av_0 &= -\frac{y}{2}(0) - (r - q - \frac{y}{2})0 - \kappa(\theta - y) + r(y + c) \\ &= (rc - \kappa\theta) + (\kappa + r)y \geq rc - \kappa\theta > 1, \end{aligned}$$

so that by Lemma 6.3.1, we can choose $h_0 = (h_{x,0}, h_{y,0})$ such that both $\mathcal{L}^h v_0(x_i, y_j) \geq 1/2$ and $\widetilde{\mathcal{L}}^h v_0(x_i, y_j) \geq 1/2$, for all $h = (h_x, h_y)$ with $h_x \leq h_{x,0}$ and $h_y \leq h_{y,0}$, and all discrete interior points $(x_i, y_j) \in \mathscr{O} \cup \Gamma_0$.

Take a solution u_h of Equation (6.21) and consider $w^\varepsilon = u_h - 2(F + \varepsilon)v_0 - G$, where $F = \max_{\mathscr{O}\cup\Gamma_0}(f(x_i, y_j))_+$, $G = \max_{\Gamma_1}(g(x_i, y_j))_+$, and ε is a positive constant. We want to prove that $w^\varepsilon \leq 0$ in $\overline{\mathscr{O}}$. For that, suppose by contradiction that $w^\varepsilon > 0$ for some points in $\overline{\mathscr{O}}$ and define $(x_i^\varepsilon, y_j^\varepsilon) \in \overline{\mathscr{O}}$ to be a node where the maximum w^ε is achieved. Notice such a point cannot lie on $\overline{\Gamma_1}$ because otherwise $w(x_i^\varepsilon, y_j^\varepsilon) = (g(x_i^\varepsilon, y_j^\varepsilon) - G) - 2(F + \varepsilon)v_0 \leq 0$, given that v_0 is clearly non-negative. That would be a contradiction. Hence, such a point must be an interior node in $\mathscr{O} \cup \Gamma_0$.

Now, by the discrete maximum principle (or by direct calculation of values of the operators on constant functions), we have that $\mathcal{L}^h G \geq 0$ and $\widetilde{\mathcal{L}}^h G \geq 0$. By letting $\mathcal{L}^h$ and $\widetilde{\mathcal{L}}^h$ act on w^ε we obtain,

$$\begin{aligned} \mathcal{L}^h w_{i,j}^\varepsilon &= \mathcal{L}^h u_h(x_i, y_j) - 2(F + \varepsilon)\mathcal{L}^h v_0(x_i, y_j) - \mathcal{L}^h G \\ &\leq f(x_i, y_j) - 2(F + \varepsilon)\frac{1}{2} \\ &\leq -\varepsilon \leq 0 \end{aligned}$$

for all $(x_i, y_j) \in \mathscr{O} \cup \Gamma_0$ and all $h = (h_x, h_y)$ with $h_x \leq h_{x,0}$ and $h_y \leq h_{y,0}$. Therefore, by the discrete maximum principle once again, we reach a contradiction to the existence of $(x_i^\varepsilon, y_j^\varepsilon)$. We then have that $w^\varepsilon \leq 0$ for all $\varepsilon > 0$ and all $h_x \leq h_{x,0}$ and $h_y \leq h_{y,0}$, and by letting ε tend to 0, we obtain Equation (6.23). □

Theorem 6.3.3 (Convergence of finite-difference schemes)**.** *Let $h = (h_x, h_y) \leq h_0 = (h_{x,0}, h_{y,0})$ and $f, g \in C(\overline{\mathscr{O}})$. Denote by u_h the discrete solution to the linear system in* (6.21)*, and by $u \in C_s^{2+\alpha}(\overline{\mathscr{O}})$ the unique solution to Problem 6.1. Then,*

$$\max_{(x_i,y_j)\in\overline{\mathscr{O}}} |(u - u_h)(x_i, y_j)| \leq K \max\{h_x, h_y\}^{\alpha/2} \|u\|_{C_s^{2+\alpha}(\overline{\mathscr{O}})},$$

for some constant K that depends on α, the height of $\mathscr{O}$, Λ, and the constant parameters that appear in A.

Proof. Consider $w_h := u_h - u$ defined on the nodes of $\overline{\mathscr{O}}$. Notice that w_h solves the following linear system,

$$\begin{cases} \mathcal{L}^h w_h(x_i, y_j) = f(x_i, y_j) - \mathcal{L}^h u(x_i, y_j), & \text{for all } (x_i, y_j) \in \mathscr{O} \cup \Gamma_0, \\ w_h(x_i, y_j) = (g - u)(x_i, y_j) = 0, & \text{for all } (x_i, y_j) \in \Gamma_1, \end{cases}$$

Hence, by Lemma 6.3.2 we have that,

$$\begin{aligned} \max_{(x_i,y_j)\in\overline{\mathscr{O}}} |w_h(x_i, y_j)| &\leq K \max_{(x_i,y_j)\in\mathscr{O}\cup\Gamma_0} |f(x_i, y_j) - \mathcal{L}^h u(x_i, y_j)| + 0 \\ &\leq K \max_{(x_i,y_j)\in\mathscr{O}\cup\Gamma_0} |Au(x_i, y_j) - \mathcal{L}^h u(x_i, y_j)| \end{aligned}$$

and by Lemma 6.3.1, we obtain,

$$\max_{(x_i,y_j)\in\overline{\mathscr{O}}} |w_h(x_i, y_j)| \leq K \max\{h_x^\alpha, h_y^\alpha\} \|u\|_{C_s^{2+\alpha}(\overline{\mathscr{O}})}$$

This completes the proof. □

6.4 Numerical solution of the Heston boundary value problem via the finite-difference method

For consistency and comparison reasons we will solve with finite differences the exact same problems we solved with finite elements.

Consider Problem 3.45,

$$\begin{cases} Au = f & \text{a.e. in } \mathscr{O} \\ u = g & \text{on } \Gamma_1 \end{cases} \tag{6.24}$$

with the same values for the parameters of the Heston operator A, the same domain $\mathscr{O} = (-L, L) \times (0, V)$ with $L = 1$ and $V = 2$, same source function $f = -\frac{1}{2}$ and boundary condition g given by the restriction to Γ_1 of $\tilde{g}$, as in Equation (3.46). Hence the Heston boundary value problem is once again given by Equation (3.47),

$$\begin{cases} Au = -\dfrac{1}{2} - A\tilde{g} & \text{a.e. in } \mathscr{O} \\ u = 0 & \text{on } \Gamma_1 \end{cases} \tag{6.25}$$

Solving this problem by the finite-difference method given by the operator $\mathcal{L}$ (same answers were achieved by using operator $\widetilde{\mathcal{L}}$), gives us the approximate solution illustrated in Figure 6.3.

Since the finite-difference solutions are defined only on mesh nodes, they need to be extended to the entire domain $\mathscr{O}$. There are multiple ways of doing so and we chose to extend them by doing a bilinear interpolation on each mesh sub-rectangle. Notice this way of interpolating keeps the function values on non mesh-nodes bounded by the function values on mesh-nodes, hence it does not perturb the L^∞ norm between the approximating function and the true solution.

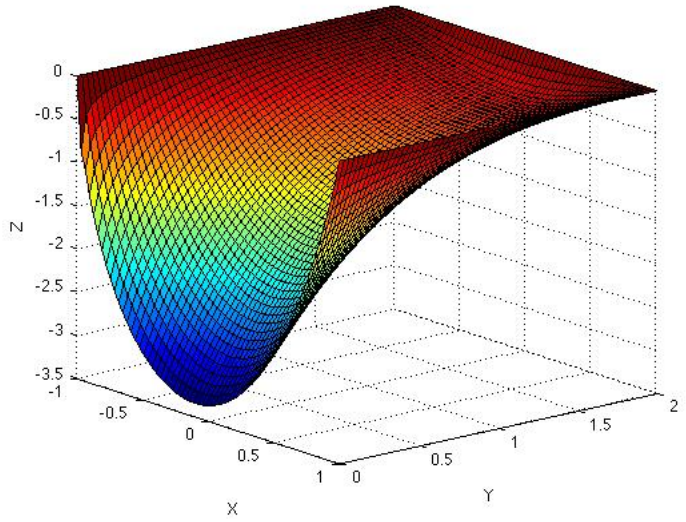


Figure 6.3: Approximate solution to the homogeneous problem, Problem 6.25

The graph in Figure (6.3) was obtained by partitioning the domain $\mathscr{O}$ in a homogenous mesh of size 64×64. The solution $u^{64\times 64}$ is identically zero along Γ_1 and takes values along Γ_0 implied by being the solution to the linear system associated to the approximating operator $\mathcal{L}$, instead of by a prescription of a boundary condition.

We have added in Figure (6.4) the graph of the non-homogeneous approximate solution to Problem 6.24 solved by the finite-difference method for completeness. The solutions to both homogenous and non-homogeneous boundary value problems are identical to the ones obtained by the finite-element method.

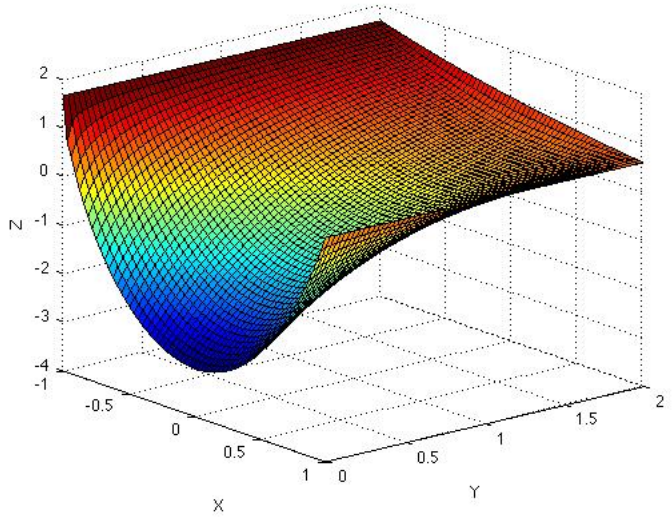


Figure 6.4: Approximate solution to the non-homogeneous problem, Problem (6.24)

As we did with the finite-element method, we include graphs of approximate solutions for increasingly finer meshes, in Figure 6.5, this time to the homogenous boundary value problem instead of the non-homogenous one, and summarize in Table 6.1 numerical evidence of convergence of the finite-difference solutions to the solution of Problem 6.25.

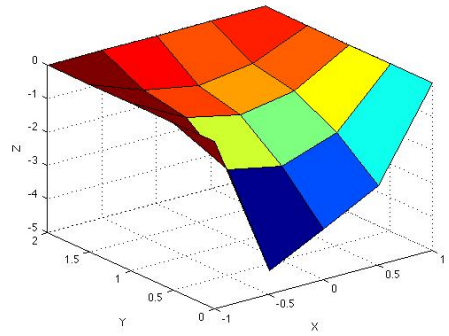
Iteration (i)	Mesh ($N_i \times N_i$)	Time [seconds]	Increment (e_i)	Order (α_i)
1	4×4	0.03	————	————
2	14×14	0.42	1.063686	————
3	24×24	1.29	0.452378	0.6825
4	34×34	2.73	0.197707	1.5357
5	44×44	4.85	0.112111	1.6287
6	54×54	8.21	0.072794	1.6750
7	64×64	13.03	0.051358	1.7032
8	74×74	21.23	0.038329	1.7223
9	84×84	44.20	0.029790	1.7360
10	94×94	($\sim$ 1.2m) 71.64	0.023875	1.7463
11	104×104	($\sim$ 1.7m) 102.48	0.019600	1.7541

Table 6.1: Numerical results for finite differences on the Heston boundary value problem

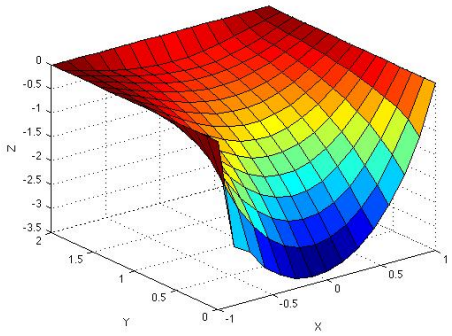
The first clear observation by comparing Tables 6.1 and 3.1 is how much faster the code implementation for finite differences is compared to the one for finite elements, while the difference in between consecutive approximate solutions is virtually the same.

Besides taking advantage of the sparseness of the matrix of the linear system of Equation (6.13) when calculating its inverse, there is no improvement left to be done on the finite-difference method to achieve a faster solvability for a given mesh size.

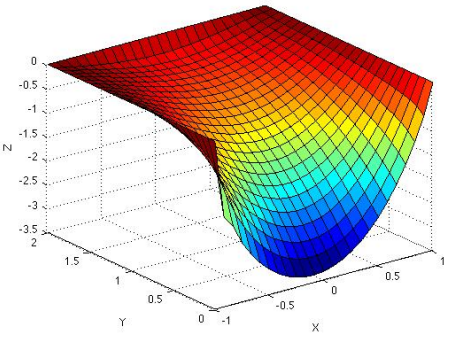
Second, the quantity e_i labeled in Table 6.1 as Increment is the same defined in Equation (3.48). By Theorem 6.3.3, when we are approximating the solution to a Heston boundary value problem that we expect to be $C_s^{2+\alpha}$, then we also expect the e_i to converge to zero with an order of convergence of at least $\alpha/2$, but by Theorem 5.2.1, given that f and g are smooth, for every positive $\alpha \in (0, 1)$ we expect this order of convergence to be at least $\alpha/2$ as the mesh gets finer, hence at least we expect this order of convergence to be of order one half. As it turns out for most classic examples of strictly elliptic second-order operators [30, Page 88], the actual convergence is faster than theoretically expected, and this is evidenced by the data collected in Table 6.1.



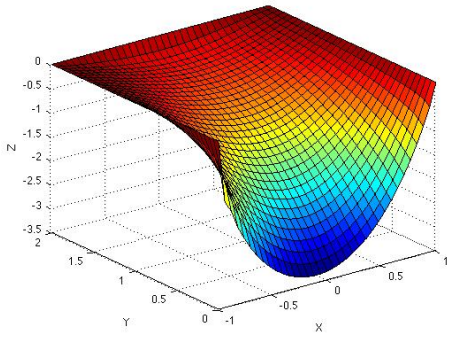
(a) 4×4 subintervals



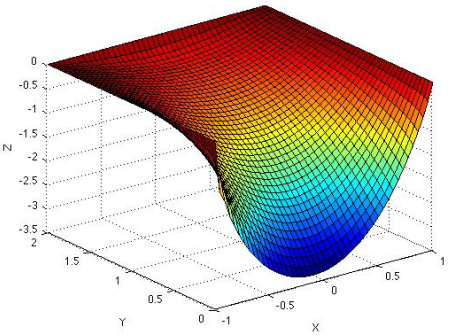
(b) 14×14 subintervals



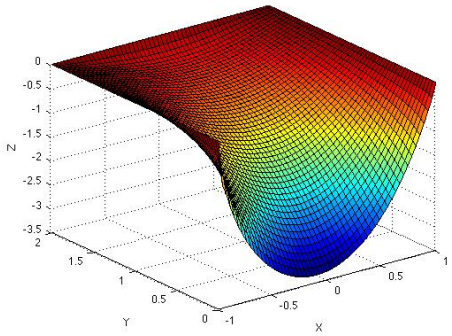
(c) 24×24 subintervals



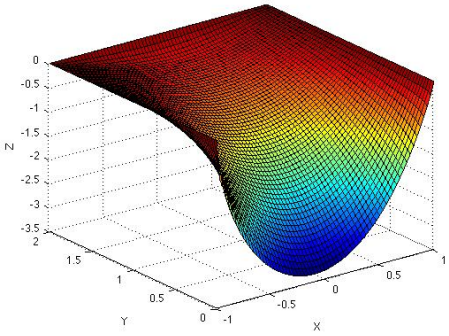
(d) 34×34 subintervals



(e) 44×44 subintervals



(f) 54×54 subintervals



(g) 64×64 subintervals

Figure 6.5: Finite-difference solutions of the boundary value problem for the homogenous elliptic Heston operator

Chapter 7

Finite-difference method for elliptic Heston obstacle value problem

The homogeneous elliptic Heston obstacle Problem 2.21 can be written equivalently in the form

$$\begin{cases} Au \geq f \text{ a.e. in } \mathscr{O} \\ u \geq \psi \text{ a.e. in } \mathscr{O} \\ (Au - f)(u - \psi) = 0 \text{ a.e. in } \mathscr{O} \\ u = 0 \text{ on } \Gamma_1 \end{cases} \tag{7.1}$$

We present two approaches for solving numerically this problem. First, by reducing Problem 7.1 to a non-linear equality problem, which gets solved by combining finite differences with the Newton method, and second by an iterative method, the projection method.

For comparison reasons, we illustrate these methods for the same input used in the homogeneous boundary value case. That is, we use the same values for the parameters of the Heston operator A, the same domain $\mathscr{O} = (-L, L) \times (0, V)$ with $L = 1$ and $V = 2$, and same source function $f = -\frac{1}{2} - A\tilde{g}$ with $\tilde{g}$, as in Equation (3.46).

We conveniently consider $\psi = -1$ and notice from Figure 6.5 that ψ will indeed be a non-trivial barrier for the homogeneous boundary value problem associated to problem 7.1. This is equivalent to having set the original barrier function for the non-homogeneous version of Problem 7.1 as $\tilde{\psi} = \tilde{g} - 1$ (so that $\tilde{\psi} \leq \tilde{g}$ on Γ_1).

7.1 Penalty method

By analogy with the penalty method for elliptic variational inequalities [25, Section I.7], we define the functional $j : H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w}) \longrightarrow \overline{\mathbb{R}}$ by

$$j(w) := \frac{1}{2} \int_{\mathscr{O}} ((\psi - w)^+)^2 \, d\mathfrak{w}$$

It is easy to see that j is a convex, proper, lower semi-continuous functional, $j(w) = 0$ if and only if $w \in \mathbb{K} = \left\{ w \in H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w}) | w \geq \psi \text{ a.e in } \mathscr{O} \right\}$, and $j(w) \geq 0$ for all $w \in H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$. Also, for each $\varepsilon > 0$ we define

$$j_\varepsilon(w) := \frac{1}{\varepsilon} j(w)$$

and the penalization operator by

$$\beta_\varepsilon(w) := j'_\varepsilon(w) = -\frac{1}{\varepsilon}(\psi - w)^+ \tag{7.2}$$

We consider the penalized (non-linear) boundary value problem associated to the elliptic Heston obstacle problem,

$$\begin{cases} Au + \beta_\varepsilon(u) = f \text{ a.e. in } \mathscr{O} \\ u = 0 \text{ on } \Gamma_1 \end{cases} \tag{P_ε}$$

In Problem 7.1 we were looking for functions $u \in H^2(\mathscr{O}, \mathfrak{w})$ such that $u \geq \psi$ a.e. in $\mathscr{O}$, and thus we can think of Problem P_ε as consisting on "penalizing" functions u_ε such that $u_\varepsilon < \psi$ on subsets of $\mathscr{O}$ with positive measure, by making $\beta_\varepsilon(u)$ a very large negative number as ε goes to zero, guaranteeing that Au_ε will be at least of size f.

If $u_\varepsilon \in H^2(\mathscr{O}, \mathfrak{w})$ solves P_ε, then u_ε solves the penalized equation for the elliptic Heston bilinear form,

$$\mathfrak{a}(u_\varepsilon, v) + (\beta_\varepsilon(u_\varepsilon), v)_H = (f, v)_H, \text{ for all } v \in H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w}) \tag{7.3}$$

and thus, given we consider only open subsets, $\mathscr{O}$, that are bounded in the x-direction, the Heston bilinear form is coercive in $H_0^1(\mathscr{O} \cup \Gamma_0, \mathfrak{w})$. By standard theory of penalty methods for elliptic variational inequalities of the first kind, [25, Theorem 7.1], it follows both that $u_\varepsilon \longrightarrow u$ in $H^1(\mathscr{O}, \mathfrak{w})$ and $j_\varepsilon(u_\varepsilon) \longrightarrow 0$ a.e. in $\mathscr{O}$, as $\varepsilon \to 0$.

Hence we can focus on solving approximately the non-linear Problem P_ε. We can do so by combining the finite-difference method with the Newton method.

First, using the same finite differences used in (6.3)-(6.8) we write (P_ε) as

$$\sum_{k=i-1}^{k=i+1}\sum_{l=j-1}^{l=j+1} m_{k,l}u_{k,l} - \frac{2h_x^2h_y^2}{\varepsilon}(\psi_{i,j}-u_{i,j})^+ = 2h_x^2h_y^2 f_{i,j} \tag{7.4}$$

for each interior point $(x_i, y_j) \in \mathscr{O}\cup\Gamma_0$. Notice that the $m_{k,l}$'s coefficients are the same coefficients of Equation (6.9). After re-indexing with the same index map I used in Equation (6.12), we have for each interior node $t = I(i,j)$ the equation

$$\sum_{s=I(k,l)} \widehat{m}_{t,s}u_s - \frac{2h_x^2h_y^2}{\varepsilon}(\psi_r-u_t)^+ = 2h_x^2h_y^2 f_t$$

This non-linear problem on the unknowns $\{u_s\}_s$ can be written as

$$\widehat{M}\vec{u} - \frac{2h_x^2h_y^2}{\varepsilon}\left(\vec{\psi}-\vec{u}\right)^+ = 2h_x^2h_y^2\vec{f} \tag{7.5}$$

where $\widehat{M} = \{\widehat{m}_{t,s}\}_{t,s}$ is a sparse matrix, and $\vec{u}, \vec{\psi}, \vec{f} \in \mathbb{R}^{(M-2)(N-1)}$.

Define,

$$F(\vec{u}) := \widehat{M}\vec{u} - \frac{2h_x^2h_y^2}{\varepsilon}\left(\vec{\psi}-\vec{u}\right)^+ - 2h_x^2h_y^2\vec{f}$$

We want to find $\vec{u} \in \mathbb{R}^{(M-2)(N-1)}$ such that $F(\vec{u}) = \vec{0}$. For that, by Newton method, all we need to do is to solve for $\vec{u}^{n+1}$ in the iterative formula

$$DF(\vec{u}^n)\left(\vec{u}^{n+1}-\vec{u}^n\right) = -F(\vec{u}^n) \tag{7.6}$$

where DF is the derivative of F and $\vec{u}^0$ is an initial guess, say the one given by $\widehat{M}\vec{u}^0 = 2h_x^2h_y^2\vec{f}$. It is easy to see that the derivative of F, DF, is nothing but

$$DF(\vec{u}) = \widehat{M} - \frac{2h_x^2h_y^2}{\varepsilon}J(\vec{u})$$

where $J(\vec{u}) \in \mathbb{R}^{(M-2)(N-1)\times(M-2)(N-1)}$ is a diagonal matrix with diagonal entries

$$J_{s,s}(\vec{u}) = \begin{cases} 0, u_s > \psi_s \\ -1, u_s < \psi_s \end{cases}$$

We implemented in MATLAB this combined method and obtained the numerical results summarized in Table 7.1 together with the graphs presented in Figure 7.1. For

each mesh size, our code took less time than the one coded for the finite-element method, for a decreasing sequence of ε's of the form $\varepsilon_k = 10^{-k}$ where we didn't have to go past $k = 4$, however this time increases fairly quickly as seen on the Time column of Table 7.1.

For each mesh size, and for a given ε_k, we used an incremental error precision of 10^{-6} between a couple of consecutive solutions as the admissible value when to consider equation (7.6) to be solved. For the example illustrated here, the maximum number of said iterations needed was of at most 17 and it obviously happened for ε_4.

The different graphs in Figure 7.1 show the approximate finite-difference solutions to the non-homogeneous obstacle problem for the elliptic Heston operator. In all of them, we are also showing the original barrier function $\tilde{\psi}$ together with an approximation to the "free boundary", $\partial\left\{u > \tilde{\psi}\right\} \cap \mathscr{O}$. As expected, this seems to be a continuous connected curve, though we did not look into proving that in general.

Iteration (i)	Mesh ($N_i \times N_i$)	Time [seconds]	Increment (e_i)	Order (α_i)
1	4×4	0.20	———	———
2	14×14	3.50	0.714309	———
3	24×24	12.72	0.416660	0.4303
4	34×34	33.23	0.214265	1.2339
5	44×44	($\sim$ 1.34m) 80.52	0.085864	2.6254
6	54×54	174.15	0.059583	1.4172
7	64×64	($\sim$ 6.65m) 399.17	0.042326	1.6698
8	74×74	($\sim$ 12.34m)740.83	0.031655	1.7098
9	84×84	($\sim$ 25.72m)1543.28	0.024625	1.7299

Table 7.1: Numerical results for finite differences, combined with the Newton method, to solve the Heston obstacle problem

Our code was not fully optimized to take advantage of the sparseness of the matrix $\widehat{M}$ and thus when we are looking for an approximate solution to Problem 7.1 over a grid of size 100×100 (roughly 10^4 nodes), which implies dealing with a sparse matrix of size 10^4 by 10^4, our personal computer would run out of memory, as it attempted to allocate memory in the stack for storing a lot of zeros that are not necessary to track of. If that enhancement to the code is done, then one can consider much finer meshes.

7.2 Relaxation method with projection

Another way of solving Problem 7.1 is by using an over-relaxation method with projection as described in [25, Section V.5] for the discretization of problem 7.1, say by using finite differences, however there is a hypothesis that is not fulfilled by the matrix in the resulting linear complimentarily problem. Namely, the matrix is not symmetrical.

Consider a homogenous mesh of of $\mathscr{O} = (-L, L) \times (0, V)$ with $M-1$ subintervals length h_x in the x direction, $N-1$ subintervals of length h_y in the y direction, and let $Q = (M-2)(N-1)$. Indeed, Problem 7.1 after being discretized by using the finite-difference approximations (6.3)-(6.8), becomes the linear complementarity problem of finding $\vec{u} \in \mathbb{R}^Q$ such that

$$\begin{cases} \widehat{M}\vec{u} \geq 2h_x^2 h_y^2 \vec{f} \\ \vec{u} \geq \vec{\psi} \\ \left(\widehat{M}\vec{u} - 2h_x^2 h_y^2 \vec{f}\right)^T \left(\vec{u} - \vec{\psi}\right) = 0 \end{cases} \tag{7.7}$$

where $\widehat{M}$ is the same matrix as in equation 6.13, and $\vec{f}$ and $\vec{\psi}$ are nothing but vectors with the valuations of f and ψ at the "interior" nodes of $\mathscr{O}_{h_x,h_y}$.

Its iterative formulation, for our case, is as follows. Let $\vec{u}^0 \in \mathbb{R}^Q$ be any seed point in $\mathcal{C} := \{(x_i, y_j) \in \mathscr{O}_{h_x,h_y} | u_t^0 \geq \psi_t \text{ with } t = I(i,j)\}$, where $\vec{u}^0 = \left(u_1^0, \ldots, u_Q^0\right)$. A perfectly first good choice could be $\vec{u}^0 = (\psi_1, \ldots, \psi_Q)$ where $\psi_t = \psi(x_k, y_l)$ and $t = I(k,l)$. Then, with $\vec{u}^n$ being known, we compute $\vec{u}^{n+1}$ component by component using the following formulas,

$$\bar{u}_s^{n+1} = \frac{1}{\widehat{m}_{s,s}} \left(b_s - \sum_{t=1}^{s-1} \widehat{m}_{s,t} u_j^{n+1} - \sum_{t=s+1}^{Q} \widehat{m}_{s,t} u_t^n \right) \tag{7.8}$$

$$u_s^{n+1} = \max\left\{\psi_s, u_s^n + w\left(\bar{u}_s^{n+1} - u_s^n\right)\right\} \tag{7.9}$$

for $s = 1, 2, \ldots, Q$, where $\vec{b} = \vec{f} - \widehat{M}\vec{\psi}$. The parameter w is known as the relaxation factor and from numerical experiments it has been found that the optimal value of w is always strictly greater than unity.

As a numerical experiment we went ahead and coded this formulation, regardless of our matrix $\widehat{M}$ not being symmetric, to look into whether convergence was achieved or

not, and whether we would find the same results as with the penalty method. We did so for $w = 1.5$ and it turns out, we found the same answers and we include in Table 7.2 a summary of the results for comparison with the penalty method.

Iteration (i)	Mesh ($N_i \times N_i$)	Time [seconds]	Increment (e_i)	Order (α_i)
1	4×4	0.06	————	————
2	14×14	0.48	0.714286	————
3	24×24	1.85	0.416667	0.4302
4	34×34	8.41	0.214267	1.2339
5	44×44	32.47	0.085863	2.6255
6	54×54	($\sim$ 1.74m) 104.30	0.059583	1.4172
7	64×64	247.67	0.042326	1.6698
8	74×74	($\sim$ 11.59m) 695.38	0.031655	1.7097
9	84×84	($\sim$ 30.15m) 1808.72	0.024624	1.7300
10	94×94	2580.99	0.019759	1.7367
11	104×104	($\sim$ 1.05h) 3793.03	0.016319	1.7003

Table 7.2: Numerical results for finite differences, combined with the relaxation method with projection, to solve the Heston obstacle problem

Notice that the Increment and Order columns of Tables 7.2 and 7.1 are pretty much identical, while the performance of the relaxation method with projection is slightly better. It is well known [25, Theorem 5.1] that the relaxation method with projection is convergent when the underlying bilinear form is symmetric, but even though that is not our case (the matrix $\widehat{M}$ is not symmetric), we have here an example for which the convergence is happening.

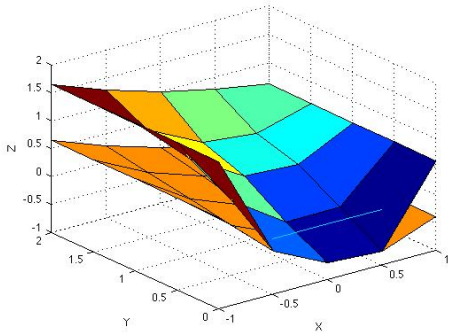
7.3 Convergence

Even though we have numerical evidence of convergence of these two methods combined with the finite-difference method, we haven't provided a convergence result for either of them, as we did it for the elliptic Heston boundary value problem.

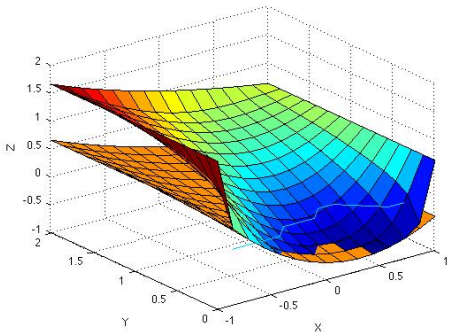
Even though we don't do it in this thesis, an approach to do so would be proving existence of viscosity solutions to viscosity solutions to the obstacle problem by adapting previous work of Barles [3] for existence of viscosity solutions to fully nonlinear boundary value problems with fully nonlinear boundary conditions, since we have a comparison principle proved by Feehan [12], and then adapting the work of Barles and Souganidis [4]

where they proved convergence of these finite difference schemes to fully nonlinear second order (non necessarily strictly elliptic) equations.

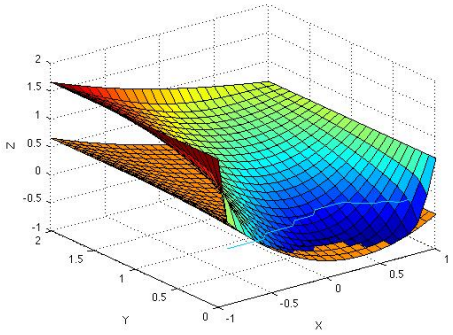
In the case of the penalty method combined with finite differences, one should be able to prove convergence and find a lower bound for the rate of convergence to the unique solution of the obstacle problem for the elliptic Heston operator, by combining the convergence results of the penalty method itself, and say, the convergence results of the Newton method. Another approach would be to adapt Krylov [30, Sections 6.6-6.7] results combined with maximum principle results proved by Feehan [17].



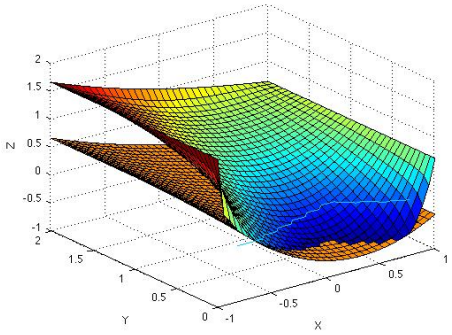
(a) 4×4 subintervals



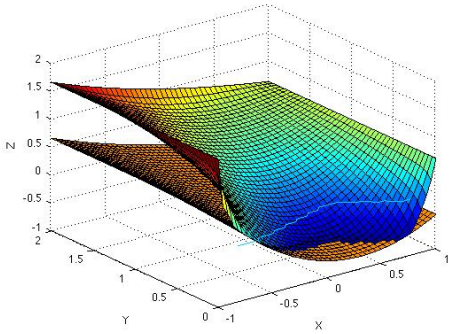
(b) 14×14 subintervals



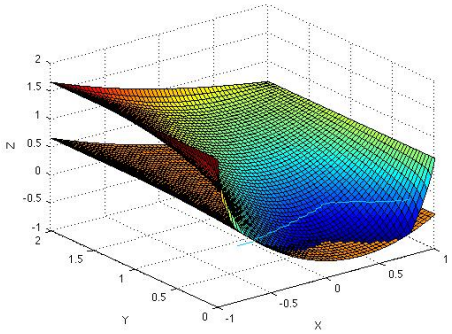
(c) 24×24 subintervals



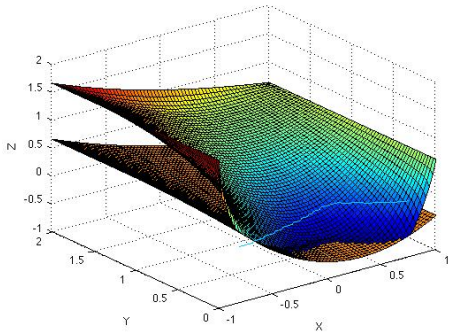
(d) 34×34 subintervals



(e) 44×44 subintervals



(f) 54×54 subintervals



(g) 64×64 subintervals

Figure 7.1: Finite-difference solutions, combined with the Newton method, of the obstacle problem for the elliptic Heston operator.

Chapter 8

Conclusions

- The idea of working with Sobolev weighted spaces, by Feehan and Daskalopoulos [12], is key to treat the degeneracy of the elliptic Heston operator. Even within that framework the coercivity of the bilinear form associated to the Heston operator, while continuous, it is not guaranteed to be coercive. We proved, under a practical condition on the underlying domain, general enough to cover most cases in applications, that this bilinear form is indeed coercive and no additional coercive bilinear form needs to be introduced to obtain existence and uniqueness results. This finding provides an appropriate framework to apply the finite-element methodology to solve numerically both the boundary value problem and the obstacle problem. Numerical simulations were obtained only for the equality case.

- Coercivity of the bilinear form of the Heston operator is highly desirable for finite elements. We extended our results for the Heston operator to a family of operators having the same kind of degeneracy and the same type of growth on its first and constant order terms.

- Performance of the finite-element implementation for the Heston boundary value problem and obstacle problems was not the goal of this work, however we observed that most of the time spent by our code is spent in calculating the fundamental integrals defining the bilinear form when calculated on the basis we chose. Either research should be done on finding more suitable bases for which we can have closed form formulas for the integrals so that they can be calculated by evaluation of functions on nodes, or the calculation of the fundamental integrals should be approximated itself with an order of accuracy that would not affect the expected

order of accuracy of the finite-element method.

- The choice of a good basis for the finite-element method does not only improve the performance of its implementation, but it simplifies the derivation of error estimates. We used Larry Schumaker's results [35, Chapter 12 and 13] about error estimates for Tensor Taylor expansions to prove convergence of the method and obtain a convergence rate, but we did so under the assumption that the solutions to the original problem were in $H^2(\mathscr{O})$ rather than in $H^2(\mathscr{O}, \mathfrak{w})$, which is a more appropriate assumption for our context. We believe Larry Schumaker's results can be easily extended to hold for weighted-L^p and Sobolev spaces, and proving them would soften our hypotheses.

- Numerical evidence for the finite-element method on the boundary value problem suggests a better rate of convergence holds, but via Cea's Lemma we were unable to prove it. Alternative and more particular approaches that look directly into the bilinear form of the variational problem and the $\|\cdot\|_{H^1(\mathscr{O},\mathfrak{w})}$ norm should be investigated.

- We presented two finite-difference schemes that were proved to be consistent. We didn't look into their stability, but we did prove they are convergent and provided an initial expected rate of convergence. Our numerical results once again indicate, not surprisingly, that the convergence is better than expected and their performance was far better than the implementation written for the finite-element method. Solutions under both methods were compared to be pretty much identical.

- There are multiple ways of approximating derivatives by finite differences and we considered only a handful of them. More approximating differences should be looked into towards finding finite difference schemes for which we can prove a discrete maximum principle for more values of the parameters of the Heston operator, while keeping their consistency, to continue getting convergence through the methods exposed. We followed a similar methodology to Krylov [30].

- The existence and uniqueness results for the Heston boundary value and obstacle problems expect no dependency on boundary information along Γ_0, hence our finite-element and finite-difference methods were designed to have that feature of not prescribing in their implementation the function values along $\{y = 0\}$ but to make them part of the unknowns to be governed by the approximation to the partial differential equality (or inequality in the case of the obstacle problem) of the problem. To our knowledge, this numerical approach hadn't been pursued before for the Heston operator.

Appendix A

Boundary value problems for the Cox-Ingersoll-Ross operator

A.1 Introduction to the Cox-Ingersoll-Ross ordinary differential equation

The generator (with killing), $-A$, of the Cox-Ingersoll-Ross process (CIR), the second stochastic differential equation in (1.1), is the CIR or Kummer operator [18], which can be written in the form,

$$Au = -yu'' - (\beta - y)u' + \alpha u$$

where $\alpha, \beta > 0$.

A.2 Analytical solution to the Cox-Ingersoll-Ross ordinary differential equation

Consider the boundary value problem

$$\begin{cases} -yu'' - (\beta - y)u' + \alpha u = 0 \quad \text{in } (0, L), \\ u(L) = \tilde{u}_L, \end{cases} \tag{A.1}$$

which is equivalent to

$$\begin{cases} yu'' + (\beta - y)u' - \alpha u = 0 \quad \text{in } (0, L), \\ u(L) = \tilde{u}_L, \end{cases} \tag{A.2}$$

where $\tilde{u}_L \in \mathbb{R}$, and suppose it has a unique solution $u \in C_s^{2+\alpha}([0, L])$. The differential equation in (A.2) is known as the *Kummer equation* and its general solution is given by

$$u(y) = c_1 U(y; \alpha, \beta) + c_2 M(y; \alpha, \beta),$$

where $c_1, c_2 \in \mathbb{R}$, and M and U are the confluent hypergeometric functions of the first and second kind [1, Section 13]. The asymptotic behavior of the functions M and U at 0 and at infinity are known [1]. Both functions M and U are analytic on $(0, \infty)$, but M is actually analytical on all of $\mathbb{R}$. If $\beta > 1$, then $U(y; \alpha, \beta) \approx y^{1-\beta}$ for y near 0, hence $c_1 = 0$, since $u \in C_s^{2+\alpha}([0, L])$. Also, if $\beta = 1$, then $U(y; \alpha, \beta) \approx \log y$ for y near 0, hence once again $c_1 = 0$. And finally, if $0 < \beta < 1$, then $U(y; \alpha, \beta) \approx y^{1-\beta}$, $U'(y; \alpha, \beta) \approx y^{-\beta}$, and $yU''(y; \alpha, \beta) \approx y^{-\beta}$, for y near 0, and thus c_1 must be 0 in this case as well given that $u \in C_s^{2+\alpha}([0, L])$. Therefore $c_1 \equiv 0$ and then,

$$u(y) = \tilde{u}_L \frac{M(y; \alpha, \beta)}{M(L; \alpha, \beta)}.$$

A.3 Finite-difference scheme for solving the Cox-Ingersoll-Ross boundary value problem

Consider the boundary value problem

$$\begin{cases} yu'' + (\beta - y)u' - \alpha u = 0 \quad \text{in } (0, L), \\ u(L) = \tilde{u}_L, \end{cases} \tag{A.3}$$

where $\alpha, \beta > 0$ and $\tilde{u}_L \in \mathbb{R}$. We are interested in setting up a finite-difference scheme for solving this problem. We will do so by using centered differences to approximate the derivatives at all points in $(0, L)$. At $y = 0$, we will need to estimate only one of the derivative terms, the one of order one since $u \in C_s^{2+\alpha}([0, L])$ and so the term yu'' vanishes (by Lemma 5.1.4). We will approximate it by using a forward difference. Consider then a uniform partition of the interval $[0, L]$,

$$\mathcal{P}_y : \{0 = y_1 < y_2 < y_3 < \cdots < y_N = L\}, \tag{A.4}$$

where $y_j = hj$ for $1 \leq j \leq N$, and $h = L/N$. Our finite-difference approximations are then

$$u'(y_j) \approx \begin{cases} \frac{1}{2h}(u_{j+1} - u_{j-1}) & \text{if } j > 1, \\ \frac{1}{h}(u_{j+1} - u_j) & \text{if } j = 1, \end{cases} \tag{A.5}$$

$$u''(y_j) \approx \frac{1}{h^2}(u_{j+1} - 2u_j + u_{j-1}), \tag{A.6}$$

where $u_j = u(y_j)$. Substituting Equations (A.5) and (A.6) in Equation (A.3), and noticing that $u_N = u(L) = \tilde{u}_L$, we then obtain have the following linear system:

$$\begin{aligned}
(-\beta - \alpha h)u_1 + (\beta)u_2 &= 0, \\
&\vdots \\
A_j u_{j-1} + B_j u_j + C_j u_{j+1} &= D_j, \\
&\vdots \\
(y_{N-1}(1 + h/2) - \beta h/2)u_{N-2} + (-2y_{N-1} - \alpha h^2)u_{N-1} &= -(y_{N-1}(1 - h/2) + \beta h/2)\tilde{u}_L,
\end{aligned}$$

in the unknowns $(u_j)_{1 \le j \le N-1}$, where

$$\begin{aligned}
A_j &:= y_j(1 + h/2) - \beta h/2, \\
B_j &:= -2y_j - \alpha h^2, \\
C_j &:= y_j(1 - h/2) + \beta h/2, \quad \text{and} \\
D_j &:= 0,
\end{aligned}$$

for $2 \le j \le N - 2$. We extend the formulas for A_j, B_j, C_j and D_j to $j = 1$ and $N - 1$ by making them zero for those case, and hence we write the preceding linear system as

$$\begin{pmatrix}
B_1 & C_1 & 0 & 0 & \cdots & 0 & 0 & 0 \\
A_2 & B_2 & C_2 & 0 & \cdots & 0 & 0 & 0 \\
0 & A_3 & B_3 & C_3 & \cdots & 0 & 0 & 0 \\
\vdots & & & & \ddots & & & \vdots \\
0 & 0 & 0 & 0 & \cdots & A_{N-2} & B_{N-2} & C_{N-2} \\
0 & 0 & 0 & 0 & \cdots & 0 & A_{N-1} & B_{N-1}
\end{pmatrix}
\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{pmatrix}
=
\begin{pmatrix} D_1 \\ D_2 \\ D_3 \\ \vdots \\ D_{N-2} \\ D_{N-1} \end{pmatrix}. \qquad \text{(A.7)}$$

We will show that this linear system has one and only one solution for all values of $\alpha > 0$, $\beta > 0$, $L > 0$, and $h > 0$, and we can prove this fact by means of the following two lemmas,

Lemma A.3.1 (Uniqueness of solutions to the finite-difference scheme for $\beta \le 2 + h$)**.** *Let α, β, L be positive numbers, and let $\mathcal{P}$ be a uniform partition of $[0, L]$ with $N > 1$ points, just like in Equation* (A.4)*, with $h \le 2$. Assume $\beta \le 2 + h$. Denote the matrix*

in Equation (A.7) *by*

$$\mathcal{M} := \begin{pmatrix} B_1 & C_1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ A_2 & B_2 & C_2 & 0 & \cdots & 0 & 0 & 0 \\ 0 & A_3 & B_3 & C_3 & \cdots & 0 & 0 & 0 \\ \vdots & & & & \ddots & & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & A_{N-2} & B_{N-2} & C_{N-2} \\ 0 & 0 & 0 & 0 & \cdots & 0 & A_{N-1} & B_{N-1} \end{pmatrix} \tag{A.8}$$

where,

$$A_j = \begin{cases} 0, & j = 1, \\ y_j(1 + h/2) - \beta h/2, & j > 1, \end{cases} \tag{A.9}$$

$$B_j = \begin{cases} -\beta - \alpha h, & j = 1, \\ -2y_j - \alpha h^2, & j > 1, \end{cases} \tag{A.10}$$

$$C_j = \begin{cases} \beta, & j = 1, \\ y_j(1 - h/2) + \beta h/2, & 1 < j < N - 1, \\ 0, & j = N - 1, \end{cases} \tag{A.11}$$

for $1 \leq j \leq N - 1$. *Then* $\mathcal{M}$ *is diagonally dominant and, in particular, invertible.*

Proof. We are interested in showing that $|B_j| > |A_j| + |C_j|$. Clearly, this holds for $j = 1$. Suppose then that $j > 1$. We will show that

$$2y_j + \alpha h^2 > |y_j(1 + h/2) - \beta h/2| + |y_j(1 - h/2) + \beta h/2|, \tag{A.12}$$

for all $j > 1$, by considering three cases.

Case 1. *Suppose* $y_j \geq \beta$. *Then Equation* (A.12) *is equivalent to*

$$2y_j + \alpha h^2 > y_j + (y_j - \beta)h/2 + |y_j(1 - h/2) + \beta h/2| \tag{A.13}$$

$$\iff \quad y_j + \alpha h^2 > (y_j - \beta)h/2 + |y_j(1 - h/2) + \beta h/2|. \tag{A.14}$$

Since $h \leq 2$, *then* (A.14) *is equivalent to*

$$y_j + \alpha h^2 > (y_j - \beta)h/2 + y_j(1 - h/2) + \beta h/2 \tag{A.15}$$

$$\iff \quad \alpha h^2 > 0, \tag{A.16}$$

which holds trivially.

Case 2. *Suppose $y_j < \beta$ but $y_j \geq \beta h/(2+h)$. Then Equation* (A.12) *is equivalent to*

$$\begin{aligned} & 2y_j + \alpha h^2 && > |y_j(1+h/2) - \beta h/2| + y_j(1-h/2) + \beta h/2 && \text{(A.17)} \\ \iff & y_j + \alpha h^2 && > |y_j(1+h/2) - \beta h/2| + (\beta - y_j)h/2 && \text{(A.18)} \\ \iff & y_j + \alpha h^2 && > y_j(1+h/2) - \beta h/2 + (\beta - y_j)h/2 && \text{(A.19)} \\ \iff & \alpha h^2 && > 0 && \text{(A.20)} \end{aligned}$$

which again is trivially true.

Case 3. *Suppose $y_j < \beta h/2 + h$. This implies that $y_j < \beta$. Then again we would have*

$$\begin{aligned} & 2y_j + \alpha h^2 && > |y_j(1+h/2) - \beta h/2| + y_j(1-h/2) + \beta h/2 && \text{(A.21)} \\ \iff & y_j + \alpha h^2 && > |y_j(1+h/2) - \beta h/2| + (\beta - y_j)h/2, && \text{(A.22)} \end{aligned}$$

and this last inequality would be equivalent to

$$\begin{aligned} \iff & y_j + \alpha h^2 && > \beta h/2 - y_j(1+h/2) + (\beta - y_j)h/2 && \text{(A.23)} \\ \iff & 2y_j + \alpha h^2 && > (\beta - y_j)h && \text{(A.24)} \\ \iff & y_j && > (\beta - \alpha h)\frac{h}{2+h}. && \text{(A.25)} \end{aligned}$$

Since $y_j \geq h$, given that $j > 1$, to verify that inequality (A.25) *holds, it is enough to verify that*

$$h > (\beta - \alpha h)\frac{h}{2+h}.$$

That happens if and only if

$$h > \frac{\beta - 2}{\alpha + 1},$$

which is equivalent to

$$h\alpha > \beta - 2 - h,$$

but that last statement is trivially true given that $\beta \leq 2 + h$.

This completes the proof. □

To prove the uniqueness of solutions to this finite-difference scheme in the case of $\beta > 2 + h$ we will use a result by L. Brugnano and D. Trigiante [7]. Consider the tridiagonal matrix,

$$\mathcal{T} = \begin{pmatrix} 1 & \tau_1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \sigma_1 & 1 & \tau_2 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \sigma_2 & 1 & \tau_3 & \cdots & 0 & 0 & 0 \\ \vdots & & & & \ddots & & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \sigma_{n-2} & 1 & \tau_{n-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & \sigma_{n-1} & 1 \end{pmatrix}. \tag{A.26}$$

It is easy to see that $\mathcal{T}$ can be factored as $\mathcal{T} = \mathcal{L}\mathcal{D}\mathcal{U}$, where $\mathcal{D} = (d_{ii})$ is a diagonal matrix, $\mathcal{L}$ is an invertible lower diagonal matrix with 1's along the diagonal, and $\mathcal{U}$ is an invertible upper diagonal matrix with 1's along the diagonal. Hence, $\mathcal{T}$ is invertible if and only if all of the d_{ii} are different from zero. Brugnano and Trigiante [7] give a sufficient condition for ensuring that this last condition holds.

Theorem A.3.2 (Brugnano-Trigiante's sufficient condition for invertibility of tridiagonal matrices)**.** [7] *Let* $\Delta_i := 1 - 4(\sigma_i\tau_i)^+$, $(\sigma\tau)^- := \min_i\{(\sigma_i\tau_i)^-\}$, *and* $m := \min_i \left\{(1 + \Delta_i^{1/2})/2\right\}$. *If* $\Delta_i \geq 0$ *for* $i = 1, \ldots, n-1$, *then*

$$m \leq d_{ii} \leq 1 - (\sigma\tau)^- m^{-1},$$

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

From this Theorem ,it follows that under the same hypotheses as in Lemma A.3.1, the original matrix $\mathcal{T}$ is invertible. Let us verify that a normalized version of our matrix, $\mathcal{M}$, for $\beta > 2 + h$, satisfies hypotheses of Theorem A.3.2.

Lemma A.3.3 (Uniqueness of solutions to the finite-difference scheme for $\beta > 2 + h$)**.** *Let* α, β, L *be positive numbers, and let* $\mathcal{P}$ *be a uniform partition of* $[0, L]$ *with* $N > 1$ *points, just like in Equation* (A.4)*, with* $h \leq 1/2$. *Assume that* $\beta > 2 + h$. *Let* $\mathcal{M}$ *denote the matrix in Equation* (A.7)*, normalized to have 1's along the diagonal, that*

is,

$$\mathcal{M} = \begin{pmatrix} 1 & \frac{C_1}{B_1} & 0 & 0 & \cdots & 0 & 0 & 0 \\ \frac{A_2}{B_2} & 1 & \frac{C_2}{B_2} & 0 & \cdots & 0 & 0 & 0 \\ 0 & \frac{A_3}{B_3} & 1 & \frac{C_3}{B_3} & \cdots & 0 & 0 & 0 \\ \vdots & & & & \ddots & & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \frac{A_{N-2}}{B_{N-2}} & 1 & \frac{C_{N-2}}{B_{N-2}} \\ 0 & 0 & 0 & 0 & \cdots & 0 & \frac{A_{N-1}}{B_{N-1}} & 1 \end{pmatrix}, \tag{A.27}$$

where,

$$A_j := \begin{cases} 0, & j = 1, \\ y_j(1 + h/2) - \beta h/2, & j > 1, \end{cases} \tag{A.28}$$

$$B_j := \begin{cases} -\beta - \alpha h, & j = 1, \\ -2y_j - \alpha h^2, & j > 1, \end{cases} \tag{A.29}$$

$$C_j := \begin{cases} \beta, & j = 1, \\ y_j(1 - h/2) + \beta h/2, & 1 < j < N - 1, \\ 0, & j = N - 1, \end{cases} \tag{A.30}$$

for $1 \leq j \leq N - 1$. *Then* $\mathcal{M}$ *is invertible.*

Proof. By Theorem A.3.2, we only need to verify that

$$\Delta_j := 1 - 4\left(\frac{A_{j+1}C_j}{B_{j+1}B_j}\right)^+ \geq 0$$

for all $j \leq N - 2$. First, notice that $B_{j+1}B_j > 0$ and $C_j > 0$ for all $j \leq N - 2$. If A_{j+1} was non-positive for all j, then we would be done as we would have $\Delta_j = 1 \geq 0$ for all j. Let us suppose then that A_{j+1} is positive for some $j \in \{1, \ldots, N - 2\}$. That is,

$$y_{j+1}(1 + h/2) - \beta h/2 > 0,$$

which is equivalent to

$$\beta < j(2 + h), \tag{A.31}$$

given that $y_{j+1} = jh$. Notice that $j > 1$, as if not then this would imply that $\beta < 2 + h$, which is not the case by hypothesis. Hence, if such a j exists then it must be true that

$j > 1$. We are interested in showing that $\Delta_j \geq 0$ in such cases as well, that is,

$$A_{j+1}C_j - \frac{1}{4}B_{j+1}B_j \leq 0.$$

Notice that for $j > 1$, the B_j and B_{j+1} do not depend on β, although the A_{j+1} and C_j still do. Let $p(\beta)$ denote the following polynomial in β,

$$p(\beta) := A_{j+1}(\beta)C_j(\beta) - \frac{1}{4}B_{j+1}B_j.$$

It follows that its derivative is given by

$$p'(\beta) = \frac{h^2}{2}\left(1 + \frac{y_j + y_{j+1}}{2} - \beta\right),$$

and p has a maximum at $\widehat{\beta} = 1 + (y_j + y_{j+1})/2$. Let us calculate that maximum. Indeed,

$$\begin{aligned}
p(\widehat{\beta}) &= A_{j+1}(\widehat{\beta})C_j(\widehat{\beta}) - \frac{1}{4}B_{j+1}B_j \\
&= h^2\left(j - \frac{1}{2} + \frac{h}{4}\right)\left(j - \frac{1}{2} + \frac{h}{4}\right) - \frac{1}{4}B_{j+1}B_j \\
&= h^2\left(j - \frac{1}{2} + \frac{h}{4}\right)\left(j - \frac{1}{2} + \frac{h}{4}\right) - \left(jh + \alpha\frac{h^2}{2}\right)\left((j-1)h + \alpha\frac{h^2}{2}\right) \\
&= h^2\left(j - \frac{1}{2} + \frac{h}{4}\right)^2 - h^2\left(j + \alpha\frac{h}{2}\right)\left((j-1) + \alpha\frac{h}{2}\right) \\
&= h^2\left(\frac{h}{2}\left(\frac{1}{2} - \alpha\right)j + \frac{h^2}{16} - \frac{1}{4} - \frac{\alpha h}{2}\left(\frac{\alpha h}{2} - 1\right)\right) \\
&= h^2\left(\left(\frac{h^2}{16} + \frac{h}{4} - \frac{1}{4}\right) - \alpha\frac{h}{2}(j-1) - \left(\alpha\frac{h}{2}\right)^2\right).
\end{aligned}$$

Let us set $\widehat{\alpha} := \alpha h/2$. Hence we have shown so far that,

$$p(\beta) \leq h^2\left(\left(\frac{h^2}{16} + \frac{h}{4} - \frac{1}{4}\right) - \widehat{\alpha}(j-1) - \widehat{\alpha}^2\right).$$

But from this it follows, since $\widehat{\alpha} > 0$ and $j > 1$, that $p(\beta) \leq 0$ for all β, say when, $h \leq \frac{1}{2}$. Thus, $\Delta_j \geq 0$ for all such j. □

Since $\mathcal{M}$ is invertible, by combining Lemmas A.3.1 and A.3.3, then we have a unique solution to the finite-difference scheme proposed. In particular this illustrates numerically how $u(0) \approx u_1$ is uniquely implied by the *smooth* approximation to the *differential equation* together with its *boundary condition* at $y = L$.

A.4 Numerical results for the Cox-Ingersoll-Ross boundary value problem

Suppose $L = 2, \tilde{u}_L = 1, \alpha = 0.05$ and $\beta = 1.25$. We solved Problem A.2 analytically and by the finite-difference method using 4, 8, 16, 32, and 64 nodes on the interval $[0, L]$. The results are presented in Table A.1. Figure A.1 shows both the analytical solution and the finite-difference solution with 16 points.

Points	Error	Time [ms]
4	0.009460	1.085620
8	0.001988	1.579736
16	0.000449	2.914346
32	0.000106	5.569578
64	0.000025	10.93454

Table A.1: Numerical results for the Cox-Ingersoll-Ross boundary value problem when $\beta = 1.25$.

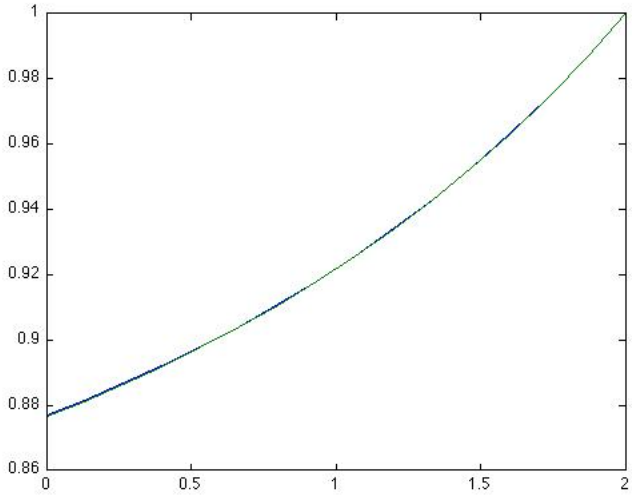


Figure A.1: Analytical and approximate solution using the finite-difference method with 16 points when $\beta = 1.25$.

Clearly, as we subdivide the interval $[0, L]$ with more and more nodes, the average time it takes the finite-difference scheme to converge increases but the maximum error with respect to the analytical solution decreases.

Similarly, for $\beta = 0.75$, we solve Problem A.2 and present the results in Table A.2

and Figure A.2. As expected, the average times do not change, however the maximum error decreases at a smaller rate.

Points	Error	Time [ms]
4	0.023140	1.168104
8	0.006263	1.590480
16	0.001787	2.918295
32	0.000523	5.547732
64	0.000155	10.90285

Table A.2: Numerical results for the Cox-Ingersoll-Ross boundary value problem when $\beta = 0.75$.

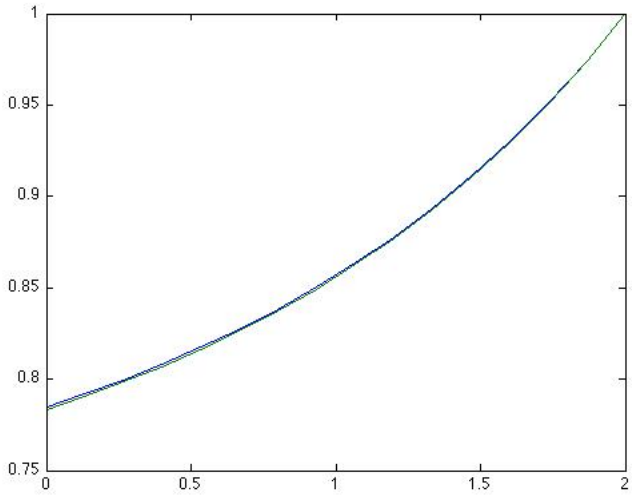


Figure A.2: Analytical and approximate solution using the finite-difference method with 16 points when $\beta = 0.75$.

This finite-difference scheme can be easily generalized to numerically solve both the boundary value problem and the obstacle problem for the elliptic Heston operator, which we do in Chapters 6 and 7.

References

[1] M. Abramovitz and I. A. Stegun, *Handbook of mathematical functions*, Dover, New York, 1972.

[2] R. A. Adams, *Sobolev spaces*, Academic Press, Orlando, FL, 1975.

[3] G. Barles, *Fully nonlinear neumann type boundary conditions for second-order elliptic and parabolic equations*, J. Differential Equations **106** (1993), 90–106.

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

[5] A. Bensoussan and J. L. Lions, *Applications of variational inequalities in stochastic control*, North-Holland, New York, 1982.

[6] F. Black and M. Scholes, *The pricing of options and corporate liabilities*, J. Political Economy **81** (1973), 637–654.

[7] L. Brugnano and D. Trigiante, *Tridiagonal Matrices: Invertibility and conditioning*, Linear Algebra and its Applications **166** (1992), 131–150.

[8] L. A. Caffarelli, *The obstacle problem revisited*, J. Fourier Anal. Appl. **4** (1998), 383–402.

[9] P.G. Ciarlet, *Discrete maximum principle for finite-difference operators*, Aequationes Math **4** (1970), 338–352.

[10] J. Cox, J. Ingersoll, and S. Ross, *A theory of the term structure of interest rates*, Econometrica **53** (1985), 385–407.

[11] P. Daskalopoulos and P. M. N. Feehan, *$c^{1,1}$ regularity for degenerate elliptic obstacle problems in mathematical finance*, `http://arxiv.org/abs/1206.0831`.

[12] ———, *Existence, uniqueness, and global regularity for variational inequalities and obstacle problems for degenerate elliptic partial differential operators in mathematical finance*, `arxiv.org/abs/1109.1075v1`.

[13] P. Daskalopoulos and R. Hamilton, *C^∞-regularity of the free boundary for the porous medium equation*, J. Amer. Math. Soc. **11** (1998), 899–965.

[14] L. C. Evans, *Partial differential equations*, American Mathematical Society, Providence, RI, 1998.

[15] Richard E. Falk, *Error estimates for the approximation of a class of variational inequalities*, Mathematics of Computation **28** (1974), 963–971.

[16] P. M. N. Feehan, *A classical Perron method for existence of smooth solutions to boundary value and obstacle problems for degenerate-elliptic operators via holomorphic maps*, http://arxiv.org/abs/1302.1849.

[17] ———, *Partial differential operators with non-negative characteristic form, maximum principles, and uniqueness for boundary value and obstacle problems*, http://arxiv.org/abs/1204.6613v1.

[18] ———, *A schauder approach to degenerate-parabolic partial differential equations with unbounded coefficients*, Journal of Differential Equations (2013), http://arxiv.org/abs/1112.4824.

[19] P. M. N. Feehan and C. A. Pop, *Degenerate elliptic operators in mathematical finance and hölder continuity for solutions to variational equations and inequalities*, arxiv.org/abs/1110.5594.

[20] ———, *Higher-order regularity for solutions to degenerate elliptic variational equations in mathematical finance*, arxiv.org/abs/1208.2658.

[21] ———, *Schauder a priori estimates and regularity of solutions to degenerate-elliptic linear second-order partial differential equations*, arxiv.org/abs/1210.6727.

[22] ———, *Stochastic representation of solutions to degenerate elliptic and parabolic boundary value and obstacle problems with dirichlet boundary conditions*, http://arxiv.org/abs/1204.1317.

[23] A. Friedman, *Variational principles and free boundary problems*, Wiley, New York, 1982, reprinted by Dover, New York, 2010.

[24] D. Gilbarg and N. Trudinger, *Elliptic partial differential equations of second order*, second ed., Springer, New York, 1983.

[25] R. Glowinski, *Numerical methods for nonlinear variational problems*, Springer, Berlin, 2008.

[26] R. Glowinski, J-L. Lions, and R. Trémolières, *Numerical analysis of variational inequalities*, North-Holland, Amsterdam, 1981.

[27] S. Heston, *A closed-form solution for options with stochastic volatility with applications to bond and currency options*, Review of Financial Studies **6** (1993), 327–343.

[28] T. Kluge, *Pricing derivatives in stochastic volatility models using the finite difference method*, Ph.D. thesis, Technische Universität Chemnitz, Fakultät für Mathematik, 2002, http://kluge.in-chemnitz.de/documents/diploma/.

[29] H. Koch, *Non-Euclidean singular integrals and the porous medium equation*, Habilitation Thesis, University of Heidelberg, 1999, www.mathematik.uni-dortmund.de/lsi/koch/publications.html.

[30] N.V. Krylov, *Lectures on elliptic and parabolic equations in Hölder spaces*, American Mathematical Society, Providence, RI, 1996.

[31] A. Kufner, *Weighted Sobolev spaces*, Wiley, New York, 1985.

[32] B. Øksendal, *Stochastic differential equations*, sixth ed., Springer, Berlin, 2003.

[33] James M. Ortega, *Numerical analysis: A second course*, Society for Industrial and Applied Mathematics, Philadelphia, 1987.

[34] Thomson Reuters, *Web of knowledge*, Internet, `wokinfo.com`.

[35] Larry L. Schumaker, *Spline functions: Basic theory*, Cambridge University Press, New York, 2007.