

# Boundary Treatment and Multigrid Preconditioning for Semi-Lagrangian Schemes Applied to Hamilton-Jacobi-Bellman Equations



Julen Rotaetxe Arto  
St Hugh's College  
University of Oxford

A thesis submitted for the degree of  
*Doctor of Philosophy*  
Trinity 2016



A mis padres Santiago y Ana María



## Acknowledgements

I am very grateful to Prof. Christoph Reisinger for giving me the opportunity to continue the work started in the MSc thesis and develop this to a DPhil under his supervision. His patience, support, availability and guidance were utterly invaluable, and essential to this project.

I am especially thankful and happy that Dr. Athena Picarelli agreed to join the project and act as a co-supervisor of this thesis. Working with her made the last months of the project really enjoyable and fun.

Throughout my five years in Oxford I was very fortunate to have the continued support of family and friends. In particular, European football nights in St Hugh's MCR served as a good way to leave research behind for a few hours. Catching up with friends from *Teleko*, Amama's family meals, and the holiday in Conil my brother organised all gave welcome respite. I was also very lucky to meet Jane during my first year at Oxford; her love and encouragement made the DPhil much easier.

*Por último, pero no por ello menos importante, quisiera expresar mi más sentido agradecimiento a mis padres Santiago y Ana María por su apoyo incondicional y su dedicación al bienestar de sus hijos. Sin su generosa ayuda nada de esto hubiera sido posible.*



# Boundary Treatment and Multigrid Preconditioning for Semi-Lagrangian Schemes Applied to Hamilton-Jacobi-Bellman Equations

Julen Rotaetxe Arto

St Hugh's College  
University of Oxford

*A thesis submitted for the degree of  
Doctor of Philosophy*

Trinity 2016

We analyse two practical aspects that arise in the numerical solution of Hamilton-Jacobi-Bellman (HJB) equations by a particular class of monotone approximation schemes known as semi-Lagrangian schemes, namely boundary treatment and multigrid preconditioning. These schemes make use of a wide stencil to achieve convergence and result in discretization matrices that are less sparse and less local than those coming from standard finite difference schemes. This leads to computational difficulties not encountered there.

We start by considering the overstepping of the domain boundary and analyse the accuracy and stability of stencil truncation. This truncation imposes a stricter CFL condition for explicit schemes in the vicinity of boundaries than in the interior, such that implicit schemes become attractive. Then, we show that for problems posed on a (semi) infinite domain whose boundary has no regular points we can avoid such truncation by means of a smooth transformation of the domain.

Next, we consider the error analysis for semi-Lagrangian schemes with truncated stencils. The stencil truncation alters properties of the semi-Lagrangian scheme that were used for the derivation of error bounds. Hence, using an alternative approach, relying on a regularization procedure due to Krylov and a switching system approximation to the HJB equation, we derive error bounds for the truncated scheme.

Finally, motivated by the stricter CFL condition of the scheme with truncated stencils, we consider implicit time stepping schemes. This involves solving non-linear systems of algebraic equations for semi-Lagrangian discretization matrices, hence, we study the use of geometric, algebraic and aggregation-based multigrid preconditioners to construct efficient solvers.





# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Contributions and outline . . . . .	13
<b>2</b>	<b>Boundary treatment for monotone SL schemes</b>	<b>17</b>
2.1	Results on Dirichlet boundary conditions for HJB equations . . . . .	20
2.2	Truncation of the LISL scheme . . . . .	23
2.2.1	Definition of truncated stencils . . . . .	25
2.2.2	Consistency conditions . . . . .	27
2.2.3	Properties of the truncated stencil . . . . .	36
2.2.4	Numerical experiments . . . . .	39
2.3	Domain transformations for the Black-Scholes equation . . . . .	51
2.3.1	Boundary regularity for the Black-Scholes equation . . . . .	53
2.3.2	Boundary regularity for a transformed Black-Scholes equation . . . . .	54
2.3.3	Numerical experiment . . . . .	61
2.3.4	Extension to the non-linear case . . . . .	64
2.4	Conclusion . . . . .	65
<b>3</b>	<b>Error bounds for monotone schemes for the Cauchy-Dirichlet problem for HJB equations</b>	<b>67</b>
3.1	Introduction . . . . .	67

3.2	Definitions and results on viscosity solutions . . . . .	73
3.3	Background results for switching systems . . . . .	78
3.4	Convergence rate for a switching system . . . . .	93
3.5	Error bounds for discretizations of the Cauchy-Dirichlet problem on bounded domains . . . . .	102
3.5.1	Upper bound by Krylov regularization . . . . .	103
3.5.2	Lower bound by switching system approximation . . . . .	107
3.6	Error bounds for some monotone finite difference schemes . . . . .	113
3.6.1	The scheme by Kushner and Dupuis . . . . .	113
3.6.2	The truncated semi-Lagrangian scheme . . . . .	115
3.6.3	The truncated semi-Lagrangian scheme with local refinement . . . . .	119
3.7	Conclusion . . . . .	124
<b>4</b>	<b>Multigrid preconditioning</b>	<b>125</b>
4.1	Introduction . . . . .	125
4.2	On the spectrum of LISL matrices . . . . .	128
4.3	Local Fourier analysis of the smoothers . . . . .	132
4.4	Performance of geometric multigrid . . . . .	136
4.5	Properties of the LISL matrix . . . . .	140
4.6	Performance of the algebraic approaches . . . . .	143
<b>5</b>	<b>Conclusion</b>	<b>149</b>
	<b>Bibliography</b>	<b>152</b>

# List of Figures

2.2.1 Truncation and extrapolation of the stencil for an elliptical domain and a mesh made of square cells. The modified stencil samples the domain boundary. . . . .	26
2.2.2 Graphical representation of the stencil over a two-dimensional Cartesian grid of size $11 \times 11$ and 10 equally spaced points from the control set $\mathcal{A}$ . The finite difference weights corresponding to some of the points are printed, for simplicity the weights are labelled $A \equiv A_{2,1}^\alpha(x)$ , $B \equiv B_{2,1}^\alpha(x)$ and $C \equiv (\mu_{2,2}^\alpha(x))^{-1}$ , following the notation in (2.2.7) and (2.2.6). To illustrate the non-locality of the scheme as the grid is refined, the second row represents the histograms of the shortest displacement from the central node for a grid of size $641 \times 641$ for both problems. The radius of the stencil in $\sigma^\alpha$ is 14.27 for this grid, given by $\frac{\ \sigma^\alpha\ _2}{\sqrt{\Delta x}} = \sqrt{640/\pi}$ . . . . .	41

2.3.1	Graphical representation of the semi-Lagrangian stencil for the two dimensional Black-Scholes equations over a Cartesian grid of size $11 \times 11$ for the domain $[0, 50]^2$ . The parameters of the model are fixed as follows: $r = 0.025$ , $q_1 = q_2 = 0$ , $\sigma_1 = 0.3$ , $\sigma_2 = 0.1$ , $\rho = 0.3$ . The lines linking the stencil nodes to the central node are colour coded according to the colours in equations (2.3.2)–(2.3.3), i.e. the drift is represented in black, $\Sigma_1$ in blue and $\Sigma_2$ in red. For $\Sigma_2$ we also represent the nodes and the weights (divided by 2) of the truncated stencil according to (2.2.7).	55
2.3.2	Sparsity plot for the discretizations matrices using the a LISL scheme (left) and a standard fixed stencil scheme (right). The rows are represented vertically and the columns horizontally. The top left corner corresponds to the first row and first column. A blue point appears for each non-zero coefficient in the matrix. The red lines correspond to nodes whose stencil oversteps the mesh.	56
2.3.3	Sparsity plots for the discretization matrices from the LISL-scheme for the linear Black-Scholes equation in the original domain and in the transformed domain. The axes of the graph denote the rows and columns of the discretization matrices. The points in blue represent non-zero coefficients of the discretization matrices, and in red nodes whose stencil oversteps. In the case of the transformed domain, two matrices for different volatility factors have been plotted, this is to highlight that the tridiagonal matrix is only obtained for small $\sigma$ .	59

2.3.4	Errors in the $L^\infty$ norm for the interval $(S_1, S_2) \in [0, 200]^2$ for the price of a call option on the maximum of two assets. We denote by <i>Exp</i> and <i>Imp</i> whether the time-stepping is explicit or implicit. <i>Trunc</i> refers to the LISL method with truncated stencil introduced in Section 2.2. <i>CLS</i> refers to the classical central space finite difference discretization as in [76]. <i>Trans</i> refers to the LISL scheme in [26] applied to the transformed equation (2.3.8). The lines labelled $\mathcal{O}(\Delta x)$ and $\mathcal{O}(\sqrt{\Delta x})$ are straight lines with slopes equal to -1 and $-\frac{1}{2}$ and shown for readability of the plots. . . . .	63
3.6.1	Locally refined grid according to the diffusion stencil shown in red. The regions in Definition 3.6.1 are shown in different colors. Brown for $\Omega_{\Delta x}^{(1)}$ , green for $\Omega_{\Delta x}^{(2)}$ and black for $\Omega_{\Delta x}^{(3)}$ . In red we plot the stencil for one node per region. . . . .	121
4.2.1	Eigenvalues of $L_{SL}^{N,m,\gamma}$ , $L_N^m$ , $L_N^{m+1}$ and $L_N$ with parameter values $N = 31$ , $m = 5$ and $\gamma = 0.346$ and the eigenvectors corresponding to three eigenvalues of the same matrices. . . . .	131
4.3.1	Representation of the smoothing factor for high frequencies, i.e. $\theta \in [-\pi, \pi]^2 \setminus [-\pi/2, \pi/2]^2$ , for the Gauss-Seidel iteration for the classical fixed stencil Finite Difference (FD) and the LISL schemes of the two-dimensional Laplacian operator. The maxima calculated numerically are 0.49 (theoretical value is 0.5) for the fixed stencil FD and 0.95 for the LISL scheme (lower is better). . . . .	135

4.4.1 Residual $\ b - Ax^k\ _2$ in the Euclidean norm at the end of the $k$ -th iteration of different geometric and algebraic multigrid cycles when solving (4.2.1) on equispaced Cartesian grid of $[0, 1]^2$ with 257 nodes per dimension and with homogeneous Dirichlet boundary conditions. Geometric $V(\nu_1, \nu_2)$ and $W(\nu_1, \nu_2)$ cycles are considered, where $\nu_1$ and $\nu_2$ denote the number of pre- and post-smoothing steps. Their performance is compared to the iterative method BICGSTAB with and without preconditioner, and to two algebraic algorithms, AMG and AGMG from [69] and [63], respectively (see also Sections 4.5 and 4.6). Notice the almost overlapping of lines for geometric $V(\nu_1, \nu_2)$ and $W(\nu_1, \nu_2)$ cycles for equal $\nu_1$ and $\nu_2$ (see also Table 4.4.1, which shows almost identical rates for $\ell = 8$ ). . . . .	137
4.6.1 Total number of seconds for solving the linear systems versus the size of the systems for each of the linear system solvers considered. We use equispaced Cartesian grids in space with 81, 161, 321 and 641 nodes per dimension and one time step. . . . .	144
4.6.2 Average number of seconds per time step for solving the linear systems versus the size of the systems. We use equispaced Cartesian grids in space with 81, 161, 321 and 641 nodes per dimension and $\Delta t = \Delta x$ . . .	145

# List of Tables

2.2.1 Results using the truncation of the stencil for explicit method with $N_\alpha = 40$ for Problem A. . . . .	44
2.2.2 Results using constant extrapolation of the boundary condition for explicit method with $N_\alpha = 40$ for Problem A. . . . .	44
2.2.3 Results using linear extrapolation for points out of the domain for explicit method with $N_\alpha = 40$ for Problem A. . . . .	45
2.2.4 Results using truncation for points out of the domain for implicit method with $N_\alpha = 40$ for Problem A. . . . .	45
2.2.5 Results using stencil truncation for explicit method with $N_\alpha = 40$ for Problem B. . . . .	46
2.2.6 Results using constant extrapolation of the boundary condition for explicit method with $N_\alpha = 40$ for Problem B. . . . .	46
2.2.7 Results using linear extrapolation for points out of the domain for explicit method with $N_\alpha = 40$ for Problem B. . . . .	47
2.2.8 Results using truncation for points out of the domain for implicit method with $N_\alpha = 40$ for Problem B. . . . .	47
2.2.9 Results using the truncation of the stencil for explicit method with $N_\alpha = 40$ for Problem A on shifted domain. . . . .	48
2.2.10 Results using the truncation of the stencil for explicit method for Problem C. . . . .	50

2.2.1 Results using the truncation of the stencil for implicit method for Problem C. . . . .	51
4.4.1 The residual reduction factor $\rho$ for different mesh sizes and different multigrid algorithms, for the two-dimensional Laplace equation; the length of the stencil $m$ as per (4.2.2). . . . .	138
4.4.2 Comparison of the residual reduction factor $\rho$ for different system sizes and different solvers for the one dimensional Laplace equation. The system size is $2^\ell + 1$ . . . . .	139
4.4.3 Comparison of the grid and algebraic complexities as per Definitions 4.1.1 and 4.1.2 for different mesh sizes and different multigrid algorithms, for the two-dimensional case. . . . .	139
4.6.1 Average seconds per time step solving linear systems. . . . .	145
4.6.2 Percentage of computational time spent in linear solvers for the Examples in Section 4.6. . . . .	146
4.6.3 Peak memory consumption statistics in gigabytes (GB) of the MATLAB process sampled using the shell command <code>top</code> . VIRT is the total amount of virtual memory used by MATLAB, whereas RES is the non-swapped physical memory (limited to 7.5). . . . .	147



4.6.4 Quantities related to the Krylov subspace iteration and multigrid coarsening. <i>Avg Krylov It</i> contains the average number of Krylov iterations over all time steps and all policy iterations; <i># levels</i> contains the average depth in the grid hierarchy; <i>C/F stencil</i> contains the ratio between the stencil at the coarsest level and that on the finest level (lower is better). On the finest level, the stencil is close to 11 for Problem A and close to 8 for Problem B. The last two columns report the grid and algebraic complexity as per Definitions 4.1.1 and 4.1.2. As the full matrix hierarchy was not available from [44] for AMG, but only the coarsest and finest matrices, the starred algebraic complexities are estimates based on the assumption of a geometrically decreasing complexity between the coarsest and finest level, which is likely to be a significant underestimate. . . . .	148
---	-----

## List of Notations

$[[a, b]]$	for $a, b \in \mathbb{R}$ denotes $[a, b] \cap \mathbb{Z}$
$\mathcal{A}$	a compact set usually representing the control set of a second order non-linear Hamilton–Jacobi–Bellman equation
$A_p^\alpha$	finite difference weight of the truncated semi-Lagrangian scheme associated to the stencil $\hat{y}_p^{\alpha,+}$
$B_p^\alpha$	finite difference weight of the truncated semi-Lagrangian scheme associated to the stencil $\hat{y}_p^{\alpha,-}$
$C^0(Q)$	the space of bounded continuous functions in $Q$
$C^k(Q)$	for $k \geq 1$ , the space of continuous and $k$ -times differentiable functions in $Q$
$\partial\Omega$	the boundary of $\Omega \subset \mathbb{R}^d$
$\mathcal{C}^{0,\delta}$	for any $\delta \in (0, 1]$ , the subset of $C^0$ with finite $ \cdot _\delta$ norm
$\partial^*Q_T$	for a domain $Q_T := (0, T] \times \Omega \subset \mathbb{R}^{d+1}$ denotes the parabolic boundary, i.e. $\partial^*Q_T := (\{0\} \times \overline{\Omega}) \cup ((0, T] \times \partial\Omega)$
$\Delta x$	grid refinement parameter for the spatial domain of a PDE
$\Delta t$	grid refinement parameter for the time-line of a PDE
$\mathcal{G}_h$	a time-space discretization of a space time domain $Q_T$ with mesh refinement parameter $h$
$\mathcal{I}_{\Delta x}$	interpolation operator defined on a grid with grid refining parameter $\Delta x > 0$
$h$	space-time mesh refinement parameter $h = (\Delta t, \Delta x)$
$\text{LSC}(Q; \mathbb{R}^d)$	the space of lower semicontinuous functions from $Q$ to $\mathbb{R}^d$
$\mathcal{M}_i r$	applied to a vector $r$ , $\mathcal{M}_i r := \min_{j \neq i} r_j + k$ , where $k > 0$

$\mu_p^{\alpha,\pm}$	positive scaling parameter associated to the semi-Lagrangian the stencil $y_p^{\alpha,\pm}$
$\mathcal{N}(x)$	in relation to an interpolation operator $\mathcal{I}_{\Delta x}$ , the set of interpolation nodes for the point $x$
$\omega(z)$	continuous and positive function $\omega : [0, \infty) \times [0, \infty)$ such that $\omega(0^+) = 0$
$\Omega$	subset of $\mathbb{R}^d$ , the spatial domain of a PDE
$\overline{\Omega}$	for $\Omega \subset \mathbb{R}^d$ , the closure of the set $\Omega$
$\Omega_{\Delta x}$	possibly unstructured discretization of a spatial domain $\Omega$ with refinement parameter $\Delta x$
$ \phi _0$	for a bounded function $\phi : Q_T \rightarrow \mathbb{R}^d$ , $ \phi _0 := \sup_{(t,y) \in Q}  \phi(t, y) $
$[\phi]_\delta$	for any $\delta \in (0, 1]$ and $\phi : Q_T \rightarrow \mathbb{R}^d$ , $[\phi]_\delta := \sup_{(t,x) \neq (s,y)} \frac{ \phi(t,x) - \phi(s,y) }{( x-y  +  t-s ^{1/2})^\delta}$
$ \phi _\delta$	for any $\delta \in (0, 1]$ and $\phi : Q_T \rightarrow \mathbb{R}^d$ , $ \phi _\delta :=  \phi _0 + [\phi]_\delta$
$\phi^*$	for a locally bounded function $\phi : Q \rightarrow \mathbb{R}^d$ , it denotes its upper-semicontinuous envelope $\phi^*(x) := \limsup_{y \rightarrow x, y \in Q} \phi(y)$
$\phi_*$	for a locally bounded function $\phi : Q \rightarrow \mathbb{R}^d$ , it denotes its lower-semicontinuous envelope $\phi_*(x) = \liminf_{y \rightarrow x, y \in Q} \phi(y)$
$Q_T$	a time space domain $Q_T := (0, T] \times \Omega \subset \mathbb{R}^{d+1}$ for any $T > 0$
$Q_T^\varepsilon$	associated to a space time domain $Q_T$ , $Q_T^\varepsilon := (\varepsilon^2, T] \times \Omega$ for any $\varepsilon, T > 0$
$\mathcal{P}^{2,+}, \mathcal{P}^{2,-}$	second-order parabolic superjet and subjet respectively
$\rho_\varepsilon$	family of mollifiers with parameter $\varepsilon > 0$

$\mathcal{S}^d$	the space of $d \times d$ real symmetric matrices
$u_h$	if $u$ is a solution to the PDE, then $u_h$ is a numerical approximation to $u$
$U_i^n$	applied to a matrix $U \in \mathbb{R}^{N \times I}$ , denotes the element of $U$ with indexes $n$ and $i$
$\text{USC}(Q; \mathbb{R}^d)$	the space of upper semicontinuous functions from $Q$ to $\mathbb{R}^d$
$[U]_{n,i}$	applied to a matrix $U \in \mathbb{R}^{N \times I}$ , denotes all the elements of $U$ except $U_{n,i}$
$y_p^{\alpha, \pm}$	stencil for the original semi-Lagrangian scheme
$\hat{y}_p^{\alpha, \pm}$	stencil for the truncated semi-Lagrangian scheme

# Chapter 1

## Introduction

### 1.1 Motivation

We consider the Hamilton-Jacobi-Bellman (HJB) equation

$$u_t(t, x) + \sup_{\alpha \in \mathcal{A}} \{-L^\alpha[u](t, x) - c^\alpha(t, x)u(t, x) - f^\alpha(t, x)\} = 0, \quad (t, x) \in (0, T] \times \Omega, \quad (1.1.1)$$

$$u(0, x) = \psi(0, x), \quad x \in \bar{\Omega}, \quad (1.1.2)$$

$$u(t, x) = \psi(t, x), \quad (t, x) \in (0, T] \times \partial\Omega, \quad (1.1.3)$$

where  $Q_T := (0, T] \times \bar{\Omega}$  with  $\bar{\Omega} := \Omega \cup \partial\Omega \subseteq \mathbb{R}^d$ ,  $\mathcal{A}$  is a compact set,

$$L^\alpha[u](t, x) = \text{tr}[a^\alpha(t, x)D^2u(t, x)] + b^\alpha(t, x)Du(t, x), \quad (1.1.4)$$

is a second order differential operator, and the function  $\psi$  contains the initial and spatial boundary conditions.

The coefficients  $a^\alpha = \frac{1}{2}\sigma^\alpha\sigma^{\alpha,T}$ ,  $b^\alpha$ ,  $c^\alpha$ ,  $f^\alpha$ , and the data  $\psi$  take their values, respectively, in  $\mathcal{S}^d$ , the space of  $d \times d$  symmetric matrices,  $\mathbb{R}^d$ ,  $\mathbb{R}$ ,  $\mathbb{R}$ , and  $\mathbb{R}$ ,  $\sigma^\alpha \in$

$\mathbb{R}^{d \times P}$ , such that  $a^\alpha$  is positive semi-definite. We also assume the usual well-posedness conditions on the PDE coefficients, i.e. Lipschitz continuity in  $x$  uniformly in  $\alpha$ , Hölder continuity with exponent  $\frac{1}{2}$  in time and continuity in  $\alpha$  for each  $(t, x) \in Q_T$  [54]. The relevant notion of solution for this type of non-linear equations is that of viscosity solutions [23] and the above conditions guarantee existence and uniqueness.

Even if we will mainly use analytic techniques to treat the HJB equation, it is worth mentioning that the problem has also a probabilistic representation, see the classical references [54, 84] for further details. In a nutshell, as pointed out in [67], the HJB equation represents the limit in time of localising the dynamic programming and therefore describes the local dynamics of the value function  $u(t, x)$ . More specifically, the value function for the exit-time control problem below is given by

$$u(T - t, x) = \inf_{\alpha(\cdot) \in \mathcal{A}} \mathbb{E} \left[ \int_t^\tau \Gamma(t, s) f^{\alpha(s)}(s, X_s^{t,x}) ds + \Gamma(t, \tau) \psi(\tau, X_\tau^{t,x}) \right], \quad (1.1.5)$$

where the state process  $X_t^{s,x}$  is the solution of the following controlled stochastic differential equation

$$X_t = x \in \Omega, \quad dX_s^{t,x} = b^{\alpha(s)}(s, X_s^{t,x}) ds + \sigma^{\alpha(s)}(s, X_s^{t,x}) dW_s \quad \text{for } s > t, \quad (1.1.6)$$

$W_s$  is a  $P$ -dimensional Brownian motion,  $\tau$  is the first exit time of the trajectory of the state process  $X_t^{s,x}$  from  $\Omega$ , and the discount factor is given by

$$\Gamma(t, s) := \exp \left( - \int_t^s c^{\alpha(r)}(r, X_r^{t,x}) dr \right).$$

Using the dynamic programming principle one can then establish that the value function in (1.1.5) is a viscosity sub- and supersolution of (1.1.1)–(1.1.3). It is precisely for this reason that we are interested in viscosity solutions to (1.1.1)–(1.1.3).

In general, the viscosity solution to (1.1.1)–(1.1.3) is unknown. Thus, in practice,

it is necessary to compute approximations numerically. Sufficient conditions for a numerical scheme to converge to the unique viscosity solution of (1.1.1)–(1.1.3) were proved by Barles and Souganidis [10] in terms of consistency,  $L^\infty$ -stability and monotonicity. For first order Hamilton–Jacobi equations, there exists a weaker formulation of the previous result replacing monotonicity by so-called  $\varepsilon$ -monotonicity, see [13, 30]. In this thesis, we restrict our attention to finite difference discretizations of the second order differential operator (1.1.4).

The requirement of monotonicity drastically affects the properties and construction of finite difference schemes. Theorem 4 in [65] proves that local monotone discretizations have consistency errors of at most first order for first-order equations and second order consistent for second-order equations. What is more, standard fixed stencil methods are able to produce monotone discretizations only under restrictions on the diffusion matrix, such as diagonal dominance [26, 35, 53]. Results from [24, 59] further illustrate the limitations of such methods for the monotone approximation of second order derivatives.

This implies that, generally, monotone approximations have to be non-local on the discrete level, i.e. the distance between mesh points involved in the scheme at a given point grows in relation to the mesh width as the mesh is refined. Such schemes are referred to as wide stencils. For general diffusion matrices, first order accurate wide stencils of the type considered here have been proposed in [19, 26, 58], and a mixed fixed- and wide-stencil scheme in [56].

The above discussion concerns methods known to converge to the viscosity solution. If stronger assumptions are made on the regularity of the solution, it is possible to construct schemes that do not suffer from the limitations imposed by monotonicity. Some of these so-called non-monotone methods are reviewed in [32], and while they may have some advantages over monotone methods, such as a higher convergence rates, major drawbacks are that they are not so generally applicable, and that

each method requires a bespoke convergence analysis. We note that the latter, in general, is not an easy task. An example of a non-monotone method with high accuracy, compact stencils and with full theoretical treatment is shown in [73, 72]. The method shows good performance on challenging problems for non-local methods, such as boundary layers and anisotropic diffusions. The analysis is done under suitable regularity conditions, i.e. Cordès conditions, to show convergence to solutions in the Sobolev space  $H^2$ .

The main motivation for the thesis is to analyse two known issues arising in practice when numerically solving (1.1.1)–(1.1.3) using the class of schemes described in [16, 19, 26, 30, 33, 58] to discretize the second order differential operator (1.1.4). This approximation combines wide stencils in the directions determined by the columns of the matrix  $\sigma^\alpha$  and the drift  $b^\alpha$ , together with monotone interpolation. Following the notation in [26], we write the matrix  $\sigma^\alpha \in \mathbb{R}^{d \times P}$  as  $(\sigma_1^\alpha, \sigma_2^\alpha, \dots, \sigma_P^\alpha)$ , where  $\sigma_p^\alpha \in \mathbb{R}^d$  for  $p \in \{1, 2, \dots, P\}$  denotes the  $p$ -th column of  $\sigma^\alpha$ , and observe that for  $k > 0$  and any smooth function  $\phi$ ,

$$\frac{1}{2} \text{tr} [\sigma^\alpha \sigma^{\alpha, T} D^2 \phi(x)] = \frac{1}{2} \sum_{p=1}^P \frac{\phi(x + k\sigma_p^\alpha) - 2\phi(x) + \phi(x - k\sigma_p^\alpha)}{k^2} + \mathcal{O}(k^2), \quad (1.1.7)$$

$$b^\alpha D\phi(x) = \frac{\phi(x + k^2 b^\alpha) - \phi(x)}{k^2} + \mathcal{O}(k^2), \quad (1.1.8)$$

where  $\mathcal{O}(k^2)$  is the local consistency error of the finite difference and for compactness we write  $b^\alpha \equiv b^\alpha(t, x)$  and  $\sigma^\alpha \equiv \sigma^\alpha(t, x)$ . As these approximations will be used for points lying on a discrete spatial grid  $\Omega_{\Delta x}$  with nodes  $\{x_j : 1 \leq j \leq N\}$ , the displaced points  $x + k^2 b^\alpha$ ,  $x \pm k\sigma_p^\alpha$  do not generally coincide with nodes of  $\Omega_{\Delta x}$ . Therefore,  $\phi$  is replaced by an interpolant  $\mathcal{I}_{\Delta x} \phi$  on that grid. It is known that these schemes can be defined on unstructured meshes. Reproducing the notation in [26], the spatial grid  $\Omega_{\Delta x} := \{x_i\}_{i \in \mathbb{N}}$  where each  $x_i$  is a vertex of a non-degenerate polyhedral subdivision



$\mathcal{T}^{\Delta x} = \{T_j^{\Delta x}\}_{j \in \mathbb{N}}$  of  $\bar{\Omega}$ . The elements of the set  $\mathcal{T}^{\Delta x}$  satisfy

$$\begin{aligned} \text{int}(T_j^{\Delta x} \cap T_i^{\Delta x}) &= \emptyset, \quad \forall i \neq j, \\ \bigcup_{j \in \mathbb{N}} T_j^{\Delta x} &= \bar{\Omega}, \\ \nu \Delta x &\leq \sup_{j \in \mathbb{N}} \{\text{diam } \mathcal{B}_{T_j^{\Delta x}}\} \leq \sup_{j \in \mathbb{N}} \{T_j^{\Delta x}\} \leq \Delta x, \end{aligned}$$

for some  $\nu \in (0, 1)$ , where  $\text{int}$  and  $\text{diam}$  are the interior and the diameter of a polyhedron, and  $\mathcal{B}_{T_j^{\Delta x}}$  is the greatest ball contained in  $T_j^{\Delta x}$ . We restrict our attention to rectangular meshes and linear interpolants, defined by the standard piecewise multilinear non-negative basis functions  $\{w_j(\cdot) : 1 \leq j \leq N\}$  associated with the mesh nodes, such that for any function  $\phi$

$$[\mathcal{I}_{\Delta x} \phi](x) = \sum_{j \in \mathcal{N}(x)} \phi(x_j) w_j(x), \quad (1.1.9)$$

for all  $x \in \Omega$ ,  $x_j \in \Omega_{\Delta x}$ , where  $w_j(x)$  is the interpolation weight, and  $\mathcal{N}(x)$  is the set of interpolation neighbours of  $x$  on the mesh  $\Omega_{\Delta x}$ , i.e. the mesh points with non-zero interpolation weight. We ignore the scenario where  $x$  lies outside the mesh (“oversteps”) for now. The resulting scheme is referred to as the Linear Interpolation Semi-Lagrangian (LISL) scheme.

It is shown in [26] that the leading order terms of the local consistency error for a smooth function are proportional to  $k^2$  and  $\frac{\Delta x^2}{k^2}$ . The first term is the consistency error for the finite difference approximation of the first and second order derivatives in (1.1.7) and (1.1.8), whereas the second term corresponds to the linear interpolation error when replacing  $\phi$  by its interpolant in the finite difference formulae (1.1.7) and (1.1.8). Therefore, by choosing  $k = \sqrt{\Delta x}$ , the resulting scheme has local consistency error proportional to  $\Delta x$ .

Following the notation in [26], the LISL finite difference approximations for the

differential operator in (1.1.4) can be expressed as

$$L_{\Delta x}^{\alpha}[\mathcal{I}_{\Delta x}\phi](t, x) := \sum_{p=1}^M \frac{[\mathcal{I}_{\Delta x}\phi](t, x + y_p^{\alpha,+}(t, x)) - 2[\mathcal{I}_{\Delta x}\phi](t, x) + [\mathcal{I}_{\Delta x}\phi](t, x + y_p^{\alpha,-}(t, x))}{2\Delta x}, \quad (1.1.10)$$

for  $x \in \Omega_{\Delta x}$ , and some  $M \geq 1$ . The functions  $y_p^{\alpha,\pm}(t, x)$  determine the stencil of the scheme at  $(t, x)$ .

Different schemes can be obtained depending on the values taken by  $M$  and  $y_p^{\alpha,\pm}(t, x)$ . In particular, [26] discusses the following three schemes for second order operators:

#### Examples of LISL schemes.

1. **Scheme 1:** The approximation of Camilli and Falcone [19], corresponding to  $y_p^{\alpha,\pm} = \pm\sqrt{\Delta x}\sigma_p^{\alpha} + \frac{\Delta x}{P}b^{\alpha}$  and  $M = P$ .
2. **Scheme 2:** The approximation in [26], corresponding to  $y_p^{\alpha,\pm} = \pm\sqrt{\Delta x}\sigma_p^{\alpha}$  for  $p \leq P$ ,  $y_{P+1}^{\alpha,\pm} = \Delta x b^{\alpha}$ , and  $M = P + 1$ .
3. **Scheme 3:** A more efficient version of the Camilli-Falcone approximation, corresponding to  $y_p^{\alpha,\pm} = \pm\sqrt{\Delta x}\sigma_p^{\alpha}$  for  $p < P$ ,  $y_P^{\alpha,\pm} = \pm\sqrt{\Delta x}\sigma_P^{\alpha} + \Delta x b^{\alpha}$ , and  $M = P$ .

The authors show that this family of discretizations of (1.1.4) is consistent and monotone. Monotonicity of the scheme is fulfilled as the discrete approximation  $L_{\Delta x}^{\alpha}[\mathcal{I}_{\Delta x}\phi]$  is the composition of monotone finite differences and a monotone interpolation operation. Once discretised in space, the final scheme arises from discretising in time using the standard  $\theta$ -time stepping scheme for  $\theta \in [0, 1]$ , where  $\theta = 0$  corresponds to the explicit Euler time stepping and  $\theta = 1$  to the implicit case.

As discussed in Section 6.4 of [26], the construction of the schemes above can be interpreted as a weak approximation to the SDE (1.1.6). We reproduce the argument

for completeness. Let  $\{t_n\}_{n=0}^{N_t+1}$  with  $t_0 = 0$ ,  $t_{N_t+1} = T$ , and  $\Delta t_n := t_n - t_{n-1} \leq \Delta t$  for all  $n$  be a strictly increasing sequence of points, then for any  $n > m$

$$\tilde{X}_m = x, \quad \tilde{X}_n = \tilde{X}_{n-1} + \sigma^\alpha(t_{n-1}, \tilde{X}_{n-1})k_n\xi_n + b^\alpha(t_{n-1}, \tilde{X}_{n-1})k_n^2\eta_n, \quad n > m,$$

where  $k_n = \sqrt{(P+1)\Delta t_n}$ ,  $\alpha \in \mathcal{A}$  and  $\xi_n = (\xi_{n,1}, \dots, \xi_{n,P})^T$  and  $\eta_n$  are mutually independent sequences of i.i.d. random variables satisfying

$$\begin{aligned} \mathbb{P}((\xi_{n,1}, \dots, \xi_{n,P}, \eta_n) = \pm e_j) &= \frac{1}{2(P+1)}, \quad \text{if } j \in \{1, \dots, P\}, \\ \mathbb{P}((\xi_{n,1}, \dots, \xi_{n,P}, \eta_n) = \pm e_{P+1}) &= \frac{1}{P+1}, \end{aligned}$$

where  $e_j$  is the  $j$ -th vector of the canonical base of  $\mathbb{R}^{P+1}$ . All other values of  $(\xi_{n,1}, \dots, \xi_{n,P}, \eta_n)$  have zero probability.

The three schemes described above have a wide stencil as the length of the stencil, being proportional to the ratio  $k/\Delta x \sim 1/\sqrt{\Delta x}$ , tends to  $\infty$  as  $\Delta x \rightarrow 0$ . Hence, when applied on a bounded discrete grid, the stencil will generally exceed the domain for points close to its boundary. As discussed in [26], the overstepping may pose a problem depending on the equation and the type of boundary conditions imposed. We consider Dirichlet boundary conditions here.

Our first goal is to present and analyse a modification of the LISL scheme to deal with overstepping for problems on bounded domains with Dirichlet boundary conditions, and general drift and diffusion coefficients. We start by describing how to truncate the LISL stencil so that the truncation remains consistent and monotone. We then prove that the resulting stencil for Scheme 2 above is monotone. This is not the case for Schemes 1 and 3. We also observe that the truncation has both local and global impacts on the properties of the scheme. Locally, the modification of the scheme leads to a loss of accuracy by half an order in the consistency error,

i.e.  $\mathcal{O}(\sqrt{\Delta x})$  instead of  $\mathcal{O}(\Delta x)$ , due to the loss of symmetry of the finite difference approximation with respect to the central node. However, as the mesh points requiring truncation of the scheme are restricted to an  $\mathcal{O}(\sqrt{\Delta x})$  layer at the boundary, convergence rates close to  $\mathcal{O}(\Delta x)$  are observed empirically for the modified scheme. The truncation has a global effect in the sense that it modifies the CFL condition of explicit schemes by at least half an order, from  $\Delta t = \mathcal{O}(\Delta x)$  to  $\Delta t = \mathcal{O}(\Delta x^{3/2})$ . However, in the worst case the truncated scheme requires  $\Delta t = \mathcal{O}(\Delta x^2)$  for stability. As the empirical error is  $\mathcal{O}(\Delta t) + \mathcal{O}(\Delta x)$  for fully implicit schemes, the computationally most efficient choice is  $\Delta t \sim \Delta x$ , outside the stability region of explicit schemes. Finally, we compare the accuracy of the truncation with extrapolations of the boundary conditions by means of numerical tests for benchmark problems.

For degenerate problems posed on a (semi) infinite domain whose boundary has no regular points, we show that it is not necessary to truncate the stencil. From the stochastic control perspective this degeneracy is due to the reversibility of the underlying stochastic process, see [85]. A common strategy when numerically solving problems on unbounded domains using finite difference methods, is to truncate the domain and impose artificial boundary conditions on the new domain. When used together with semi-Lagrangian schemes this approach suffers from two undesirable effects. On the one hand, the errors introduced by the truncation of the domain, see [5, 6], and, on the other hand, the modification of the properties of the semi-Lagrangian scheme due to overstepping. We show that an alternative and effective strategy is to apply suitable smooth transformations to the unbounded domain mapping it to a bounded one, as in [85]. Furthermore, due to the reversibility of the underlying stochastic process the semi-Lagrangian scheme does not overstep the domain. Therefore, this avoids the necessity to modify the scheme and having to deal with a stricter CFL condition for explicit time-stepping schemes, and the reduction of the local consistency error as discussed previously.

Next, we consider the error analysis for the truncated semi-Lagrangian scheme, that is, estimating upper and lower bounds for the difference between the solution of (1.1.1)–(1.1.3) and an approximation computed by means of a finite difference scheme. The cornerstone of the error analysis for finite difference approximations to (1.1.1) is a regularization procedure, also known as “shaking coefficients”, introduced by Krylov in [50]. This regularization procedure allows the construction of smooth subsolutions to (1.1.1)–(1.1.3) from which we infer one of the bounds. For the remaining bound two approaches are observed in the literature. On the one hand, constructing a smooth subsolution for the scheme as in [50, 26], and, on the other hand, considering an auxiliary approximation to (1.1.1)–(1.1.3) as in [8]. As noted in [8], the first approach has the advantage of yielding better error rates, but the second one is more general. In particular, the first approach relies on continuous dependence estimates for the solution of the scheme with respect to the coefficients of the equation and the boundary data. We note that such results have been established for the viscosity solution of the initial value problem (1.1.1)–(1.1.2) with  $\Omega = \mathbb{R}^d$  by probabilistic methods in [34, 84] and by analytic methods in [47]. For the semi-Lagrangian schemes described in [19, 26] either of the approaches yield fractional convergence rates, more specifically  $\mathcal{O}(\Delta x^{1/5})$  and  $\mathcal{O}(\Delta x^{1/10})$  using the first and second approach respectively, see [27].

The truncation of the stencil modifies the scheme in such a way that the approach based on constructing smooth subsolutions for the equation and the scheme is not directly applicable. In particular, the numerical solution does not have the regularity required to show that solutions of the scheme depend continuously on the coefficients and the boundary data. Hence, we adapt the approach in [8] to the Cauchy-Dirichlet problem on bounded spatial domains. This is not straightforward. The “shaking coefficients” regularization requires defining the solution of a perturbed HJB equation outside the spatial domain  $\Omega$ . An enlargement of the domain is then necessary.

The extension of the spatial domain was used in [15] to prove error estimates for semi-Lagrangian schemes approximating HJB equations with an oblique derivative boundary condition. In our case, under suitable assumptions on the shape of the domain, we couple the “shaking coefficients” technique with a stretching of the domain in order to maintain the regularization inside the stretched domain.

Finally, driven by the worsening of the CFL condition for explicit time-stepping schemes, our second goal is the use of implicit schemes and the efficient solution of the algebraic systems of equations on LISL discretization matrices. For this we consider applying multigrid as preconditioner. Given that the coefficient matrix of the discrete system stems from finite difference discretizations of a PDE, we know that these matrices are generally ill-conditioned, in fact the condition number grows with the size of the matrix, and that, as noted in [82], the matrix may inherit some of the properties of the PDE it approximates.

The former indicates that to build efficient solvers we ought to start by considering preconditioners. Preconditioners can be defined as suitable transformations that, when applied to a problem, yield a transformed problem that is easier to solve numerically and whose solution is closely related to the original problem. In particular, it is known that ill-conditioned problems are difficult to solve numerically, so, as noted in [82], applying a preconditioner allows solving large problems in reasonable computational time. Furthermore, due to the growth of the condition number of the discretization matrix with the size of the system, we seek preconditioners yielding mesh independent performance. As a consequence, among all the different preconditioning techniques, see [82] for a review on the subject, we consider multigrid for its PDE related origins, its sound theoretical background, mesh-size independent results, and because it is known to produce efficient solvers in practice. We refer to [77] for a detailed overview on the subject.

For time-stepping schemes with parameter  $\theta \neq 0$  approximating (1.1.1), the cou-

pling of the optimal control and the coefficients imply that to find the numerical solution we must solve a non-linear system of algebraic equations like,

$$\max_{\alpha \in \mathcal{A}} (A_i^\alpha X - F_i^\alpha) = 0, \quad i = 1, \dots, N, \quad (1.1.11)$$

where  $A_i^\alpha$  is the  $i$ -th row of a matrix  $A^\alpha$  with elements  $A_{i,j}^\alpha$ ,  $i, j = 1, \dots, N$ , and control  $\alpha \in \mathcal{A}$ . In our case the coefficient matrix  $A^\alpha$  stems from wide stencil finite difference discretization of (1.1.1),  $F_i^\alpha$  is the right hand side term, and  $X = (X_i)$  is the solution vector for a given time step. The maximisation over  $\alpha$  in (1.1.11) is row-wise and usually done by linear search. By construction of the LISL scheme,  $A^\alpha$  is an M-matrix with non-negative row sum. Therefore, following results in [14], we can use policy iteration to compute  $X$ . Then, within each policy iteration, a linear system  $A_i^{\alpha_i} X = F_i^{\alpha_i}$ ,  $i = 1, \dots, N$ , with fixed control vector  $(\alpha_i)_{1 \leq i \leq N}$  has to be solved. We find that this last step is the computationally most costly part of the overall algorithm if direct linear solvers or standard iterative solvers are used. We therefore study multigrid preconditioners for this step to reduce the overall computational expense of the algorithm in a general setting. This is because, in the absence of a special structure on the control set or the equation, such as in the case of option pricing with uncertain volatility [40, p. 230], it is difficult to consider a general approach to reduce the computational expense of the second step in the iteration.

In the literature on multigrid preconditioners for HJB equations, two main approaches are observed: on the one hand, applying multigrid preconditioners directly to the non-linear problem, as in [12, 41, 43]; and on the other hand, applying multigrid preconditioners to a linearised problem, as in [3]. In particular, [12, 43] provide the first multigrid algorithms for HJB equations and prove convergence, while [41] presents a novel smoother for HJB equations based on damped value iteration [53]. These articles have in common the use of standard fixed stencil finite difference ap-

proximations and the use of a geometric structure when building the hierarchy of multigrid subspaces. We find that geometric multigrid is not well suited for matrices coming from non-local approximation schemes.

We then investigate algebraic multigrid methods. The basis for the specific algorithm we use was introduced in [62] for linear elliptic PDEs. It empirically showed that “aggregation based methods could yield robust<sup>1</sup> and convergent schemes if used as preconditioners of a Krylov method, and were part of an enhanced multigrid cycle, not simple V- or W-cycles” as considered in [75]. By enhanced multigrid cycles, the authors refer to recursive schemes in which at each coarse level the solution to the residual equation is computed using a number of Krylov subspace iterations as in [64] or with a semi-iterative method based on Chebyshev polynomials called the AMLI cycle, see Section 5.6 of [79]. The aggregates were formed using heuristic criteria following coupling in the strongest direction.

In [60] the authors introduced an aggregation-based multigrid method with guaranteed convergence rate for symmetric M-matrices with non-negative row sum. A LISL discretization matrix is only symmetric in very specific cases with limited practical interest. For non-symmetric matrices, in [63] convergence of a simplified two-grid scheme using aggregation is proved for non-singular M-matrices with non-negative row and column sums. This requirement ensures that the symmetric part of the coefficient matrix  $A$  given by  $A + A^T$  meets the assumptions in [60] and allows the use of its theoretically justified algorithms. We will derive conditions on the coefficients of the HJB equation such that this theory applies, and show empirically that aggregation-based multigrid gives roughly mesh-size independent convergence.

---

<sup>1</sup>In this context a robust method is referred to as one showing good performance for a large range of problems without changing the smoother.



## 1.2 Contributions and outline

The main contributions of this thesis are:

- to provide the first consistent and monotone scheme for HJB equations on bounded domains with Dirichlet boundary conditions,
- analysis of the method's consistency and stability,
- analysis of the convergence order for general approximation schemes on bounded domains with Dirichlet boundary conditions,
- analysis of the spectral properties and empirical study of geometric multigrid for wide-stencil schemes,
- theoretical and empirical analysis of two algebraic multigrid algorithms for matrices arising from semi-Lagrangian discretizations.

The thesis is organised as follows. Chapter 2 analyses the domain overstepping issue of wide-stencil schemes. First of all, we consider the application of the LISL scheme to the Cauchy-Dirichlet problem for HJB equations on bounded spatial domains. To deal with the domain overstepping, we propose a truncation of the stencil in the direction of the diffusion or the drift and recalculating the finite difference weights for consistency. We prove that this truncation yields a consistent and monotone scheme but that it negatively affects, on the one hand, the CFL condition for explicit time-stepping schemes, and, on the other hand, the local consistency error. Then, we consider the Black-Scholes equation and show that appropriate domain transformations prevent the scheme from overstepping the domain, avoiding its modification near the boundary.

Chapter 3 considers the error analysis of monotone numerical schemes for the Cauchy-Dirichlet problem for HJB equations on bounded domains. We start by reviewing the existing approaches in the literature to analyse the error of monotone

numerical schemes converging to viscosity solutions of non-linear equations. Next, we justify the choice of adapting the approach in [8] to the problem of interest due to the properties of the scheme presented in Chapter 2. Having a bounded spatial domain introduces additional difficulties in the error analysis. Compared to the unbounded case, additional assumptions are required to ensure the regularity of the viscosity solutions, see [46]. Further assumptions are also required on the coefficients of the equation and the spatial domain to ensure that the solution satisfied the Dirichlet boundary conditions pointwise, see [9]. Moreover, when considering bounded domains the mollification requires points outside the domain, hence the need to expand the original domain, which may in turn impose further restrictions on the solution and/or the coefficients. We demonstrate that for star shaped domains, it is possible to avoid imposing such restrictions by means of an affine transformation of the domain. Finally, we compute the error rates for three example schemes.

Chapter 4 studies the efficient solution of algebraic systems of equations with co-efficient matrices given by monotone semi-Lagrangian discretizations. We start by justifying the use of fully implicit time stepping schemes as these have nicer CFL conditions than their explicit counterpart. Indeed, as seen in Chapter 2, if the problem to solve is such that the semi-Lagrangian stencil needs truncation, then explicit schemes have a restrictive CFL condition. When applied to HJB equations, implicit time stepping schemes result in non-linear systems of algebraic equations due to the coupling of the optimal control and the equation coefficients. To solve these systems, we employ policy iteration [14]. Within each policy iteration, first a linear system of algebraic equations with fixed control vector has to be solved, followed by a search of the optimal control. We study the use of multigrid preconditioners to efficiently solve the linear systems in the first step of the policy iteration. We start by considering standard geometric multigrid cycles and illustrate the necessity to use multigrid cycles based on algebraic ideas. Next, we argue that aggregation based methods are well

suited for this kind of discretizations, and provide theoretical justification. Finally, we compare aggregation based multigrid (AGMG) against the classical algebraic multigrid (AMG) and show that, for the examples considered, AGMG outperforms AMG according to complexity and scalability benchmarks.

The final chapter, Chapter 5, concludes by summarising the main results of the thesis and suggesting possible directions for future research.

We note that Chapter 2 and 4 have been published in [68].



## Chapter 2

# Boundary treatment for monotone SL schemes

In Section 6.2 of [26] the authors state the following regarding the known issue of the Linear Interpolation Semi-Lagrangian (LISL) scheme overstepping the boundary of the domain:

1. “For Dirichlet boundary conditions the proposed approach is either a modification of the scheme near the boundary or an extrapolation of the boundary conditions. This may result in a loss of accuracy or monotonicity near the boundary.
2. Homogeneous Neumann boundary conditions can be implemented exactly by extending to the exterior in the normal direction the values of the solution on the boundary to the exterior.
3. If the boundary has no regular points, no boundary conditions can be imposed. In this case the SL schemes will not leave the domain if the normal diffusion tends to zero fast enough when the boundary is approached. Typical examples are equations of Black-Scholes type.”

However, no further details are given on how to modify the scheme in the first situation, or in the Black-Scholes case, how to deal with the fact that the domain is unbounded. The author in [1] discusses discretizations of first order Hamilton-Jacobi equations. However, it is only when considering the second order case that we encounter the overstepping problem. For instance, [16] only considers problems with periodic boundary conditions, [2] considers the non-linear Neumann and oblique boundary conditions. To account for the overstepping, the authors in [2] apply the scheme on a strictly interior region and the boundary conditions on a layer close to the boundary.

The loss of consistency for points near the boundary was indicated in [17, Section 7] for a different wide stencil for a two-dimensional HJB equation. In a nutshell, the scheme in [17] approximates the diffusion matrix  $a^\alpha$  by other matrices  $\tilde{a}_k^\alpha$  that are “easier” to approximate monotonically using finite difference discretizations. The parameter  $k$  in  $\tilde{a}_k^\alpha$  represents the length of the stencil and, as in the semi-Lagrangian case, the consistency error is inversely proportional to it. For points near the boundary the maximum value of  $k$  is limited by the distance to the boundary. This constraint limits the quality of the approximation to the diffusion matrix and, as a consequence, it may result in a loss of consistency.

Other authors have proposed modifications of wide stencil schemes near the boundary of the domain, see [56]. The scheme in [56] is a hybrid between local and wide-stencil, i.e. non-local, schemes. The main idea is to use a local stencil whenever the coefficients satisfy a positivity condition, see [80], and use a wide-stencil scheme otherwise. The resulting scheme is proved to be unconditionally monotone for a non-linear version of the two dimensional Black-Scholes equation. However, the unconditional monotonicity of the scheme is specific to this problem. This is due to the implicit assumption that if a stencil of length proportional to  $\mathcal{O}(\sqrt{\Delta x})$  oversteps, then one of length  $\mathcal{O}(\Delta x)$  will not overstep.

The positivity conditions in [80], reflect that fixed stencil discretizations are monotone if the diffusion matrix is diagonally dominant, see [35, 53, 80]. Thus, naively replacing the wide-stencil scheme for a fixed stencil one near the boundary, may result in a loss of monotonicity. The loss of monotonicity and consistency is clearly undesirable, as it is under these conditions that the scheme is known to converge to the unique viscosity solution under the theory of Barles and Souganidis [10].

Regarding the situation in item 3. above, in practice, even PDEs defined on (semi) infinite domains with no regular boundary points are subject to the domain overstepping issue. Indeed, in practice it is common to localise (semi) infinite domains into bounded ones and then numerically solve the discrete equation. For example, in Proposition 3.1 of [19] the authors bound the error introduced by localising the domain. In general, such domain localisations lead to the introduction of artificial boundary conditions. In financial problems, for example, it is common to localise the domain and impose approximate boundary conditions, see [6, 5]. The approximate boundary condition is inferred from the asymptotic behaviour of the solution, however, as commented in [6, 5], for certain problems this is not obvious. Moreover, for the Black-Scholes PDE, as shown in [83], these artificial boundary conditions may impact the properties of the numerical schemes and requires further analysis.

The aim of this chapter is to discuss items 1. and 3. in Section 6.2 of [26] in detail and demonstrate two possible ways to deal with these situations. For bounded and regular domains with Dirichlet boundary conditions, we propose a truncation of the semi-Lagrangian stencil together with a modification of the associated finite difference weights. We then show that this modification, preserves monotonicity and consistency at the expense of a lower local consistency error and a worse CFL condition for explicit time-stepping schemes. The resulting truncated scheme is therefore known to converge to the unique viscosity solution of the problem by the theory of Barles and Souganidis [10].

For the Black-Scholes equation, we use the approach in [85]. In particular, we show that it is possible to map the (semi) infinite domain into a bounded one avoiding the domain localisation and the introduction of artificial boundary conditions. Such mappings modify the equation coefficients such that the semi-Lagrangian stencil does not overstep.

The rest of the chapter is organised as follows: Section 2.1 reproduces known results on the satisfaction of Dirichlet boundary conditions for the solution of the HJB equation in the viscosity sense. Section 2.2 discusses the truncation of the LISL scheme for points whose stencil exceeds the domain and compares its performance to naïve extrapolations of the boundary conditions. Section 2.3 discusses a possible transformations for (semi) infinite domains whose boundaries have no regular points. Section 2.4 contains the final remarks.

## 2.1 Results on Dirichlet boundary conditions for HJB equations

This section briefly discusses the treatment of Dirichlet boundary conditions in the framework of viscosity solutions. The motivation for this is that viscosity solutions are defined for degenerate PDEs and it is known that such degeneracies affect the behaviour of the solution at the boundary. In particular, boundary conditions may not be satisfied by the solution.

One of the first studies on degeneracy in parabolic PDEs was done in [31]. The article studies the boundary behaviour at  $x = 0$  for the following parabolic PDE

$$\frac{\partial u}{\partial t} = \frac{\partial^2}{\partial x^2} (axu) - \frac{\partial}{\partial x} ((bx + c)u), \quad \text{for } (t, x) \in (0, T] \times (0, \infty), \quad (2.1.1)$$

where  $a, b, c$  are real-valued constants with  $a > 0$ . The author defines a regular



boundary point  $x$  as a boundary point for which the solution satisfies the initial and boundary conditions pointwise, and reports conditions, in terms of the values of the parameters  $a, b, c$ , for  $x = 0$  to be regular. For instance, in finance, the conclusions of [31] were used to derive the positivity conditions of the Cox-Ingersoll-Ross (CIR) model for interest rates, see [21]. Let  $V(t, r)$  be the price of a contingent claim whose value depends on time  $t$  and the interest rate value  $r$  where  $r$  follows a CIR process, then  $V(t, r)$  is the solution of

$$\frac{\partial V}{\partial t} = \frac{1}{2}\sigma^2 r \frac{\partial^2 V}{\partial r^2} + a(b-r) \frac{\partial V}{\partial r} - rV, \quad \text{for } (t, r) \in (0, T] \times (0, \infty),$$

where  $a, b \in \mathbb{R}$  and  $\sigma \in \mathbb{R}^+$  are the parameters of the model. The results in [31] state that we do not need to prescribe a boundary condition at  $r = 0$  if  $2ab \geq \sigma^2$ . This is because, from the stochastic perspective, as noted in [21], the drift pushes the process upwards away from zero<sup>1</sup>. However, if the above condition is not satisfied, then  $r = 0$  is a regular point and boundary conditions can be prescribed.

The results in [31] have been generalised for linear parabolic or elliptic PDEs with non-negative characteristic forms in [66, 36, 37]. For HJB equations on smooth spatial domains the relevant reference is [9], which generalises the linear case to the controlled case. The later article [20] relaxes some of the domain smoothness assumptions in [9]. The rest of this section reproduces the results in [9, 20], where the ideas in [31] are extended to more general coefficients, controlled equations and higher dimensions with curvilinear boundaries.

Following the notation in [20], for any  $x \in \partial\Omega$  we introduce the function set  $Z(x)$

$$Z(x) = \{\zeta \in C^2(\mathbb{R}^d) \mid \forall x \in \partial\Omega \zeta(x) = 0, D\zeta(x) \neq 0 \text{ and } \zeta > 0 \text{ in } \Omega\}. \quad (2.1.2)$$

---

<sup>1</sup>Excluding negative interest rates was a desirable feature of an interest rate model at the time the article was published.

Let  $d$  be the distance function to the boundary  $\partial\Omega$ , similarly to [9, 20], we define the following subsets of  $\partial\Omega$  for  $t_0 \in (0, T]$ :

$$\Gamma_{in}(t_0) := \left\{ x \in \partial\Omega \left| \begin{array}{l} d \in C^2 \text{ in a neighbourhood of } x \text{ and} \\ \forall \alpha \in \mathcal{A}, \sigma^{\alpha, T}(t_0, x) Dd(x) = 0 \text{ and} \\ \text{tr}[a^\alpha(t_0, x) D^2 d(x)] - b^\alpha(t_0, x) Dd(x) \geq 0 \end{array} \right. \right\}, \quad (2.1.3)$$

and

$$\Gamma_{out}(t_0) := \left\{ x \in \partial\Omega \left| \begin{array}{l} \exists \zeta \in Z(x) \text{ such that} \\ \forall \alpha \in \mathcal{A}, \sigma^{\alpha, T}(t_0, x) D\zeta(x) \neq 0 \text{ or} \\ \text{tr}[a^\alpha(t_0, x) D^2 \zeta(x)] - b^\alpha(t_0, x) D\zeta(x) < 0 \end{array} \right. \right\}. \quad (2.1.4)$$

As noted in [20], the set  $\Gamma_{in}$  is a subset of the smooth part of  $\partial\Omega$ , whereas  $\Gamma_{out}$  is part of the subset of the domain where  $Z(x)$  is non-empty. Nonemptiness of  $Z(x)$ , as noted in [20], holds under the exterior sphere condition:

$$\forall x \in \partial\Omega, \exists \varepsilon \in \mathbb{R}^d \setminus \{0\} \text{ such that } \overline{\mathcal{B}(x + \varepsilon, |\varepsilon|)} \cap \overline{\Omega} = \{x\}.$$

These two sets are related to the regularity of the points on the boundary  $\partial\Omega$ . More specifically, for points in  $\Gamma_{in}$  the equation holds up to the boundary whereas for points in  $\Gamma_{out}$  the Dirichlet boundary conditions are satisfied pointwise.

In terms of the underlying stochastic process, if the equation holds up to the boundary for  $x \in \partial\Omega$ , it means that all the paths of the underlying controlled stochastic process departing from  $x \in \partial\Omega$  stay within the domain for a small time regardless of the value of the control, see [9]. To characterize this part of the boundary we use assumption (H4) in [9]:

**(H4)**  $\Omega$  is a smooth, bounded domain of  $\mathbb{R}^d$  with a  $W^{3,\infty}$ -boundary  $\partial\Omega$ .

This assumption is used together with Lemma 14.16 in [38] to deduce that the

distance function “ $d$  is a  $W^{3,\infty}$ -function in a neighbourhood of  $\partial\Omega$ , and therefore  $d \in C^2$  in this neighbourhood”, ensuring the regularity to define  $\Gamma_{in}$ . We note that the use of weak derivatives in the definition allows for domain with corners.

Following the notation in [9], we denote by  $\mathcal{A}_{in}$  the set of controls for which a path of the stochastic process with initial state  $x \in \partial\Omega$  stays in  $\overline{\Omega}$  for small time.

$$\mathcal{A}_{in}(t, x) = \{\alpha \in \mathcal{A} : \sigma^{\alpha, T}(t, x)n(x) = 0 \text{ and } \text{tr}[a^\alpha(t, x)D^2d(x)] + b^\alpha(t, x)Dd(x) \geq 0\}, \quad (2.1.5)$$

where  $n(x)$  is the outward unit normal to  $\partial\Omega$  at  $x$ . We can therefore establish a correspondence for the regularity of a boundary point in terms of  $\mathcal{A}_{in}(t, x)$ .

**Definition 2.1.1.** A boundary point  $x \in \partial\Omega$  is **not regular** at time  $t \in (0, T]$ , i.e.  $x \in \Gamma_{in}(t)$ , if  $\mathcal{A}_{in}(t, x)$  equals the whole control set  $\mathcal{A}$ . Then, at this point  $(t, x)$  the equation holds up to the boundary. Alternatively, if for  $(t, x) \in (0, T] \times \partial\Omega$  we have that  $\mathcal{A}_{in}(t, x) = \emptyset$  then the point is **regular**, i.e.  $x \in \Gamma_{out}(t)$ , and boundary conditions are satisfied pointwise.

## 2.2 Truncation of the LISL scheme

In this section, we analyse adaptations of the Linear Interpolation Semi-Lagrangian (LISL) scheme for initial-boundary value problems on bounded domains.

Following the notation in [26], the LISL finite difference approximations for the differential operator in (1.1.4) can be expressed as

$$L_{\Delta x}^\alpha[\mathcal{I}_{\Delta x}\phi](t, x) := \sum_{p=1}^M \frac{[\mathcal{I}_{\Delta x}\phi](t, x + y_p^{\alpha,+}(t, x)) - 2\phi(t, x) + [\mathcal{I}_{\Delta x}\phi](t, x + y_p^{\alpha,-}(t, x))}{2\Delta x}, \quad (2.2.1)$$

for  $x \in \Omega_{\Delta x}$ , and some  $M \geq 1$ .

For convenience, we repeat the three schemes from [26] already discussed in the introduction.

### Examples of LISL schemes.

1. **Scheme 1:** The approximation of Camilli and Falcone [19], corresponding to  $y_p^{\alpha,\pm} = \pm\sqrt{\Delta x}\sigma_p^\alpha + \frac{\Delta x}{P}b^\alpha$  and  $M = P$ .
2. **Scheme 2:** The approximation in [26], corresponding to  $y_p^{\alpha,\pm} = \pm\sqrt{\Delta x}\sigma_p^\alpha$  for  $p \leq P$ ,  $y_{P+1}^{\alpha,\pm} = \Delta x b^\alpha$ , and  $M = P + 1$ .
3. **Scheme 3:** A more efficient version of the Camilli-Falcone approximation, corresponding to  $y_p^{\alpha,\pm} = \pm\sqrt{\Delta x}\sigma_p^\alpha$  for  $p < P$ ,  $y_P^{\alpha,\pm} = \pm\sqrt{\Delta x}\sigma_P^\alpha + \Delta x b^\alpha$ , and  $M = P$ .

As described in the introduction, for points  $x$  close to the boundaries of the domain, the stencil points  $x + y_p^{\alpha,\pm}(t, x)$  in (2.2.1) may lie outside the domain. In the following, we discuss a modification of (2.2.1) to prevent the overstepping and so that the resulting scheme remains monotone, consistent, and  $L^\infty$ -stable. The proposed modification is a truncation of the stencil that samples the spatial boundary in the direction of  $y_p^{\alpha,\pm}(t, x)$ , and adjusts the finite difference weights for consistency.

The final scheme arises from discretising in time using the standard  $\theta$ -time stepping scheme for  $\theta \in [0, 1]$ , where  $\theta = 0$  corresponds to the explicit Euler time stepping and  $\theta = 1$  to the implicit case, on a time grid represented by a strictly increasing sequence of points  $\{t_n\}_{n=0}^{N_t+1}$  with  $t_0 = 0$ ,  $t_{N_t+1} = T$ , and  $\Delta t_n := t_n - t_{n-1} \leq \Delta t$  for all  $n$ . The scheme being monotone, it can be written as described in the following definition, where for any grid function  $V : \{t_n\}_{n=0}^{N_t+1} \times \Omega_{\Delta x} \rightarrow \mathbb{R}$ ,  $V_i^n \equiv V(t_n, x_i)$ .

**Definition 2.2.1** (Equation (4.1) in [26]). A scheme is said to be of positive type, if

it can be written as

$$\max_{\alpha \in \mathcal{A}} \left\{ \mathcal{B}_{j,j}^{\alpha,n,n} U_j^n - \sum_{i \neq j} \mathcal{B}_{j,i}^{\alpha,n,n} U_i^n - \sum_{i=1}^N \mathcal{B}_{j,i}^{\alpha,n,n-1} U_i^{n-1} - F_j^{\alpha,n-1+\theta} \right\} = 0, \quad (2.2.2)$$

for  $j = 1, \dots, N$ , on the discrete domain  $\{t_n\}_{n=0}^{N_t+1} \times \Omega_{\Delta x}$ , where  $U_i^n$  is the numerical solution at node  $(t_n, x_i)$  and all the coefficients  $\mathcal{B}$  are non-negative.

For completeness, we reproduce the expressions for  $\mathcal{B}_{j,\cdot}^{\alpha,n,\cdot}$  of the LISL schemes as in [26], for all  $1 \leq i \neq j \leq N$ ,  $x_i, x_j \notin \partial\Omega$ ,

$$\begin{aligned} \mathcal{B}_{j,j}^{\alpha,n,n} &= 1 + \theta \Delta t_n \left( \frac{M}{2\Delta x} - l_{j,j}^{\alpha,n} - c_j^{\alpha,n-1+\theta} \right), & \mathcal{B}_{j,i}^{\alpha,n,n} &= \theta \Delta t_n l_{j,i}^{\alpha,n}, \\ \mathcal{B}_{j,j}^{\alpha,n,n-1} &= 1 - (1 - \theta) \Delta t_n \left( \frac{M}{2\Delta x} - l_{j,j}^{\alpha,n-1} - c_j^{\alpha,n-1+\theta} \right), & \mathcal{B}_{j,i}^{\alpha,n,n-1} &= (1 - \theta) \Delta t_n l_{j,i}^{\alpha,n-1}, \end{aligned}$$

where  $c_j^{\alpha,n-1+\theta} = c^\alpha(t_{n-1} + \theta \Delta t, x_j)$  and

$$l_{j,i}^{\alpha,n} = \sum_{p=1}^M \frac{w_i(x_j + y_p^{\alpha,+}(t_n, x_j)) + w_i(x_j + y_p^{\alpha,-}(t_n, x_j))}{2\Delta x}.$$

### 2.2.1 Definition of truncated stencils

We take  $\Omega \subset \mathbb{R}^d$  for  $d \geq 2$ . We first outline how the method can be defined on a general domain with curved boundary, but later (especially in the numerical tests) focus for simplicity on rectangular domains. We start with a Cartesian mesh on  $\mathbb{R}^d$  with uniform mesh width  $\Delta x$  and then choose  $\Omega_{\Delta x}$  as all the points which lie inside  $\Omega$ . See Figure 2.2.1.

We now fix a mesh node  $x \in \Omega_{\Delta x}$ . There are two distinct situations where interpolation at the point  $x + y_p^{\alpha,\pm}(t, x)$  as per (2.2.1) is not possible for given  $t, \alpha$  and  $p$ :

- A.  $x + y_p^{\alpha,\pm}(t, x) \notin \bar{\Omega}$  (bottom left in Fig. 2.2.1);

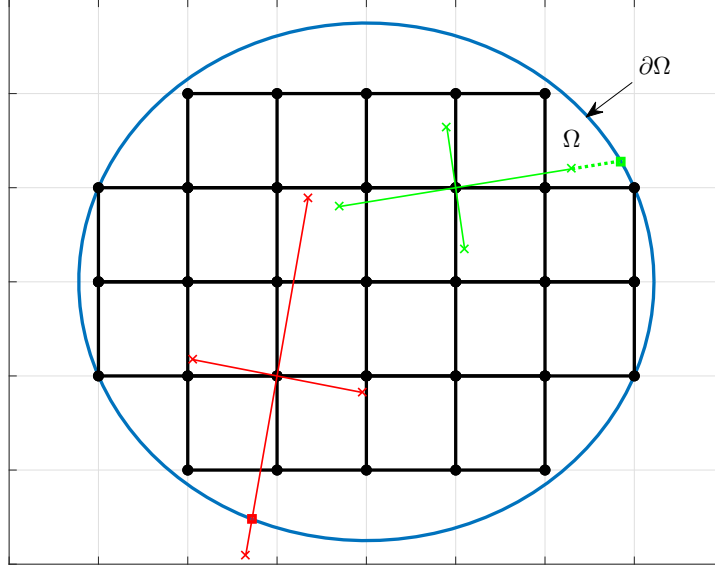


Figure 2.2.1: Truncation and extrapolation of the stencil for an elliptical domain and a mesh made of square cells. The modified stencil samples the domain boundary.

B.  $x + y_p^{\alpha, \pm}(t, x) \in \bar{\Omega}$ , but the element it is contained in has vertices outside  $\bar{\Omega}$  (top right).

We say the stencil “oversteps”. In such cases, the objective is to find truncated or extended stencil vectors  $\hat{y}_p^{\alpha, \pm}(t, x)$  and corresponding finite difference weights  $A_p^\alpha \equiv A_p^\alpha(t, x)$  and  $B_p^\alpha \equiv B_p^\alpha(t, x)$ , such that  $x + \hat{y}_p^{\alpha, \pm}(t, x) \in \partial\Omega$  and the truncated scheme

$$\begin{aligned} \hat{L}_{\Delta x}^\alpha[\mathcal{I}_{\Delta x}\phi](t, x) := \\ \sum_{p=1}^M \frac{A_p^\alpha(\mathcal{I}_{\Delta x}\phi)(t, x + \hat{y}_p^{\alpha, +}(t, x)) - (A_p^\alpha + B_p^\alpha)\phi(t, x) + B_p^\alpha(\mathcal{I}_{\Delta x}\phi)(t, x + \hat{y}_p^{\alpha, -}(t, x))}{2\Delta x} \end{aligned} \quad (2.2.3)$$

is a consistent approximation of (1.1.4) as  $\Delta x \rightarrow 0$ . If the stencil does not overstep, we have that  $\hat{y}_p^{\alpha, \pm}(t, x) = y_p^{\alpha, \pm}(t, x)$  and  $A_p^\alpha = B_p^\alpha = 1$ . If it does, for any  $t$  we define

$$\hat{y}_p^{\alpha, \pm}(t, x) = \mu_p^{\alpha, \pm}(t, x) y_p^{\alpha, \pm}(t, x),$$

where

$$\mu_p^{\alpha,\pm}(t, x) = \min \{ \mu \geq 0 : x + \mu y_p^{\alpha,\pm}(t, x) \in \partial\Omega \}.$$

In case A, this means  $\mu < 1$ , while in case B we have  $\mu > 1$ .

In the remainder of this section we restrict our attention to the truncation of the scheme on rectangular domains, in which case the elements of the Cartesian mesh cover exactly the domain and case B does not occur. Moreover, this means that interior mesh points cannot be arbitrarily close to the boundary, but are always at least  $\Delta x$  away<sup>2</sup>. This allows the derivation of CFL conditions for the explicit schemes as given below in Section 2.2.3.

## 2.2.2 Consistency conditions

In the truncated scheme (2.2.3) there are  $M$  pairs of weights, which can be chosen freely, subject to positivity, in order to obtain a consistent scheme. As we will see below, this is only possible for Scheme 2.

In the following, we denote  $[[1, j]] \equiv [1, j] \cap \mathbb{Z}$  and for a vector  $v \in \mathbb{R}^d$ ,  $(v)_i$  denotes its  $i$ -th element. As in the introduction, we have that  $b^\alpha \in \mathbb{R}^d$ , and  $\sigma^\alpha = (\sigma_1^\alpha, \dots, \sigma_p^\alpha, \dots, \sigma_P^\alpha) \in \mathbb{R}^{d \times P}$  where  $\sigma_p^\alpha \in \mathbb{R}^d$  denotes the  $p$ -th column vector. For compactness, we omit the dependence of the coefficients and the stencil related functions with respect to the position, that is  $b^\alpha \equiv b^\alpha(t, x)$ ,  $\sigma_p^\alpha \equiv \sigma_p^\alpha(t, x)$ ,  $y_p^{\alpha,\pm} \equiv y_p^{\alpha,\pm}(t, x)$  and  $\mu_p^{\alpha,\pm} \equiv \mu_p^{\alpha,\pm}(t, x)$ . We add a second subscript taking values 1, 2 or 3 to  $A_p^\alpha$ ,  $B_p^\alpha$  and  $y_p^{\alpha,\pm}$  to make the discretization scheme explicit. For example,  $A_{1,p}^\alpha$  and  $B_{1,p}^\alpha$  are the finite difference weights for scheme 1.

**Proposition 2.2.1.** *The truncated version of Schemes 1 and 3 is generally not consistent.*

---

<sup>2</sup>This can also be enforced in the general case by removing the outermost layer of cells, such that again a distance of  $\Delta x$  between non-boundary mesh points and the domain boundary is ensured.

*Proof.* By Taylor expansion of any smooth function  $\phi$  and  $\hat{y}_p^{\alpha,\pm}(t, x) \equiv \hat{y}_p^{\alpha,\pm}$ , we have that

$$\phi(t, x + \hat{y}_p^{\alpha,\pm}) = \phi(t, x) + \sum_{i=1}^d \phi_{x_i}(\hat{y}_p^{\alpha,\pm})_i + \frac{1}{2} \sum_{i_1=1}^d \sum_{i_2=1}^d \phi_{x_{i_1} x_{i_2}}(\hat{y}_p^{\alpha,\pm})_{i_1} (\hat{y}_p^{\alpha,\pm})_{i_2} + \mathcal{O}((\mu_p^\pm)^3 \Delta x^{3/2}).$$

Hence, the consistency conditions for the first and second order terms for Scheme 1 are

$$\begin{aligned} \sum_{p \in \mathcal{P}} (A_{1,p}^\alpha (\hat{y}_{1,p}^{\alpha,+})_i + B_{1,p}^\alpha (\hat{y}_{1,p}^{\alpha,-})_i) &= 2\Delta x \frac{|\mathcal{P}|}{P} (b^\alpha)_i + o(\Delta x), \\ \sum_{p \in \mathcal{P}} (A_{1,p}^\alpha (\hat{y}_{1,p}^{\alpha,+})_{i_1} (\hat{y}_{1,p}^{\alpha,+})_{i_2} + B_{1,p}^\alpha (\hat{y}_{1,p}^{\alpha,-})_{i_1} (\hat{y}_{1,p}^{\alpha,-})_{i_2}) &= 2\Delta x \sum_{p \in \mathcal{P}} (\sigma_p^\alpha)_{i_1} (\sigma_p^\alpha)_{i_2} + o(\Delta x), \end{aligned}$$

where  $\mathcal{P} \subseteq [[1, P]]$  denotes the set of stencils overstepping the domain and  $i, i_1, i_2 \in [[1, d]]$ .

In Scheme 1, there are  $2|\mathcal{P}| \leq 2d$  variables, but  $(d^2 + 3d)/2$  equations,  $d$  from the condition on the Jacobian and  $(d^2 + d)/2$  from the condition on the Hessian. For  $d \geq 2$  this overdetermined system has a solution only if there is linear dependence between the equations. Except for special cases, e.g.  $|\mathcal{P}| = 0$  or  $\sigma_p^\alpha$  parallel to  $b^\alpha$  for some  $p$ , this is not the case. Hence, in general the truncated Scheme 1 is not consistent.

We observe that the same principle applies to Scheme 3 for  $y_{3,p}^{\alpha,\pm} = \pm \sqrt{\Delta x} \sigma_p^\alpha + \Delta x b^\alpha$ .

□

The following examples illustrate the result in Proposition 2.2.1.

**Example 2.2.1.** Let  $x = (0, 0)^T$ ,  $\mathcal{A} = \{\alpha\}$ ,  $\bar{\Omega} = [-5, 1]^2$ , and  $\sqrt{\Delta x} \sigma^\alpha = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ ,



and  $\Delta x b^\alpha = (0, 1)^T$ . For Scheme 1 we have that

$$y_{1,1}^{\alpha,\pm} = \pm \begin{pmatrix} 2 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix}, \quad y_{1,2}^{\alpha,\pm} = \pm \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix},$$

we truncate the  $+$  side of the stencils  $\hat{y}_{1,1}^{\alpha,+} = \begin{pmatrix} 1 \\ 1/4 \end{pmatrix}$ , and  $\hat{y}_{1,2}^{\alpha,\pm} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ , which gives  $\mu_{1,1}^{\alpha,+} = \frac{1}{2}$  and  $\mu_{1,2}^{\alpha,+} = \frac{2}{3}$ . For the  $-$  side of the stencils  $y_{1,1}^{\alpha,-} = \begin{pmatrix} -2 \\ 1/2 \end{pmatrix}$ , and  $y_{1,2}^{\alpha,-} = \begin{pmatrix} 0 \\ -1/2 \end{pmatrix}$ . From the gradient we obtain the consistency conditions on the left and for the Hessian the ones on the right

$$\begin{cases} 1A_{1,1}^\alpha - 2B_{1,1}^\alpha + 0A_{1,2}^\alpha + 0B_{1,2}^\alpha &= 0 \\ \frac{1}{4}A_{1,1}^\alpha + \frac{1}{2}B_{1,1}^\alpha + 1A_{1,2}^\alpha - \frac{1}{2}B_{1,2}^\alpha &= 1 \end{cases}, \quad \begin{cases} 1A_{1,1}^\alpha + (-2)^2B_{1,1}^\alpha + 0A_{1,2}^\alpha + 0B_{1,2}^\alpha &= 4 \\ \frac{1}{4^2}A_{1,1}^\alpha + \frac{1}{2^2}B_{1,1}^\alpha + 1A_{1,2}^\alpha + \frac{1}{2^2}B_{1,2}^\alpha &= 2, \\ \frac{1}{4}A_{1,1}^\alpha - 2\frac{1}{2}B_{1,1}^\alpha + 0 \cdot 1A_{1,2}^\alpha + 0\frac{1}{2}B_{1,2}^\alpha &= 0 \end{cases}$$

the system does not have a solution as equation 1 from the gradient is not compatible with equations 1 and 3 from the Hessian.

For Scheme 3 the stencils are

$$y_{3,1}^{\alpha,\pm} = \pm \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \quad y_{3,2}^{\alpha,\pm} = \pm \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

we truncate the  $+$  side of the stencils  $\hat{y}_{3,1}^{\alpha,+} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ , and  $\hat{y}_{3,2}^{\alpha,\pm} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ , which gives  $\mu_{3,1}^{\alpha,+} = \frac{1}{2}$  and  $\mu_{3,2}^{\alpha,+} = \frac{1}{2}$ . For the  $-$  sides we have  $y_{3,1}^{\alpha,-} = \begin{pmatrix} -2 \\ 0 \end{pmatrix}$  and  $y_{3,2}^{\alpha,-} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ .

From the gradient we obtain the consistency conditions on the left and for the Hessian the ones on the right

$$\begin{cases} 1A_{3,1}^\alpha - 2B_{3,1}^\alpha + 0A_{3,2}^\alpha + 0B_{3,2}^\alpha &= 0 \\ 0A_{3,1}^\alpha + 0B_{3,1}^\alpha + 1A_{3,2}^\alpha + 0B_{3,2}^\alpha &= 1 \end{cases}, \quad \begin{cases} 1A_{3,1}^\alpha + 2^2B_{3,1}^\alpha + 0A_{3,2}^\alpha + 0B_{3,2}^\alpha &= 4 \\ 0A_{3,1}^\alpha + 0B_{3,1}^\alpha + 1A_{3,2}^\alpha + 0B_{3,2}^\alpha &= 2, \\ 0A_{3,1}^\alpha + 0B_{3,1}^\alpha + 0A_{3,2}^\alpha + 0B_{1,2}^\alpha &= 0 \end{cases}$$

Removing the drift from the consistency it yields a solvable system, i.e.

$$\begin{cases} 1A_{3,1}^\alpha - 2B_{3,1}^\alpha + 0A_{3,2}^\alpha + 0B_{3,2}^\alpha &= 0 \\ 0A_{3,1}^\alpha + 0B_{3,1}^\alpha + 1A_{3,2}^\alpha + 0B_{3,2}^\alpha &= 1 \end{cases}, \quad \begin{cases} 1A_{3,1}^\alpha + 2^2B_{3,1}^\alpha + 0A_{3,2}^\alpha + 0B_{3,2}^\alpha &= 4 \\ 0A_{3,1}^\alpha + 0B_{3,1}^\alpha + 1A_{3,2}^\alpha + 0B_{3,2}^\alpha &= 1, \\ 0A_{3,1}^\alpha + 0B_{3,1}^\alpha + 0A_{3,2}^\alpha + 0B_{1,2}^\alpha &= 0 \end{cases}$$

where  $A_{3,1}^\alpha = 8/6$ ,  $B_{3,1}^\alpha = 4/6$  and  $A_{3,2}^\alpha = B_{3,2}^\alpha = 1$ .

Changing the drift to  $\Delta x b^\alpha = (1, 1)^T$ . The stencils for Scheme 1 are

$$y_{1,1}^{\alpha,\pm} = \pm \begin{pmatrix} 2 \\ 0 \end{pmatrix} + \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}, \quad y_{1,2}^{\alpha,\pm} = \pm \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix},$$

we truncate the  $+$  side of both stencils  $\hat{y}_{1,1}^{\alpha,+} = \begin{pmatrix} 1 \\ 1/5 \end{pmatrix}$ , and  $\hat{y}_{1,2}^{\alpha,\pm} = \begin{pmatrix} 1/3 \\ 1 \end{pmatrix}$ , which

gives  $\mu_{1,1}^{\alpha,+} = \frac{2}{5}$  and  $\mu_{1,2}^{\alpha,+} = \frac{2}{3}$ . For the  $-$  side of the stencils  $y_{1,1}^{\alpha,-} = \begin{pmatrix} -3/2 \\ 1/2 \end{pmatrix}$ , and

$$y_{1,2}^{\alpha,-} = \begin{pmatrix} 1/2 \\ -1/2 \end{pmatrix}.$$

From the gradient we obtain the consistency conditions on the left and for the

Hessian the ones on the right

$$\begin{cases} 1A_{1,1}^\alpha - \frac{3}{2}B_{1,1}^\alpha + \frac{1}{3}A_{1,2}^\alpha + \frac{1}{2}B_{1,2}^\alpha &= 1 \\ \frac{1}{5}A_{1,1}^\alpha + \frac{1}{2}B_{1,1}^\alpha + 1A_{1,2}^\alpha - \frac{1}{2}B_{1,2}^\alpha &= 1 \end{cases}, \quad \begin{cases} 1A_{1,1}^\alpha + \frac{3^2}{2^2}B_{1,1}^\alpha + \frac{1}{3^2}A_{1,2}^\alpha + \frac{1}{2^2}B_{1,2}^\alpha &= 4 \\ \frac{1}{5^2}A_{1,1}^\alpha + \frac{1}{2^2}B_{1,1}^\alpha + 1A_{1,2}^\alpha + \frac{1}{2^2}B_{1,2}^\alpha &= 1, \\ \frac{1}{5}A_{1,1}^\alpha - \frac{3}{4}B_{1,1}^\alpha + \frac{1}{3}A_{1,2}^\alpha - \frac{1}{4}B_{1,2}^\alpha &= 0 \end{cases}$$

the augmented matrix has rank 5.

The stencils for Scheme 3 are

$$y_{3,1}^{\alpha,\pm} = \pm \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \quad y_{3,2}^{\alpha,\pm} = \pm \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

we truncate the + side of both stencils  $\hat{y}_{3,1}^{\alpha,+} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ , and  $\hat{y}_{3,2}^{\alpha,\pm} = \begin{pmatrix} 1/2 \\ 1 \end{pmatrix}$ , which gives  $\mu_{3,1}^{\alpha,+} = \frac{1}{2}$  and  $\mu_{3,2}^{\alpha,+} = \frac{1}{2}$ . For the - side of the stencils  $y_{3,1}^{\alpha,-} = \begin{pmatrix} -2 \\ 0 \end{pmatrix}$ , and

$$y_{3,2}^{\alpha,-} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

From the gradient we obtain the consistency conditions on the left and for the Hessian the ones on the right

$$\begin{cases} 1A_{3,1}^\alpha - 2B_{3,1}^\alpha + \frac{1}{2}A_{3,2}^\alpha + 1B_{3,2}^\alpha &= 1 \\ 0A_{3,1}^\alpha + 0B_{3,1}^\alpha + 1A_{3,2}^\alpha + 0B_{3,2}^\alpha &= 1 \end{cases}, \quad \begin{cases} 1A_{3,1}^\alpha + 2^2B_{3,1}^\alpha + \frac{1}{2^2}A_{3,2}^\alpha + 1B_{3,2}^\alpha &= 4 \\ 0A_{3,1}^\alpha + 0B_{3,1}^\alpha + 1A_{3,2}^\alpha + 0B_{3,2}^\alpha &= 1, \\ 0A_{3,1}^\alpha + 0B_{3,1}^\alpha + \frac{1}{2}A_{3,2}^\alpha + 0B_{3,2}^\alpha &= 0 \end{cases}$$

the system does not have a solution as the second equation from the gradient and the last two from the Hessian are not compatible.

**Remark 2.2.1.** Even when the set of consistency conditions for Schemes 1 and 3

admits a solution, it is possible that the resulting discretization is not monotone. For simplicity consider a one-dimensional spatial domain. In this case, Schemes 1 and 3 are equivalent as the stencil is  $y_{3,1}^{\alpha,\pm} = y_{1,1}^{\alpha,\pm} = \pm\sqrt{\Delta x}\sigma_1^\alpha + \Delta x b^\alpha$ . Hence, for simplicity  $y_1^{\alpha,\pm} \equiv y_{1,1}^{\alpha,\pm}$  throughout this remark.

The truncation weights are given by

$$A_1^\alpha = \frac{2\Delta x(\sigma^\alpha)^2 + 2\Delta x^2(b^\alpha)^2 - 2\Delta x(b^\alpha)\hat{y}_1^{\alpha,-}}{\hat{y}_1^{\alpha,+}(\hat{y}_1^{\alpha,+} - \hat{y}_1^{\alpha,-})}, \quad (2.2.4)$$

$$B_1^\alpha = \frac{2\Delta x(\sigma^\alpha)^2 + 2\Delta x^2(b^\alpha)^2 - 2\Delta x(b^\alpha)\hat{y}_1^{\alpha,+}}{\hat{y}_1^{\alpha,-}(\hat{y}_1^{\alpha,-} - \hat{y}_1^{\alpha,+})}. \quad (2.2.5)$$

The term  $2\Delta x^2(b^\alpha)^2$  in the numerator does not affect the consistency condition, but it facilitates the analysis and if the stencils do not overstep  $A_1^\alpha = B_1^\alpha = 1$  is a solution.

Using the definition of the truncated stencils  $\hat{y}_1^{\alpha,+}$  in (2.2.4)–(2.2.5), we see that the truncated scheme is monotone with  $A_1^\alpha, B_1^\alpha \geq 1$  if  $\text{sign}(y_1^{\alpha,+}) \neq \text{sign}(y_1^{\alpha,-})$ . However, if the drift term dominates the diffusion in  $y_1^{\alpha,\pm}$ , then  $\text{sign}(y_1^{\alpha,+}) = \text{sign}(y_1^{\alpha,-})$  and the truncated scheme is monotone with  $A_1^\alpha, B_1^\alpha \geq 0$  providing that

$$y_1^{\alpha,+} = y_1^{\alpha,-}, \text{ or } (|\hat{y}_1^{\alpha,+}|, |\hat{y}_1^{\alpha,-}|) \in \left[ \frac{(\sigma^\alpha)^2 + \Delta x(b^\alpha)^2}{|b^\alpha|}, |y_1^{\alpha,+}| \right] \times \left[ \frac{(\sigma^\alpha)^2 + \Delta x(b^\alpha)^2}{|b^\alpha|}, |y_1^{\alpha,-}| \right].$$

The case when  $y_1^{\alpha,+} = y_1^{\alpha,-}$  corresponds to the first order Hamilton–Jacobi case and the scheme is monotone by construction<sup>3</sup>.

For example, let  $\sqrt{\Delta x}\sigma_1^\alpha = 1.5$ ,  $\Delta x b^\alpha = 3.5$ ,  $\hat{y}_1^{\alpha,+} = 3$  and  $\hat{y}_1^{\alpha,-} = y_1^{\alpha,-} = 2$ , then applying (2.2.4)–(2.2.5) we have that  $A_1^\alpha = 5$  and  $B_1^\alpha = -4$ .

We conclude that for points whose stencil oversteps the boundary, the approximations of the first and second derivative should be considered separately, as done in Scheme 2.

---

<sup>3</sup>Equations (2.2.4)–(2.2.5) are not applicable in the first order case as they are derived assuming that the diffusion is non zero. For this case we can use (2.2.6) below.

**Proposition 2.2.2.** *For Scheme 2 and all  $p \in [[1, P + 1]]$ , let  $\mu_p^{\alpha, \pm} \in [0, 1]$  be the smallest constant such that  $x + \mu_p^{\alpha, \pm} y_{2,p}^{\alpha, \pm} \in \overline{\Omega}$  and define*

$$A_{2,P+1}^\alpha = B_{2,P+1}^\alpha = \frac{1}{\mu_{P+1}^{\alpha,+}} \left( = \frac{1}{\mu_{P+1}^{\alpha,-}} \right), \quad (2.2.6)$$

and, for  $p \in [[1, P]]$ ,

$$A_{2,p}^\alpha = \frac{2}{(\mu_p^{\alpha,+})^2 + \mu_p^{\alpha,+} \mu_p^{\alpha,-}}, \quad B_{2,p}^\alpha = \frac{2}{(\mu_p^{\alpha,-})^2 + \mu_p^{\alpha,-} \mu_p^{\alpha,+}}. \quad (2.2.7)$$

Then the scheme defined by (2.2.3) is consistent.

*Proof.* If the stencils overstep, then the truncated stencil consists of the point at the intersection between the boundary  $\partial\Omega$  and one of the segments  $\{x, x + \sqrt{\Delta x} \sigma_p^\alpha\}$ ,  $\{x, x - \sqrt{\Delta x} \sigma_p^\alpha\}$ , or  $\{x, x + \Delta x b^\alpha\}$ . For each point  $(t, x)$  Scheme 2 requires the calculation of at most  $2P + 1$  different weights, i.e.  $2P$  for the second order term and one for the first order term. For the latter we have that  $\hat{y}_{2,P+1}^{\alpha,+} = \hat{y}_{2,P+1}^{\alpha,-}$ , therefore  $A_{2,P+1}^\alpha = B_{2,P+1}^\alpha$ . The coefficients are obtained from the following consistency conditions,

$$(A_{2,P+1}^\alpha + B_{2,P+1}^\alpha)(\hat{y}_{2,P+1}^{\alpha, \pm})_i = 2\Delta x(b^\alpha)_i, \quad \forall i \in [[1, d]], \quad (2.2.8)$$

for the first order term, and

$$A_{2,p}^\alpha(\hat{y}_{2,p}^{\alpha,+})_i + B_{2,p}^\alpha(\hat{y}_{2,p}^{\alpha,-})_i = 0, \quad \forall i \in [[1, d]], \quad (2.2.9)$$

$$A_{2,p}^\alpha(\hat{y}_{2,p}^{\alpha,+})_{i_1}(\hat{y}_{2,p}^{\alpha,+})_{i_2} + B_{2,p}^\alpha(\hat{y}_{2,p}^{\alpha,-})_{i_1}(\hat{y}_{2,p}^{\alpha,-})_{i_2} = 2\Delta x(\sigma_p^\alpha)_{i_1}(\sigma_p^\alpha)_{i_2}, \quad \forall (i_1, i_2) \in [[1, d]]^2, \quad (2.2.10)$$

for the second order term.

By construction of the truncated stencil (2.2.8) and (2.2.9) are linearly dependent

across  $i$ , and (2.2.10) across  $i_1$  and  $i_2$ , resulting in one (linearly independent) equation for the first order term weights and two for  $A_{2,p}^\alpha$ ,  $B_{2,p}^\alpha$ , with solutions given by

$$A_{2,P+1}^\alpha = B_{2,P+1}^\alpha = \Delta x \frac{(b^\alpha)_i}{(\hat{y}_{2,P+1}^{\alpha,\pm})_i}, \quad (2.2.11)$$

and

$$A_{2,p}^\alpha = \frac{2\Delta x (\sigma_p^\alpha)_i^2}{(\hat{y}_{2,p}^{\alpha,+})_i ((\hat{y}_{2,p}^{\alpha,+})_i - (\hat{y}_{2,p}^{\alpha,-})_i)}, \quad B_{2,p}^\alpha = \frac{2\Delta x (\sigma_p^\alpha)_i^2}{(\hat{y}_{2,p}^{\alpha,-})_i ((\hat{y}_{2,p}^{\alpha,-})_i - (\hat{y}_{2,p}^{\alpha,+})_i)}, \quad (2.2.12)$$

which are seen to be equivalent to equations (2.2.6) and (2.2.7).

The contribution to the consistency error of (2.2.3) from the bilinear interpolation operator  $\mathcal{I}$  is bounded by  $(\Delta x)^{-1} \sum_p (|A_p| + |B_p|)(\Delta x)^2$ , which goes to 0 if  $|A_p| + |B_p| = o((\Delta x)^{-1})$  for all  $p$ , which is violated if and only if  $\mu_p^{\alpha,+}, \mu_p^{\alpha,-} \sim \mathcal{O}(\sqrt{\Delta x})$ .  $\square$

**Corollary 2.2.3.** *For the truncated Scheme 2, (2.2.3), (2.2.6) and (2.2.7), the following holds:*

a) *The scheme is of positive type and monotone with  $A_{2,p}^\alpha, B_{2,p}^\alpha \geq 1$  for all  $p \in [[1, P+1]]$ .*

b) *For points  $x$  within a distance  $\mathcal{O}(\Delta x)$  of the boundary and  $p \neq P+1$ , as  $\Delta x \rightarrow 0$ ,*

$$\text{if } |\hat{y}_{2,p}^{\alpha,+}| < \sqrt{\Delta x} |\sigma_p^\alpha| \text{ and } |\hat{y}_{2,p}^{\alpha,-}| = \sqrt{\Delta x} |\sigma_p^\alpha| \implies A_{2,p}^\alpha \sim \mathcal{O}(\Delta x^{-1/2}) \text{ and } \lim_{\Delta x \rightarrow 0} B_{2,p}^\alpha = 2,$$

$$\text{if } |\hat{y}_{2,p}^{\alpha,-}| < \sqrt{\Delta x} |\sigma_p^\alpha| \text{ and } |\hat{y}_{2,p}^{\alpha,+}| = \sqrt{\Delta x} |\sigma_p^\alpha| \implies \lim_{\Delta x \rightarrow 0} A_{2,p}^\alpha = 2 \text{ and } B_{2,p}^\alpha \sim \mathcal{O}(\Delta x^{-1/2}),$$

$$\text{if } |\hat{y}_{2,p}^{\alpha,\pm}| < \sqrt{\Delta x} |\sigma_p^\alpha| \implies A_{2,p}^\alpha, B_{2,p}^\alpha \sim \mathcal{O}(\Delta x^{-1}).$$

c) *The local consistency error for points with truncation and  $p \neq P+1$  is  $\mathcal{O}(\sqrt{\Delta x})$*

*if only one side of the stencil oversteps, and  $\mathcal{O}(1)$  if both sides overstep.*

*Proof.* The claim in a) follows from (2.2.6), (2.2.7), and the fact that  $\mu_{2,p}^{\alpha,\pm} \in (0, 1]$  and the coefficients  $A_{2,p}^\alpha, B_{2,p}^\alpha$  do not depend on the numerical solution  $U$ . The limits

in *b*) follow from (2.2.7) and noting that if the stencil oversteps for a point  $x$  lying  $\mathcal{O}(\Delta x)$  away from the boundary, but at least  $\Delta x$  by the assumption made on the mesh, then  $\mu_{2,p}^{\alpha,+} \sim \mathcal{O}(\sqrt{\Delta x})$  and/or  $\mu_{2,p}^{\alpha,-} \sim \mathcal{O}(\sqrt{\Delta x})$ , but not  $o(\sqrt{\Delta x})$ .

To prove *c*) we use Taylor expansions for each  $p$  and conclude using the limits in *b*). Let  $\phi : \bar{\Omega} \rightarrow \mathbb{R}$  be a smooth function and for any  $p \in (\mathcal{P} \cap [[1, P]])$ , where  $\mathcal{P}$  denotes the set of stencils overstepping the domain, then by Taylor expansion and the consistency conditions (2.2.9)–(2.2.10) the local consistency error  $\tau$  for the  $p$ -th addend of (2.2.3) using multi-index notation is given by

$$\begin{aligned} \tau &:= \frac{A_{2,p}^{\alpha} \phi(t, x + \hat{y}_{2,p}^{\alpha,+}) - (A_{2,p}^{\alpha} + B_{2,p}^{\alpha}) \phi(t, x) + B_{2,p}^{\alpha} \phi(t, x + \hat{y}_{2,p}^{\alpha,-})}{2\Delta x} - \frac{1}{2} \text{tr}[\sigma_p^{\alpha} \sigma_p^{\alpha,T} D^2 \phi] \\ &= \frac{1}{2\Delta x} \sum_{|\beta| \geq 3} \frac{1}{|\beta|!} (A_{2,p}^{\alpha} (\hat{y}_{2,p}^{\alpha,+})^{\beta} + B_{2,p}^{\alpha} (\hat{y}_{2,p}^{\alpha,-})^{\beta}) D^{\beta} \phi, \end{aligned}$$

where, due to the truncation of the stencil, the scheme is not central and therefore the terms for odd  $|\beta|$  do not cancel out. If only one side of the stencil oversteps then for  $|\beta| = 3$

$$\frac{A_{2,p}^{\alpha} (\hat{y}_{2,p}^{\alpha,+})^{\beta} + B_{2,p}^{\alpha} (\hat{y}_{2,p}^{\alpha,-})^{\beta}}{\Delta x} \sim \mathcal{O}(\sqrt{\Delta x}),$$

whereas if both sides overstep then the error from interpolation dominates and is  $\mathcal{O}(1)$  for points  $\mathcal{O}(\Delta x)$  from the boundary, as seen at the end of the proof of Proposition 2.2.2.  $\square$

**Remark 2.2.2** (Two-sided overstepping). We note that it is possible for both sides of the stencil to overstep if the diffusion direction  $\sigma_p^{\alpha}$  is (almost) parallel to the domain boundary, for points close to a locally convex smooth boundary with high curvature in that direction, as well as close to corners; see Remark 2.2.5 and Table 2.2.9 below.

The scheme is consistent at points with two-sided overstepping if the truncated scheme is not interpolated at the boundary but uses the exact boundary values. In that case, the consistency error for those points is  $\mathcal{O}(\Delta x)$ .

### 2.2.3 Properties of the truncated stencil

The changes in the finite difference weights of scheme (2.2.3) introduced by the truncation, modify the positivity conditions given in Lemma 4.1 in [26]. We will show that the scheme remains conditionally  $L_\infty$ -stable and monotone, but the CFL conditions are more restrictive in the truncated case for time-stepping schemes with  $\theta < 1$ . We start by writing the scheme on a discrete time-space grid with mesh parameters  $\Delta t$  and  $\Delta x$  as

$$\begin{aligned}
& \hat{L}_{\Delta x}^\alpha [\mathcal{I}_{\Delta x} \phi(t, \cdot)](t_n, x_j) \\
&= \sum_{p=1}^M \frac{1}{2\Delta x} [A_p^{\alpha,n} (\mathcal{I}_{\Delta x} \phi(t_n, \cdot))(x_j + \hat{y}_p^{\alpha,+}) - (A_p^{\alpha,n} + B_p^{\alpha,n}) \phi(t_n, x_j) \\
&\quad + B_p^{\alpha,n} (\mathcal{I}_{\Delta x} \phi(t_n, \cdot))(x_j + \hat{y}_p^{\alpha,-})] \\
&= \sum_{p=1}^M \left\{ \sum_{i \in \mathcal{N}(x_j + \hat{y}_p^{\alpha,+})} \frac{1}{2\Delta x} [A_p^{\alpha,n} w_i(x_j + \hat{y}_p^{\alpha,+})] (\phi(t_n, x_i) - \phi(t_n, x_j)) + \right. \\
&\quad \left. \sum_{i \in \mathcal{N}(x_j + \hat{y}_p^{\alpha,-})} \frac{1}{2\Delta x} [B_p^{\alpha,n} w_i(x_j + \hat{y}_p^{\alpha,-})] (\phi(t_n, x_i) - \phi(t_n, x_j)) \right\} \\
&= \sum_{i=1}^N \sum_{p=1}^M \frac{A_p^{\alpha,n} w_i(x_j + \hat{y}_p^{\alpha,+}) + B_p^{\alpha,n} w_i(x_j + \hat{y}_p^{\alpha,-})}{2\Delta x} (\phi(t_n, x_i) - \phi(t_n, x_j)) \\
&= \sum_{i=1}^N \hat{l}_{j,i}^{\alpha,n} (\phi(t_n, x_i) - \phi(t_n, x_j)), \tag{2.2.13}
\end{aligned}$$

where  $\mathcal{N}$  is the set of interpolation points as in (1.1.9), and

$$\hat{l}_{j,i}^{\alpha,n} = \sum_{p=1}^M \frac{A_p^{\alpha,n} w_i(x_j + \hat{y}_p^{\alpha,+}(t_n, x_j)) + B_p^{\alpha,n} w_i(x_j + \hat{y}_p^{\alpha,-}(t_n, x_j))}{2\Delta x}.$$

In deriving (2.2.13) we have used, first the definition of the truncated scheme (2.2.3) and then the fact that the interpolation operator  $\mathcal{I}_{\Delta x}$  is monotone and therefore



satisfies that for all  $1 \leq i, j \leq N$

$$w_j(x) \geq 0, \quad w_i(x_j) = \delta_{ij}, \quad \text{and} \quad \sum_{i \in \mathcal{N}(x)} w_i(x) \equiv 1. \quad (2.2.14)$$

Here,

$$\sum_{i=1}^N \hat{l}_{j,i}^{\alpha,n} = \sum_{p=1}^M \frac{A_p^{\alpha,n} + B_p^{\alpha,n}}{2\Delta x} \geq \frac{M}{\Delta x},$$

with equality only in the absence of domain overstepping for all  $p \in [[1, M]]$  at  $(t_n, x_j, \alpha)$ .

Writing the overall scheme in the form (2.2.2) of Definition 2.2.1, we have that

$$\begin{aligned} \sup_{\alpha} \left\{ \left[ 1 + \theta \Delta t_n \left( \sum_{p=1}^M \frac{A_p^{\alpha,n} + B_p^{\alpha,n}}{2\Delta x} - \hat{l}_{j,j}^{\alpha,n} - c_j^{\alpha,n-1+\theta} \right) \right] U_j^n - \theta \Delta t_n \sum_{i \neq j} \hat{l}_{j,i}^{\alpha,n} U_i^n + \right. \\ \left. - \left[ 1 - (1 - \theta) \Delta t_n \left( \sum_{p=1}^M \frac{A_p^{\alpha,n-1} + B_p^{\alpha,n-1}}{2\Delta x} - \hat{l}_{j,j}^{\alpha,n-1} - c_j^{\alpha,n-1+\theta} \right) \right] U_j^{n-1} + \right. \\ \left. - (1 - \theta) \Delta t_n \sum_{i \neq j} \hat{l}_{j,i}^{\alpha,n-1} U_i^{n-1} - \Delta t_n f_j^{\alpha,n-1+\theta} \right\} = 0. \end{aligned} \quad (2.2.15)$$

It is straightforward to write down the expressions for the coefficients in (2.2.2):

$$\begin{aligned} \mathcal{B}_{j,j}^{\alpha,n,n} &= 1 + \theta \Delta t_n \left( \sum_{p=1}^M \frac{A_p^{\alpha,n} + B_p^{\alpha,n}}{2\Delta x} - \hat{l}_{j,j}^{\alpha,n} - c_j^{\alpha,n-1+\theta} \right), \\ \mathcal{B}_{j,j}^{\alpha,n,n-1} &= 1 - (1 - \theta) \Delta t_n \left( \sum_{p=1}^M \frac{A_p^{\alpha,n-1} + B_p^{\alpha,n-1}}{2\Delta x} - \hat{l}_{j,j}^{\alpha,n-1} - c_j^{\alpha,n-1+\theta} \right), \\ \mathcal{B}_{j,i}^{\alpha,n,n} &= \theta \Delta t_n \hat{l}_{j,i}^{\alpha,n}, \quad \mathcal{B}_{j,i}^{\alpha,n,n-1} = (1 - \theta) \Delta t_n \hat{l}_{j,i}^{\alpha,n-1}. \end{aligned}$$

**Remark 2.2.3.** In writing down (2.2.13), we assumed that the value at the boundary is interpolated from other mesh points, which is feasible on rectangular cuboids, but not for general domain boundaries. In both cases, the Dirichlet boundary value at  $x_j + \hat{y}_p^{\alpha,\pm}$  can be used. This has the advantage that interpolation error is avoided.

Moreover, as this value then contributes to the right-hand-side  $f$  of equation (2.2.15) instead of the off-diagonal matrix elements, the system matrix becomes more diagonally dominant. This is advantageous for the iterative solution, see Section 4.5.

The next proposition contains the positivity conditions for the coefficients  $\mathcal{B}$  defined above.

**Proposition 2.2.4.** *The scheme (2.2.15) is of positive type if the following conditions hold,*

$$(1 - \theta)\Delta t_n \left[ \sum_{p=1}^M \frac{A_p^{\alpha,n-1} + B_p^{\alpha,n-1}}{2\Delta x} - c_i^{\alpha,n-1+\theta} \right] \leq 1, \quad \text{and} \quad \theta\Delta t_n c_i^{\alpha,n-1+\theta} \leq 1, \quad (2.2.16)$$

for all  $\alpha, n, i$ .

**Corollary 2.2.5.** *In the case of overstepping and  $\theta < 1$ , monotonicity requires that  $\Delta t \leq C_1\Delta x^{3/2}$  if only one side of the diffusion stencils oversteps, or  $\Delta t \leq C_2\Delta x^2$  if both sides overstep. However, if the stencil is not truncated, the positivity condition remains as in [26], that is  $\Delta t \leq C_3\Delta x$ . Where  $C_1, C_2, C_3 > 0$  are sufficiently small constants depending of the coefficients  $\sigma^\alpha, b^\alpha$  and  $c^\alpha$ , but independent of  $\Delta x$  and  $\Delta t$ .*

*Proof.* From Corollary 2.2.3, if the corresponding stencil is truncated on one side  $A_{\cdot}^{\alpha,n-1} + B_{\cdot}^{\alpha,n-1} \sim \mathcal{O}(\Delta x^{-1/2})$  for sufficiently small  $\Delta x$ ,  $A_{\cdot}^{\alpha,n-1} + B_{\cdot}^{\alpha,n-1} \sim \mathcal{O}(\Delta x^{-1})$  if both sides are truncated, whereas if there is no overstepping,  $A_{\cdot}^{\alpha,n-1} + B_{\cdot}^{\alpha,n-1} \sim \mathcal{O}(1)$ .  $\square$

The  $L^\infty$ -stability follows from the proof of Lemma 4.1 in [26] and the new CFL conditions in Proposition 2.2.4.

### 2.2.4 Numerical experiments

To test the truncation of the stencil, we consider Problems A and B in Section 9.3 from [26]. Both problems follow the formulation in (1.1.1)–(1.1.3) with homogeneous Dirichlet boundary conditions and have smooth solutions.

**Problem A** (see Section 9.3 from [26]). It has exact solution

$$u(t, x_1, x_2) = \left(\frac{3}{2} - t\right) \sin x_1 \sin x_2,$$

and coefficients and control set are given by

$$\begin{aligned} f &= \left(\frac{1}{2} - t\right) \sin x_1 \sin x_2 + \left(\frac{3}{2} - t\right) \left[ \sqrt{\cos^2 x_1 \sin^2 x_2 + \sin^2 x_1 \cos^2 x_2} + \right. \\ &\quad \left. - 2 \sin(x_1 + x_2) \cos(x_1 + x_2) \cos x_1 \cos x_2 \right], \\ c^\alpha &= 0, \quad b^\alpha = \alpha, \quad \sigma = \sqrt{2} \begin{pmatrix} \sin(x_1 + x_2) \\ \cos(x_1 + x_2) \end{pmatrix}, \quad \mathcal{A} = \{\alpha \in \mathbb{R}^2 : \alpha_1^2 + \alpha_2^2 = 1\}. \end{aligned}$$

The resulting equation is

$$\begin{aligned} u_t - \frac{1}{2} \text{tr}[\sigma \sigma^T D^2 u] - \inf_{\alpha \in \mathcal{A}} \{\alpha D u\} &= f, & (t, x_1, x_2) &\in (0, T] \times \Omega, \\ u(0, x_1, x_2) &= \frac{3}{2} \sin x_1 \sin x_2, & (x_1, x_2) &\in \bar{\Omega}, \\ u(t, x_1, x_2) &= 0, & (t, x_1, x_2) &\in (0, T] \times \partial\Omega. \end{aligned}$$

**Problem B** (see Section 9.3 from [26]). It has exact solution

$$u(t, x_1, x_2) = (2 - t) \sin(x_1) \sin(x_2),$$

and coefficients and control set

$$f^\alpha = (1 - t) \sin x_1 \sin x_2 - 2\alpha_1 \alpha_2 (2 - t) \cos x_1 \cos x_2,$$

$$c^\alpha = 0, \quad b^\alpha = 0, \quad \sigma^\alpha = \sqrt{2} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}, \quad \mathcal{A} = \{\alpha \in \mathbb{R}^2 : \alpha_1^2 + \alpha_2^2 = 1\}.$$

The resulting equation is

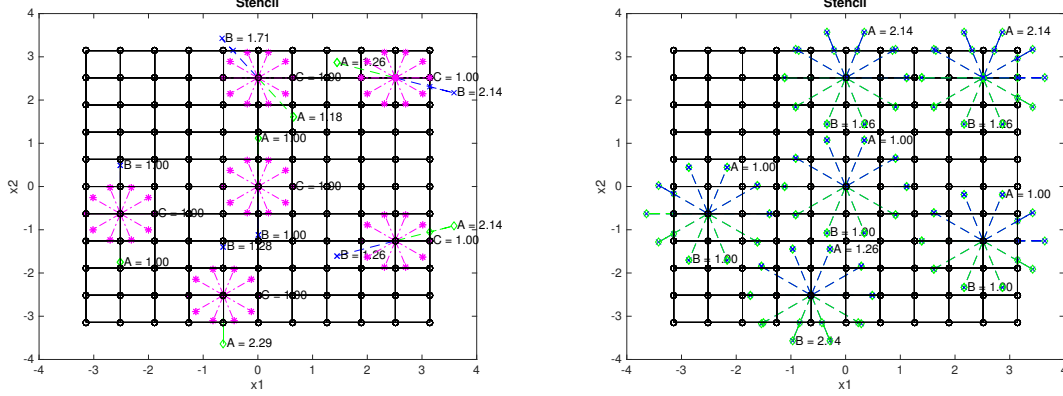
$$u_t - \inf_{\alpha \in \mathcal{A}} \left\{ \frac{1}{2} \text{tr}[\sigma \sigma^T D^2 u] + f^\alpha \right\} = 0, \quad (t, x_1, x_2) \in (0, T] \times \Omega,$$

$$u(0, x_1, x_2) = 2 \sin x_1 \sin x_2, \quad (x_1, x_2) \in \bar{\Omega},$$

$$u(t, x_1, x_2) = 0, \quad (t, x_1, x_2) \in (0, T] \times \partial\Omega,$$

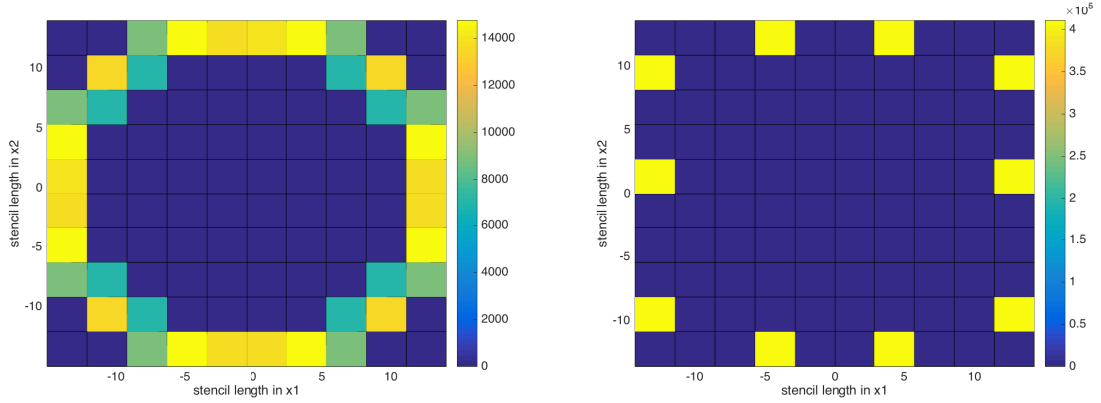
Both problems are solved on the domain  $(t, x_1, x_2) \in [0, T] \times [-\pi, \pi]^2$  with  $T = \frac{1}{2}$ . We discretize the spatial domain using Cartesian grids with  $N_x \times N_x$  equispaced nodes and for the control set  $\mathcal{A}$  we take  $N_\alpha$  equally spaced points. Here,  $\mathcal{I}_{\Delta x}$  is the usual bilinear interpolator on rectangles.

For illustration of the stencil and its non-locality, the top row of Figure 2.2.2 represents the stencil for Problems A and B on a Cartesian grid of  $11 \times 11$  points and 10 points in the control set  $\mathcal{A}$ . Colour coded lines link the stencil points with the node where the numerical solution is computed, the different colours correspond to the different  $\hat{y}^{\alpha_i}$ . On top of some of the stencil points we print the value of the finite difference weights, for compactness we set  $A \equiv A_{2,1}^\alpha(x)$ ,  $B \equiv B_{2,1}^\alpha(x)$  and  $C \equiv (\mu_{2,2}^\alpha(x))^{-1}$ , following the notation in (2.2.7) and (2.2.6). The bottom row of Figure 2.2.2 represents the non-locality of the diffusion stencil by counting the number of stencil points at a given distance from the central node. The distance is measured in multiples of  $\Delta x$  and given by  $\left\lfloor \frac{(\sigma^\alpha(x))_i}{\sqrt{\Delta x}} \right\rfloor$ , where the grid is of size  $641 \times 641$  and 10 points in the control set  $\mathcal{A}$ .



(a) Stencil for Problem A in [26] on a Cartesian  $11 \times 11$  grid for 10 sample points in the control set  $\mathcal{A}$ .

(b) Stencil for Problem B in [26] on a Cartesian  $11 \times 11$  grid for 10 sample points in the control set  $\mathcal{A}$ .



(c) Histogram of  $\left\lfloor \frac{(\sigma^\alpha(x))_i}{\sqrt{\Delta x}} \right\rfloor$  in Problem A for all  $x \in \Omega_{\Delta x}$  where  $\Omega_{\Delta x}$  is a Cartesian grid with  $\Delta x = \frac{2\pi}{640}$ , 10 points in the control set  $\mathcal{A}$ , and  $i \in \{1, 2\}$  is the dimension index.

(d) Histogram of  $\left\lfloor \frac{(\sigma^\alpha(x))_i}{\sqrt{\Delta x}} \right\rfloor$  in Problem B for all  $x \in \Omega_{\Delta x}$  where  $\Omega_{\Delta x}$  is a Cartesian grid with  $\Delta x = \frac{2\pi}{640}$ , 10 points in the control set  $\mathcal{A}$ , and  $i \in \{1, 2\}$  is the dimension index.

Figure 2.2.2: Graphical representation of the stencil over a two-dimensional Cartesian grid of size  $11 \times 11$  and 10 equally spaced points from the control set  $\mathcal{A}$ . The finite difference weights corresponding to some of the points are printed, for simplicity the weights are labelled  $A \equiv A_{2,1}^\alpha(x)$ ,  $B \equiv B_{2,1}^\alpha(x)$  and  $C \equiv (\mu_{2,2}^\alpha(x))^{-1}$ , following the notation in (2.2.7) and (2.2.6). To illustrate the non-locality of the scheme as the grid is refined, the second row represents the histograms of the shortest displacement from the central node for a grid of size  $641 \times 641$  for both problems. The radius of the stencil in  $\sigma^\alpha$  is 14.27 for this grid, given by  $\frac{\|\sigma^\alpha\|_2}{\sqrt{\Delta x}} = \sqrt{640/\pi}$ .

Problems A and B were obviously chosen in [26] for their periodic solutions, to be able to analyse the convergence of the scheme without the complication of boundary conditions. Here, we do not make use of the periodicity but only use the values at the boundary and not outside the domain.

We note that the problems being linear in  $t$ , a single time step with  $\Delta t = T$  suffices to obtain an exact solution in  $t$ . However, in order to check the effect of the truncation on the stability, in addition to  $\Delta t = T$ , we also investigate  $\Delta t$  equal to  $\frac{\Delta x}{4}$ ,  $\Delta x^{3/2}$ , and  $\Delta x^2$ . We report the  $\infty$ -norm of the errors over two regions: the first one comprising the whole domain, and the second one comprising part of the interior of the domain.

We consider explicit and implicit time stepping schemes, corresponding to  $\theta = 0$  and  $\theta = 1$  respectively. For the explicit scheme in the case of overstepping we test the following modifications of the scheme:

1. truncation of the stencil as discussed in Section 2.2.2 (Table 2.2.1 for Problem A and Table 2.2.5 for Problem B);
2. constant extrapolation of the boundary condition using the value at the nearest point in the boundary (Table 2.2.2 for Problem A and Table 2.2.6 for Problem B);
3. linear extrapolation of the function for points overstepping the domain (Table 2.2.3 for Problem A and Table 2.2.7 for Problem B). This means that if  $x + y_{2,p}^\alpha \notin \bar{\Omega}$ , then

$$\phi(t, x + y_{2,p}^\alpha) = \phi(t, z) + D\phi(t, z)(x + y_{2,p}^\alpha - z),$$

where  $z := \operatorname{argmin}_{z \in \Omega} \|x + y_{2,p}^\alpha - z\|_2$ .

For the implicit case we only consider the first modification, i.e. truncation of the

stencil (Table 2.2.4 for Problem A and Table 2.2.8 for Problem B).

The results confirm the impact of the truncation on the stability of the scheme, when  $\theta = 0$ . However, when  $\theta = 1$ , we do not observe any instability regardless of the size of the time step. When stable, the truncation of the stencil outperforms the two extrapolations of the boundary conditions considered. Furthermore, as the mesh and time steps are refined, only the truncated scheme, if stable, achieves convergence orders close to  $\mathcal{O}(\Delta x)$  when the error at  $t = T$  is measured on the entire spatial grid. This can be explained without rigorous proof by the observation that the truncation error of order  $\sqrt{\Delta x}$  is restricted to a boundary layer of width  $\sqrt{\Delta x}$ . Therefore, as seen from the last two columns in Table 2.2.4, choosing  $\Delta t$  of order higher than 1 in  $\Delta x$  does not improve the accuracy of the numerical results and leads to computational inefficiency.

**Remark 2.2.4.** Regarding the discretization of the control set, we take  $N_\alpha = 40$  equally spaced points as, for the problems and space-time mesh sizes considered, the discretization error of the LISL scheme dominates. It is to be expected that if we continued to refine the space-time mesh keeping  $N_\alpha$  constant, then the error in the control discretization would end up dominating, stalling the error convergence of the method. A more efficient way to discretize the control set is described in [29].

**Remark 2.2.5.** Corollary 2.2.5 shows two different CFL conditions for the truncated stencil, the first one for diffusion stencils where only one side oversteps and a second one when both sides overstep. The results in Table 2.2.1 for Problem A and Table 2.2.5 for Problem B correspond to the former situation. To check the sharpness of the latter, we shift the spatial domain in Problem A in both directions by  $\frac{7\pi}{8}$ . The new spatial domain is thus  $\bar{\Omega} = [-\frac{\pi}{8}, \frac{15\pi}{8}]^2$ . Note that the solution itself is periodic with period  $2\pi$ . This problem differs from the original one in that it has non-homogeneous Dirichlet boundary conditions and that both sides of the diffusion stencil overstep

(a) Error in  $L^\infty$ -norm over  $\Omega_{\Delta x}$

$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
41	1.42e-01	-	4.39e-02	-	4.39e-02	-	4.36e-02	-
81	1.04e-01	0.45	2.12e-02	1.05	2.11e-02	1.06	2.11e-02	1.05
161	7.36e-02	0.50	1.10e-02	0.94	1.10e-02	0.94	1.10e-02	0.94
321	5.28e-02	0.48	1.34e+23	-83.33	5.77e-03	0.93	5.76e-03	0.93
641	3.77e-02	0.48	5.07e+89	-221.17	3.10e-03	0.90	3.10e-03	0.89

(b) Error in  $L^\infty$ -norm over  $\Omega_{\Delta x} \cap [-\pi/2, \pi/2]^2$

$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
41	8.61e-02	-	4.38e-02	-	4.42e-02	-	4.35e-02	-
81	4.22e-02	1.03	2.12e-02	1.05	2.11e-02	1.06	2.11e-02	1.05
161	2.14e-02	0.98	1.10e-02	0.94	1.10e-02	0.95	1.10e-02	0.94
321	1.10e-02	0.96	1.84e+13	-50.57	5.71e-03	0.95	5.70e-03	0.95
641	5.96e-03	0.88	1.06e+72	-195.20	3.08e-03	0.89	3.08e-03	0.89

Table 2.2.1: Results using the truncation of the stencil for explicit method with  $N_\alpha = 40$  for Problem A.

(a) Error in  $L^\infty$ -norm over  $\Omega_{\Delta x}$

$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
41	1.36e+00	-	3.68e-01	-	3.72e-01	-	3.65e-01	-
81	1.89e+00	-0.48	2.61e-01	0.49	2.62e-01	0.51	2.60e-01	0.49
161	2.67e+00	-0.49	1.80e-01	0.54	1.80e-01	0.54	1.80e-01	0.53
321	3.77e+00	-0.50	1.27e-01	0.51	1.27e-01	0.51	1.27e-01	0.51
641	5.34e+00	-0.50	9.18e-02	0.47	9.18e-02	0.47	9.18e-02	0.46

(b) Error in  $L^\infty$ -norm over  $\Omega_{\Delta x} \cap [-\pi/2, \pi/2]^2$

$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
41	1.59e-01	-	1.04e-01	-	1.05e-01	-	1.03e-01	-
81	8.15e-02	0.96	5.25e-02	0.99	5.26e-02	1.00	5.22e-02	0.98
161	4.22e-02	0.95	2.67e-02	0.98	2.66e-02	0.98	2.66e-02	0.97
321	2.18e-02	0.95	1.36e-02	0.97	1.36e-02	0.97	1.36e-02	0.97
641	1.21e-02	0.85	8.21e-03	0.73	8.20e-03	0.73	8.19e-03	0.73

Table 2.2.2: Results using constant extrapolation of the boundary condition for explicit method with  $N_\alpha = 40$  for Problem A.



(a) Error in $L^\infty$ -norm over $\Omega_{\Delta x}$								
$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
41	1.59e-01	-	1.04e-01	-	1.05e-01	-	1.03e-01	-
81	8.15e-02	0.96	5.25e-02	0.99	5.26e-02	1.00	5.22e-02	0.98
161	4.28e-02	0.93	5.62e-01	-3.42	5.63e-01	-3.42	5.58e-01	-3.42
321	2.75e-02	0.64	4.41e+03	-12.94	6.00e+03	-13.38	8.00e+03	-13.81
641	1.85e-02	0.57	2.77e+20	-55.80	2.70e+20	-55.32	1.37e+21	-57.25

(b) Error in $L^\infty$ -norm over $\Omega_{\Delta x} \cap [-\pi/2, \pi/2]^2$								
$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
41	1.59e-01	-	1.04e-01	-	1.05e-01	-	1.03e-01	-
81	8.15e-02	0.96	5.25e-02	0.99	5.26e-02	1.00	5.22e-02	0.98
161	4.22e-02	0.95	2.67e-02	0.98	2.66e-02	0.98	2.66e-02	0.97
321	2.18e-02	0.95	1.96e+00	-6.20	2.07e+00	-6.28	2.23e+00	-6.39
641	1.21e-02	0.85	9.26e+14	-48.75	3.18e+15	-50.45	3.01e+15	-50.26

Table 2.2.3: Results using linear extrapolation for points out of the domain for explicit method with  $N_\alpha = 40$  for Problem A.

(a) Error in $L^\infty$ -norm over $\Omega_{\Delta x}$								
$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
41	3.25e-02	-	4.21e-02	-	4.17e-02	-	4.24e-02	-
81	1.59e-02	1.03	2.08e-02	1.02	2.08e-02	1.01	2.09e-02	1.02
161	8.39e-03	0.92	1.09e-02	0.93	1.09e-02	0.93	1.10e-02	0.93
321	4.38e-03	0.94	5.75e-03	0.93	5.75e-03	0.93	5.76e-03	0.93
641	2.37e-03	0.89	3.09e-03	0.89	3.10e-03	0.89	3.10e-03	0.89

(b) Error in $L^\infty$ -norm over $\Omega_{\Delta x} \cap [-\pi/2, \pi/2]^2$								
$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
41	3.25e-02	-	4.21e-02	-	4.17e-02	-	4.24e-02	-
81	1.59e-02	1.03	2.08e-02	1.02	2.08e-02	1.01	2.09e-02	1.02
161	8.39e-03	0.92	1.09e-02	0.93	1.09e-02	0.93	1.10e-02	0.93
321	4.35e-03	0.95	5.68e-03	0.94	5.69e-03	0.94	5.70e-03	0.95
641	2.37e-03	0.88	3.07e-03	0.89	3.08e-03	0.89	3.08e-03	0.89

Table 2.2.4: Results using truncation for points out of the domain for implicit method with  $N_\alpha = 40$  for Problem A.

(a) Error in  $L^\infty$ -norm over  $\Omega_{\Delta x}$

$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
41	1.73e-01	-	3.91e-02	-	3.95e-02	-	3.88e-02	-
81	1.39e-01	0.32	1.84e-02	1.09	1.84e-02	1.10	1.83e-02	1.09
161	1.07e-01	0.38	8.71e-03	1.08	8.70e-03	1.08	8.68e-03	1.07
321	8.05e-02	0.41	1.39e+43	-150.16	4.12e-03	1.08	4.11e-03	1.08
641	5.95e-02	0.44	1.77e+153	-365.76	2.17e-03	0.92	2.17e-03	0.92

(b) Error in  $L^\infty$ -norm over  $\Omega_{\Delta x} \cap [-\pi/2, \pi/2]^2$

$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
41	5.71e-02	-	3.91e-02	-	3.95e-02	-	3.88e-02	-
81	2.74e-02	1.06	1.84e-02	1.09	1.84e-02	1.10	1.83e-02	1.09
161	1.31e-02	1.06	8.71e-03	1.08	8.70e-03	1.08	8.68e-03	1.07
321	6.57e-03	0.99	8.34e+28	-102.92	4.12e-03	1.08	4.11e-03	1.08
641	3.28e-03	1.00	1.09e+127	-325.93	2.17e-03	0.92	2.17e-03	0.92

Table 2.2.5: Results using stencil truncation for explicit method with  $N_\alpha = 40$  for Problem B.

(a) Error in $L^\infty$ -norm over $\Omega_{\Delta x}$								
$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
41	1.25e+00	-	3.79e-01	-	3.82e-01	-	3.75e-01	-
81	1.99e+00	-0.67	3.55e-01	0.09	3.55e-01	0.11	3.53e-01	0.09
161	3.04e+00	-0.61	2.92e-01	0.28	2.92e-01	0.28	2.92e-01	0.27
321	4.52e+00	-0.57	2.35e-01	0.32	2.35e-01	0.32	2.35e-01	0.31
641	6.62e+00	-0.55	1.77e-01	0.41	1.77e-01	0.41	1.77e-01	0.41

(b) Error in $L^\infty$ -norm over $\Omega_{\Delta x} \cap [-\pi/2, \pi/2]^2$								
$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
41	5.71e-02	-	6.38e-02	-	6.34e-02	-	6.40e-02	-
81	2.74e-02	1.06	5.72e-02	0.16	5.72e-02	0.15	5.68e-02	0.17
161	1.31e-02	1.06	4.51e-02	0.34	4.51e-02	0.34	4.49e-02	0.34
321	6.57e-03	0.99	3.71e-02	0.28	3.71e-02	0.28	3.70e-02	0.28
641	3.28e-03	1.00	2.89e-02	0.36	2.89e-02	0.36	2.88e-02	0.36

Table 2.2.6: Results using constant extrapolation of the boundary condition for explicit method with  $N_\alpha = 40$  for Problem B.

(a) Error in $L^\infty$ -norm over $\Omega_{\Delta x}$								
$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
41	5.71e-02	-	8.46e-02	-	8.29e-02	-	8.60e-02	-
81	3.12e-02	0.87	2.43e+02	-11.49	1.82e+02	-11.10	1.67e+03	-14.25
161	2.89e-02	0.11	7.90e+18	-54.85	8.95e+20	-62.10	1.64e+31	-92.99
321	2.38e-02	0.28	1.51e+70	-170.36	9.26e+93	-242.55	1.51e+164	-441.69
641	1.87e-02	0.35	1.14e+207	-454.70	NaN	NaN	NaN	NaN

(b) Error in $L^\infty$ -norm over $\Omega_{\Delta x} \cap [-\pi/2, \pi/2]^2$								
$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
41	5.71e-02	-	3.91e-02	-	3.95e-02	-	3.88e-02	-
81	2.74e-02	1.06	1.84e-02	1.09	1.84e-02	1.10	5.19e-02	-0.42
161	1.31e-02	1.06	1.36e+09	-36.10	1.54e+11	-42.92	2.82e+21	-75.52
321	6.57e-03	0.99	5.30e+52	-144.81	3.24e+76	-217.00	5.27e+146	-416.15
641	3.28e-03	1.00	6.39e+176	-412.19	NaN	NaN	NaN	NaN

Table 2.2.7: Results using linear extrapolation for points out of the domain for explicit method with  $N_\alpha = 40$  for Problem B.

(a) Error in $L^\infty$ -norm over $\Omega_{\Delta x}$								
$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
41	3.00e-02	-	3.76e-02	-	3.72e-02	-	3.79e-02	-
81	1.40e-02	1.10	1.80e-02	1.06	1.80e-02	1.05	1.45e-02	1.38
161	6.34e-03	1.15	6.36e-03	1.50	6.37e-03	1.50	7.72e-03	0.91
321	3.04e-03	1.06	3.38e-03	0.91	3.50e-03	0.86	3.01e-03	1.36
641	1.53e-03	0.99	1.77e-03	0.93	1.76e-03	1.00	1.66e-03	0.85

(b) Error in $L^\infty$ -norm over $\Omega_{\Delta x} \cap [-\pi/2, \pi/2]^2$								
$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
41	3.00e-02	-	3.76e-02	-	3.72e-02	-	3.79e-02	-
81	1.40e-02	1.10	1.80e-02	1.06	1.80e-02	1.05	1.45e-02	1.38
161	6.34e-03	1.15	6.31e-03	1.51	6.37e-03	1.50	7.72e-03	0.91
321	3.04e-03	1.06	3.38e-03	0.90	3.50e-03	0.86	3.01e-03	1.36
641	1.53e-03	0.99	1.77e-03	0.93	1.76e-03	1.00	1.66e-03	0.85

Table 2.2.8: Results using truncation for points out of the domain for implicit method with  $N_\alpha = 40$  for Problem B.

for mesh points within a distance of  $\mathcal{O}(\sqrt{\Delta x})$  to the bottom left corner, located at  $(\frac{-\pi}{8}, \frac{-\pi}{8})$ , where  $\sigma^\alpha = (-1, 1)^T$ . In Table 2.2.9 we report the results for the explicit method using the truncation of the stencil. As expected, we find that we now need  $\Delta t \sim \Delta x^2$  for stability.

(a) Error in $L^\infty$ -norm over $\Omega_{\Delta x} \cap [-\pi/8, 15\pi/8]^2$								
$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
41	1.55e-01	-	4.71e-02	-	4.76e-02	-	4.67e-02	-
81	1.12e-01	0.47	1.57e+05	-21.67	7.90e+05	-23.98	2.11e-02	1.15
161	8.04e-02	0.47	1.02e+33	-92.39	1.30e+35	-97.06	1.10e-02	0.94
321	5.80e-02	0.47	6.73e+103	-235.26	5.96e+138	-344.35	5.76e-03	0.93
641	4.22e-02	0.46	8.17e+276	-574.97	NaN	NaN	3.10e-03	0.89

(b) Error in $L^\infty$ -norm over $\Omega_{\Delta x} \cap [3\pi/8, 11\pi/8]^2$								
$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
41	8.65e-02	-	4.70e-02	-	4.74e-02	-	4.66e-02	-
81	4.22e-02	1.04	2.07e-02	1.18	2.07e-02	1.19	2.06e-02	1.18
161	2.14e-02	0.98	1.18e+06	-25.76	1.18e+09	-35.73	1.08e-02	0.93
321	1.10e-02	0.96	7.99e+47	-138.96	4.94e+84	-251.21	5.59e-03	0.95
641	5.96e-03	0.88	9.81e+165	-392.28	NaN	NaN	3.02e-03	0.89

Table 2.2.9: Results using the truncation of the stencil for explicit method with  $N_\alpha = 40$  for Problem A on shifted domain.

**Remark 2.2.6.** For the explicit method using the truncation of the stencil, i.e. Tables 2.2.1, 2.2.5 and 2.2.9, focusing on the  $\Delta t = T$  case, we notice that the error convergence rate over the whole mesh  $\Omega_{\Delta x}$  is approximately 0.5, whereas it is approximately 1.0 when the error is measured in the interior of the mesh. We also notice that there is a significant difference between the magnitude of the errors if measured over the whole grid or on a region in the interior. The difference in the magnitude of the errors may be due to the fact that  $\Delta t = T$  does not satisfy the CFL condition and that the CFL condition is more restrictive for points where the stencil is truncated. It is also at these points that the local consistency error is of order  $\sqrt{\Delta x}$  as shown in

Corollary 2.2.3. The situation is different for  $\Delta t = \Delta x^2$  in the explicit case, or for any  $\Delta t$  in the implicit case. In these cases, the error convergence rates are approximately 1.0 when measured over the whole mesh and the errors over the whole grid and in the interior are comparable in magnitude.

For the last example we consider a linear problem in a circular domain with a structured grid.

**Problem C** It has exact solution

$$u(t, x_1, x_2) = \left(\frac{3}{2} - t\right) \sin x_1 \sin x_2,$$

the coefficients are

$$\begin{aligned} f &= \left(\frac{1}{2} - t\right) \sin x_1 \sin x_2 - (3 - 2t) \sin(x_1 + x_2) \cos(x_1 + x_2) \cos x_1 \cos x_2, \\ c &= 0, \quad b = 0, \quad \sigma = \sqrt{2} \begin{pmatrix} \sin(x_1 + x_2) \\ \cos(x_1 + x_2) \end{pmatrix}. \end{aligned}$$

The problem to solve is therefore

$$\begin{aligned} u_t - \frac{1}{2} \text{tr}[\sigma \sigma^T D^2 u] &= f, & (t, x_1, x_2) &\in (0, T] \times \Omega, \\ u(0, x_1, x_2) &= \frac{3}{2} \sin x_1 \sin x_2, & (x_1, x_2) &\in \bar{\Omega}, \\ u(t, x_1, x_2) &= \left(\frac{3}{2} - t\right) \sin x_1 \sin x_2, & (t, x_1, x_2) &\in (0, T] \times \partial\Omega. \end{aligned}$$

We solve this equation on  $(t, x_1, x_2) \in [0, T] \times \bar{\Omega}$  with  $T = \frac{1}{2}$  and  $\Omega := \{x \in \mathbb{R}^2 : x_1^2 + x_2^2 < \pi^2\}$ .

The spatial grid is a structured grid constructed as in Figure 2.2.1. First, we discretise the square where the circle is inscribed with parameter  $\Delta x > 0$ . The final mesh is made up from all the nodes in the interior of the circle that lie at least  $\Delta x$

away from the circle's boundary. Tables 2.2.10 and 2.2.11 contain the convergence results for the scheme with the truncated stencil using explicit and implicit time stepping schemes. The results are as expected from the theory. We highlight that the explicit time stepping scheme with  $\Delta t = T$  (first column in Table 2.2.10) has stability issues. Indeed the error is higher for  $\Delta x = 2\pi/640$  than for  $\Delta x = 2\pi/320$ .

(a) Error in  $L^\infty$ -norm over  $\Omega_{\Delta x}$

$\Delta_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
$2\pi/40$	3.13e-01	-	1.81e-01	-	1.84e-01	-	1.79e-01	-
$2\pi/80$	2.61e-01	0.26	9.39e-02	0.95	9.41e-02	0.97	9.27e-02	0.95
$2\pi/160$	2.05e-01	0.35	4.60e-02	1.03	4.59e-02	1.04	4.57e-02	1.02
$2\pi/320$	1.05e-01	0.96	1.16e+12	-44.52	1.05e-02	2.12	1.05e-02	2.12
$2\pi/640$	1.32e-01	-0.32	4.73e+77	-217.96	6.18e-03	0.77	6.18e-03	0.77

(b) Error in  $L^\infty$ -norm over  $\Omega_{\Delta x} \cap [-\pi/2, \pi/2]^2$

$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
$2\pi/40$	3.82e-02	-	2.23e-02	-	2.26e-02	-	2.20e-02	-
$2\pi/80$	1.87e-02	1.03	1.10e-02	1.02	1.10e-02	1.04	1.09e-02	1.02
$2\pi/160$	9.63e-03	0.96	5.28e-03	1.06	5.27e-03	1.06	5.24e-03	1.05
$2\pi/320$	4.71e-03	1.03	1.36e-01	-4.69	2.64e-03	1.00	2.63e-03	0.99
$2\pi/640$	2.37e-03	0.99	9.12e+60	-205.38	1.32e-03	1.00	1.31e-03	1.00

Table 2.2.10: Results using the truncation of the stencil for explicit method for Problem C.

(a) Error in $L^\infty$ -norm over $\Omega_{\Delta x}$								
$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
$2\pi/40$	1.16e-01	-	1.69e-01	-	1.67e-01	-	1.71e-01	-
$2\pi/80$	6.04e-02	0.94	9.05e-02	0.90	9.03e-02	0.88	9.16e-02	0.90
$2\pi/160$	3.21e-02	0.91	4.54e-02	0.99	4.55e-02	0.99	4.56e-02	1.00
$2\pi/320$	8.22e-03	1.96	1.05e-02	2.11	1.05e-02	2.11	1.05e-02	2.11
$2\pi/640$	5.59e-03	0.56	6.18e-03	0.77	6.17e-03	0.77	6.18e-03	0.77

(b) Error in $L^\infty$ -norm over $\Omega_{\Delta x} \cap [-\pi/2, \pi/2]^2$								
$N_x$	$\Delta t = T$		$\Delta t = \frac{\Delta x}{4}$		$\Delta t = \Delta x^{\frac{3}{2}}$		$\Delta t = \Delta x^2$	
	error	rate	error	rate	error	rate	error	rate
$2\pi/40$	1.60e-02	-	2.09e-02	-	2.07e-02	-	2.12e-02	-
$2\pi/80$	7.49e-03	1.09	1.06e-02	0.98	1.06e-02	0.96	1.08e-02	0.98
$2\pi/160$	3.56e-03	1.07	5.19e-03	1.04	5.20e-03	1.03	5.23e-03	1.04
$2\pi/320$	1.82e-03	0.97	2.62e-03	0.99	2.62e-03	0.99	2.63e-03	0.99
$2\pi/640$	9.07e-04	1.00	1.31e-03	1.00	1.31e-03	1.00	1.31e-03	1.00

Table 2.2.11: Results using the truncation of the stencil for implicit method for Problem C.

## 2.3 Domain transformations for the Black-Scholes equation

Models for common quantities in finance, e.g. the spot price of a stock, or the short-term interest rate, consist of diffusion processes valued on a (semi) infinite domain. The price of contingent claims - derivatives - on these quantities, commonly called underlyings, is often given by a parabolic PDE whose analytic solution is generally unknown.

The price is a function of time and the value of the underlyings, which are also referred as the spatial variables. Depending on the nature of the problem the parabolic PDE could be linear, as in the case of the Black and Scholes model for option pricing [11], or non-linear of Bellman type, as in the case of option pricing under uncertain volatility [4, 55]. The analytic solution to these equations being generally unknown, we make use of numerical techniques to compute an approximate solution.

When pricing derivatives via finite differences, a common approach is to truncate the infinite domain and apply the numerical method on the truncated (bounded) one [6, 5]. As a result of the domain truncation, and to ensure the uniqueness of the solution, the introduction of artificial boundary conditions is required. We refer to [83] for an account of a boundary condition generally employed in Black-Scholes like equations after the domain truncation. Such domain truncation modifies the original problem, hence it may be a source of errors, as shown in [19, 6, 5], and may even introduce instabilities, see [83]. However, as highlighted in [6, 5], at least theoretically the errors introduced by the artificial boundary conditions are small on interior regions far from the artificial boundaries.

An alternative approach is to transform the (semi) infinite domain into a bounded domain, avoiding the truncation and the addition of artificial boundary conditions. This is the approach followed in [85] when solving pricing problems for linear parabolic PDEs. The transformations in [85] have the peculiarity that some coefficients of the parabolic PDE vanish resulting in a degeneracy, i.e. the equation changes its character.

Such degeneracies may pose theoretical difficulties and affect the behaviour of the solution at the boundary. The former can be dealt with considering solutions in the viscosity sense, whereas the latter was briefly discussed in Section 2.1.

In this section we are interested in problems without regular boundary points. In the next section we review the remarks about semi-Lagrangian schemes for the Black-Scholes equation in item 3. in Section 6.2 of [26] and show how to avoid the stencil truncation by using domain transformations as in [85]. We then compare the domain transformation to the stencil truncation for a linear two-dimensional problem. Finally, we show that the results in Section 2.3.1 can be extended to the non-linear case.



### 2.3.1 Boundary regularity for the Black-Scholes equation

In the remark about how to treat boundary conditions in Section 6.2 of [26], it is mentioned that the Black-Scholes equation posed on the spatial domain  $[0, \infty)^d$  has no regular points. For illustration purposes, consider the two dimensional Black-Scholes equation. The underlying stochastic processes for this equation follow the following SDE

$$d \begin{pmatrix} S_t^{(1)} \\ S_t^{(2)} \end{pmatrix} = b(t, S_t^{(1)}, S_t^{(2)}) dt + \sigma(t, S_t^{(1)}, S_t^{(2)}) \cdot d \begin{pmatrix} B_t \\ Z_t \end{pmatrix}, \quad (2.3.1)$$

$$b(t, S_t^{(1)}, S_t^{(2)}) = \begin{pmatrix} (r - q_1)S_t^{(1)} \\ (r - q_2)S_t^{(2)} \end{pmatrix}, \quad (2.3.2)$$

$$\sigma(t, S_t^{(1)}, S_t^{(2)}) = \begin{pmatrix} \Sigma_1 & \Sigma_2 \end{pmatrix} = \begin{pmatrix} \sigma_1 S_t^{(1)} & 0 \\ \rho \sigma_2 S_t^{(2)} & \sqrt{(1 - \rho^2)} \sigma_2 S_t^{(2)} \end{pmatrix}, \quad (2.3.3)$$

where  $B_t, Z_t$  are two uncorrelated Brownian motions,  $r \in \mathbb{R}$  is the risk-free rate,  $q_i \in \mathbb{R}$  is the dividend yield for asset  $i$ ,  $\sigma_i \in \mathbb{R}^+$  is the volatility of asset  $S_t^{(i)}$ , and  $\rho \in (-1, 1)$  is the correlation between  $S_t^{(1)}$  and  $S_t^{(2)}$ . The colour coding of the drift in (2.3.2) and the columns of the diffusion matrix in (2.3.3) will be explained in the caption of Figure 2.3.1.

The price of a contingent claim on assets whose dynamics are given by (2.3.1) is the solution to the following linear parabolic PDE

$$\begin{aligned} \frac{\partial V}{\partial t} + \frac{1}{2} \sigma_1^2 S_1^2 \frac{\partial^2 V}{\partial S_1^2} + \frac{1}{2} \sigma_2^2 S_2^2 \frac{\partial^2 V}{\partial S_2^2} + \rho \sigma_1 \sigma_2 S_1 S_2 \frac{\partial^2 V}{\partial S_1 \partial S_2} \\ + (r - q_1) S_1 \frac{\partial V}{\partial S_1} + (r - q_2) S_2 \frac{\partial V}{\partial S_2} - rV = 0, \end{aligned} \quad (2.3.4)$$

where  $(t, S_1, S_2) \in [0, T) \times (0, \infty)^2$ .

It is straightforward to check that as  $S_t^{(i)} \downarrow 0$  for any  $i \in \{1, 2\}$ , as stated in [26,

Section 6.2], the normal diffusion tends to zero and, therefore, for small enough grid refinement parameter  $\Delta S$ , the semi-Lagrangian stencil does not overstep. Indeed, consider the diffusion stencil for the grid node  $S_{1,j}$ , the stencil is  $y_1 \sim \sigma_1 S_{1,j} \sqrt{\Delta S}$ , then the stencil does not overstep provided  $\sigma_1 \sqrt{\Delta S} \leq 1$ .

However, this is not the case when  $S_t^{(1)}, S_t^{(2)} \rightarrow \infty$  due to the linear form of the drift and the diffusion coefficients in  $S_t^{(1)}$  and  $S_t^{(2)}$ . This is shown in Figure 2.3.1.

Figure 2.3.1 represents the semi-Lagrangian stencil for  $(S_1, S_2) \in [0, 50]^2$ . The points on the truncated boundaries ( $S_i = 50$  for any  $i \in \{1, 2\}$ ) are regular, as discussed in Section 2.1, as any trajectory of the process, starting at any point of the truncated boundary will exit the domain. Thus, the need to introduce artificial boundary conditions.

As discussed in Section 2.2, in order to incorporate Dirichlet boundary conditions in the scheme, we may need to truncate the stencil. This results in a worse CFL condition, requiring at least  $\Delta t \sim \Delta S^{3/2}$  for explicit schemes, and reducing by half an order the local consistency error. For the same model as in Figure 2.3.1, Figure 2.3.2 shows the sparsity plot of the discretization matrix and the nodes where the stencil oversteps.

In the next section we discuss a domain transformation so that all points on the boundary are not regular.

### 2.3.2 Boundary regularity for a transformed Black-Scholes equation

To avoid having to truncate the stencil consider the following smooth transformation

$$\xi_i = \frac{S_i}{S_i + P_i}, \quad (2.3.5)$$

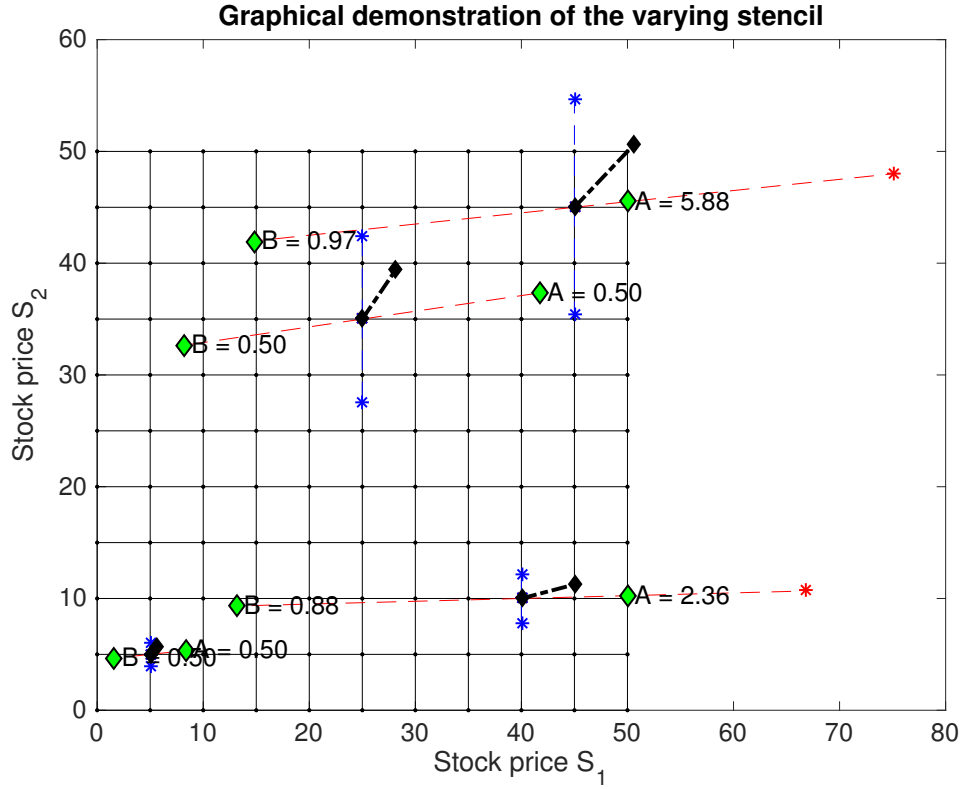
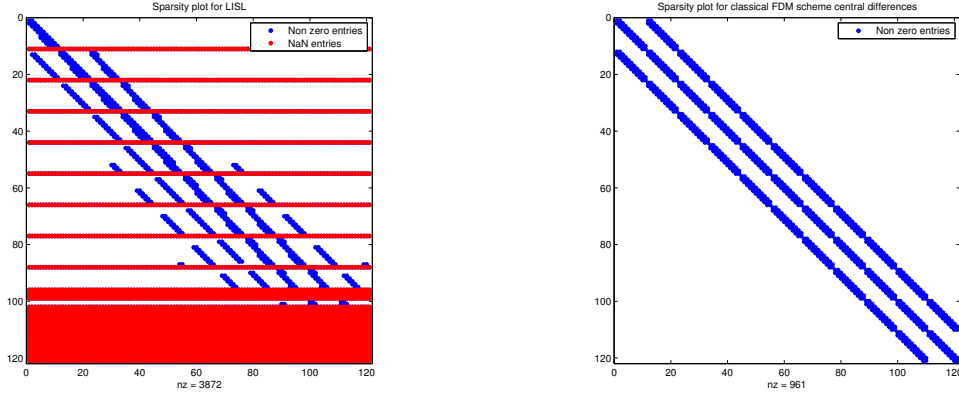


Figure 2.3.1: Graphical representation of the semi-Lagrangian stencil for the two dimensional Black-Scholes equations over a Cartesian grid of size  $11 \times 11$  for the domain  $[0, 50]^2$ . The parameters of the model are fixed as follows:  $r = 0.025$ ,  $q_1 = q_2 = 0$ ,  $\sigma_1 = 0.3$ ,  $\sigma_2 = 0.1$ ,  $\rho = 0.3$ . The lines linking the stencil nodes to the central node are colour coded according to the colours in equations (2.3.2)–(2.3.3), i.e. the drift is represented in black,  $\Sigma_1$  in blue and  $\Sigma_2$  in red. For  $\Sigma_2$  we also represent the nodes and the weights (divided by 2) of the truncated stencil according to (2.2.7).



(a) Sparsity plot for the discretization matrix of the two dimensional Black-Scholes equation using a LISL scheme. (b) Sparsity plot for the discretization matrix of the two dimensional Black-Scholes equation using a standard Finite Difference scheme.

Figure 2.3.2: Sparsity plot for the discretizations matrices using the a LISL scheme (left) and a standard fixed stencil scheme (right). The rows are represented vertically and the columns horizontally. The top left corner corresponds to the first row and first column. A blue point appears for each non-zero coefficient in the matrix. The red lines correspond to nodes whose stencil oversteps the mesh.

where  $i \in \{1, 2\}$  and  $P_i > 0$  is a location parameter. This transformation is applied to each spatial variable  $S$  in the Black-Scholes equation (2.3.4) resulting in

$$\begin{aligned} \frac{\partial V}{\partial t} + \frac{1}{2}\sigma_1^2\xi_1(1-\xi_1)\frac{\partial^2 V}{\partial \xi_1^2} + \frac{1}{2}\sigma_2^2\xi_2(1-\xi_2)\frac{\partial^2 V}{\partial \xi_2^2} + \rho\sigma_1\sigma_2\xi_1(1-\xi_1)\xi_2(1-\xi_2)\frac{\partial^2 V}{\partial \xi_1\partial \xi_2} \\ + [(r - q_1 - \sigma_1^2\xi_1)\xi_1(1-\xi_1)]\frac{\partial V}{\partial \xi_1} + [(r - q_2 - \sigma_2^2\xi_2)\xi_2(1-\xi_2)]\frac{\partial V}{\partial \xi_2} - rV = 0, \end{aligned} \quad (2.3.6)$$

for  $(t, \xi_1, \xi_2) \in [0, T) \times (0, 1)^2$ . Hence, under the transformed coefficients it is easy to see that all points in the smooth part of  $\partial([0, 1]^2)$  are not regular according to (2.1.5).

Indeed, let us consider the case  $\xi_1 \in \{0, 1\}$ , in this case the PDE coefficients are

$$b(t, \xi_1, \xi_2) = \begin{pmatrix} 0 \\ (r - q_2 - \sigma_2^2 \xi_2) \xi_2 (1 - \xi_2) \end{pmatrix},$$

$$\sigma(t, \xi_1, \xi_2) = \begin{pmatrix} 0 & 0 \\ \rho \sigma_2 \xi_2 (1 - \xi_2) & \sqrt{1 - \rho^2} \sigma_2 \xi_2 (1 - \xi_2) \end{pmatrix},$$

and  $n(\cdot)$ , the unit inner normal vector, is  $n(\xi_1, \xi_2) = (1, 0)^T$  for  $\xi_1 = 0$  and  $n(\xi_1, \xi_2) = (-1, 0)^T$  for  $\xi_1 = 1$ . Therefore,

$$\sigma^T(t, 0, \xi_2) n(0, \xi_2) = \sigma^T(t, 1, \xi_2) n(1, \xi_2) = 0,$$

for all values of  $t$  and  $\xi_2$ . To check the second condition in (2.1.5) we need the Jacobian and Hessian matrices of the distance function  $d(\xi_1, \xi_2)$ . From Section 14.6 in [38] we have that

$$Dd(\xi_1, \xi_2) = n(\xi_1, \xi_2), \quad D^2d(\xi_1, \xi_2) = 0,$$

where for the Hessian we have used the fact that the boundary is not curved in the direction of  $\xi_2$ . This allows us to conclude that

$$\sigma^T(t, 0, \xi_2) n(0, \xi_2) = \sigma^T(t, 1, \xi_2) n(1, \xi_2) = 0,$$

and therefore

$$\frac{1}{2} \text{tr}[\sigma(t, \xi_1, \xi_2) \sigma^T(t, \xi_1, \xi_2) D^2d(\xi_1, \xi_2)] + b(t, \xi_1, \xi_2) Dd(\xi_1, \xi_2) = 0,$$

for  $\xi_1 \in \{0, 1\}$  and for all  $t$  and  $\xi_2$ .

This argument is limited to the smooth part, i.e. away from the corners, as the conditions (2.1.5) require the distance function to be of class  $C^2$  in a neighbourhood of

the point, as discussed in Section 2.1, excluding the corners of the domain. However, as we approach the corners the equation degenerates to the following an ODE in  $t$

$$\frac{\partial V}{\partial t} - rV = 0.$$

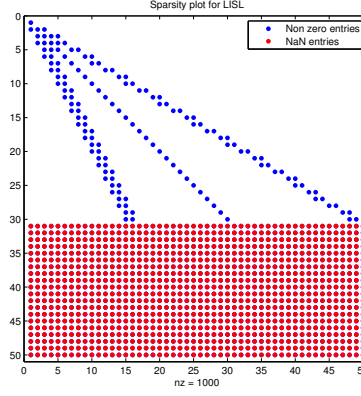
Thus, the transformation makes the whole boundary behave as in item 3. in Section 6.2 of [26] and the semi-Lagrangian stencil no longer oversteps the domain.

**Remark 2.3.1.** A curious effect of the transformation, is that it can yield tridiagonal matrices for one dimensional domains. However, this only holds for a given range of values of the mesh refinement parameter depending on the drift and diffusion coefficients. Even with the tridiagonal structure, this is not the best method for solving one dimensional problems. For example, we can use the classical central finite difference scheme to approximate the second order derivative as the method is monotone and second order accurate in this setting.

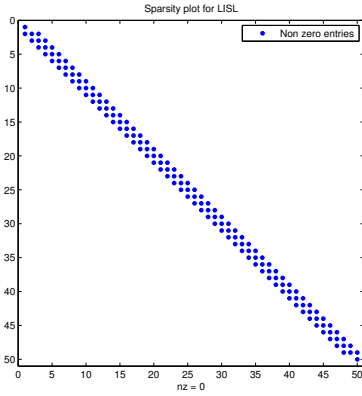
In Figure 2.3.3 we show sparsity plots for the discretization matrix in the original ( $S$ ) and the transformed domain ( $\xi$ ). We observe that if the scheme is applied in the original domain approximations need to be introduced for 20 out of 50 nodes, whereas, the method does not overstep the grid in the transformed domain. The discretization matrix is tridiagonal for some values of  $\sigma$ . For example, as shown in Figure 2.3.3 for  $\sigma = 4$  the matrix is not tridiagonal.

For Scheme 2 described in Section 2.2, assuming a uniform grid in  $\xi$  with refinement parameter  $h > 0$ , straightforward calculations show that the matrix has tridiagonal structure if  $\sigma \leq 2\sqrt{h}$ .

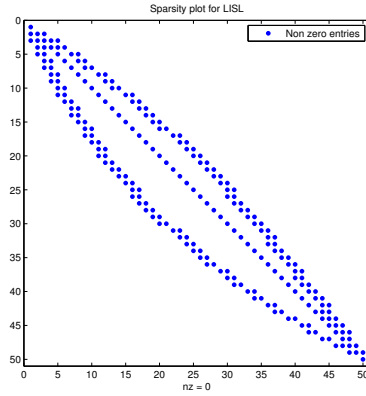
**Remark 2.3.2.** We note that the transformation in (2.3.5) is not the only one for which this technique holds. Another example is given by the map  $S \mapsto \frac{2}{\pi} \arctan(S)$ . For a one dimensional process the drift and volatility coefficients in the transformed



(a) Discretization matrix sparsity plot for L1SL-scheme Black and Scholes model.



(b) Discretization matrix sparsity plot for L1SL-scheme with variable transformation  $\xi = \frac{S}{S+E}$  with  $\sigma = 0.4$ .



(c) Discretization matrix sparsity plot for L1SL-scheme with variable transformation  $\xi = \frac{S}{S+E}$  with  $\sigma = 4$ .

Figure 2.3.3: Sparsity plots for the discretization matrices from the L1SL-scheme for the linear Black-Scholes equation in the original domain and in the transformed domain. The axes of the graph denote the rows and columns of the discretization matrices. The points in blue represent non-zero coefficients of the discretization matrices, and in red nodes whose stencil oversteps. In the case of the transformed domain, two matrices for different volatility factors have been plotted, this is to highlight that the tridiagonal matrix is only obtained for small  $\sigma$ .

variable  $\eta$  are given by

$$b(t, \eta) = \frac{2}{\pi}(r - q - \sigma^2) \sin\left(\frac{\pi\eta}{2}\right) \cos\left(\frac{\pi\eta}{2}\right), \quad \sigma(t, \eta) = \frac{2}{\pi}\sigma \sin\left(\frac{\pi\eta}{2}\right) \cos\left(\frac{\pi\eta}{2}\right).$$

An undesired consequence of the transformation is that terminal conditions with linear growth, such as call payoffs, become unbounded as  $\xi_i \rightarrow 1$ . This can be addressed by rescaling the value function by a smooth function such that the terminal conditions become bounded. A possible rescaling function is  $\sum_i S_i$ . Let  $\bar{V}$  be the rescaled value function

$$\bar{V}(t, S_1, S_2) = \frac{V(t, S_1, S_2)}{S_1 + S_2}, \quad \bar{V}(t, \xi_1, \xi_2) = \frac{V(t, \xi_1, \xi_2)}{\frac{P_1}{1-\xi_1} + \frac{P_2}{1-\xi_2}}. \quad (2.3.7)$$

For the PDE in the original coordinates  $(S_1, S_2)$  after the rescaling we have that

$$\begin{aligned} \frac{\partial \bar{V}}{\partial t} &+ \frac{1}{2}\sigma_1^2 S_1^2 \frac{\partial^2 \bar{V}}{\partial S_1^2} + \frac{1}{2}\sigma_2^2 S_2^2 \frac{\partial^2 \bar{V}}{\partial S_2^2} + \rho\sigma_1\sigma_2 S_1 S_2 \frac{\partial^2 \bar{V}}{\partial S_1 \partial S_2} \\ &+ \left( \frac{\sigma_1^2 S_1^2}{S_1 + S_2} + \rho\sigma_1\sigma_2 \frac{S_1 S_2}{S_1 + S_2} + (r - q_1)S_1 \right) \frac{\partial \bar{V}}{\partial S_1} \\ &+ \left( \frac{\sigma_2^2 S_2^2}{S_1 + S_2} + \rho\sigma_1\sigma_2 \frac{S_1 S_2}{S_1 + S_2} + (r - q_2)S_2 \right) \frac{\partial \bar{V}}{\partial S_2} - \left( \frac{q_1 S_1 + q_2 S_2}{S_1 + S_2} \right) \bar{V} = 0, \end{aligned}$$

for  $(t, S_1, S_2) \in [0, T) \times (0, \infty)^2$ . For the PDE in the transformed coordinates  $(\xi_1, \xi_2)$

the rescaling results in

$$\begin{aligned} \frac{\partial \bar{V}}{\partial t} &+ \frac{1}{2}\sigma_1^2 \xi_1^2 (1 - \xi_1)^2 \frac{\partial^2 \bar{V}}{\partial \xi_1^2} + \frac{1}{2}\sigma_2^2 \xi_2^2 (1 - \xi_2)^2 \frac{\partial^2 \bar{V}}{\partial \xi_2^2} + \rho\sigma_1\sigma_2 \xi_1 (1 - \xi_1) \xi_2 (1 - \xi_2) \frac{\partial^2 \bar{V}}{\partial \xi_1 \partial \xi_2} \\ &+ \frac{\partial \bar{V}}{\partial \xi_1} \left[ \frac{\sigma_1^2 \xi_1^2 P_1 (1 - \xi_2) + \rho\sigma_1\sigma_2 \xi_1 \xi_2 P_2 (1 - \xi_1)^2}{\xi_1 P_1 (1 - \xi_2) + \xi_2 P_2 (1 - \xi_1)} + (r - q_1 - \sigma_1^2 \xi_1) \xi_1 (1 - \xi_1) \right] \\ &+ \frac{\partial \bar{V}}{\partial \xi_2} \left[ \frac{\sigma_2^2 \xi_2^2 P_2 (1 - \xi_1) + \rho\sigma_1\sigma_2 \xi_1 \xi_2 P_1 (1 - \xi_2)^2}{\xi_1 P_1 (1 - \xi_2) + \xi_2 P_2 (1 - \xi_1)} + (r - q_2 - \sigma_2^2 \xi_2) \xi_2 (1 - \xi_2) \right] \\ &- \bar{V} \left[ \frac{q_1 \xi_1 P_1 (1 - \xi_2) + q_2 \xi_2 P_2 (1 - \xi_1)}{\xi_1 P_1 (1 - \xi_2) + \xi_2 P_2 (1 - \xi_1)} \right] = 0. \end{aligned} \quad (2.3.8)$$

**Remark 2.3.3.** Even if we have limited the discussion to the two-dimensional case,



these ideas can be extended to options on 3 or more assets. This is because, the coefficients on the  $i$ -th row of the volatility matrix are linear with respect to  $S_i$  and do not depend on the value of the other assets  $S_j$  for  $j \neq i$ .

### 2.3.3 Numerical experiment

For benchmarking purposes we solve the problem of pricing a call option on the maximum of two assets. The terminal condition is given by

$$V(T, S_1, S_2) = \max \left\{ \max_{i \in \{1,2\}} S_i - K \right\}, \quad \text{for } (S_1, S_2) \in [0, \infty)^2, \quad (2.3.9)$$

where  $K \geq 0$  is called the strike price. Under the Black-Scholes model this problem admits a closed-form solution as shown in [74].

We approximate the solution using three different approaches. First, we consider two different schemes for the two-dimensional Black-Scholes equation, as in (2.3.4), with terminal value (2.3.9) on a truncated spatial domain  $(S_1, S_2) \in [0, S_{max}]^2$ . At the regular part of the boundary, that is the part where for any  $i \in \{1, 2\}$   $S_i = S_{max}$ , we set the asymptotic value

$$V(t, S_1, S_2) = S_{max} - e^{-r(T-t)}K, \quad \text{for } S_1 = S_{max} \text{ or } S_2 = S_{max}.$$

We approximate this problem using two finite difference schemes: the standard central difference fixed stencil finite difference method, see [76], and the truncated LISL scheme discussed in Section 2.2.

Second, we consider the LISL method applied to (2.3.8) for the domain  $(0, 1)^2$  with no boundary conditions and rescaling terminal condition (2.3.9) as in (2.3.7).

We consider explicit and implicit time stepping schemes for both of the semi-Lagrangian methods. As expected, both of the implicit schemes and the explicit one for the transformed PDE are stable for  $\Delta t = \mathcal{O}(\Delta x)$ , while the explicit truncated

LISL scheme requires  $\Delta t = \mathcal{O}(\Delta x^{3/2})$  for stability.

The values for the parameters of the model are chosen as  $r = 0.05$ ,  $q_1 = q_2 = 0$ ,  $T = 0.25$ ,  $K = 40$ ,  $S_{max} = 400$ ,  $\sigma_1 = \sigma_2 = 0.5$ , and  $\rho = 0.3$ . We use an equal number of grid points per spatial dimension given by  $N_S = 10 \cdot 2^n + 1$  with  $n \in \{1, 2, \dots, 7\}$ . For the number of time steps  $N_t$ , we choose  $N_t = \mathcal{O}(N_S^{3/2})$  for the truncated LISL scheme and explicit time stepping, and  $N_t = \mathcal{O}(N_S)$  for the rest. For the transformation we set  $P_1 = P_2 = 40$ , ensuring that the strike price is a mesh point.

It is common practice to choose  $S_{max}$  far away from the region of interest to limit the impact the approximate boundary condition has on the error, see [6, 5]. Therefore, for  $\Delta S > 0$ , let  $S_{\Delta S} \subset [0, S_{max}]^2$  be the grid in the original domain. For each method we compute the errors with respect to the analytic solution in the  $L^\infty$ -norm for the grid  $S_{\Delta S} \cap [0, 200]^2$ . For the scheme in the transformed domain, we undo the transformations to the value function and for the error we use linear interpolation to obtain approximate prices for the grid  $S_{\Delta S} \cap [0, 200]^2$ .

We observe that all three schemes show, approximately, convergence proportional to  $\mathcal{O}(N_S^{-1})$ . The classical central space finite difference discretization is the most accurate. Convergence proportional to  $N_S^{-1}$  shows that the time discretization error dominates as the spatial discretization is second order accurate. Comparing both of the approaches using semi-Lagrangian schemes, we observe that the scheme approximating the transformed equation is the most accurate. This may be due to an efficient use of the grid nodes. Indeed, by the shape of the transformation, a uniform grid in  $(\xi_1, \xi_2)$  results in a non-uniform grid in  $(S_1, S_2)$ . The resulting grid in  $(S_1, S_2)$  is finest for points in the region  $[0, P_1] \times [0, P_2]$ , whereas it is coarse for points in the region where  $S_1 \gg P_1$  and  $S_2 \gg P_2$ . Recall that for the numerical tests we chose  $P_1 = P_2 = K$ .

Despite being the most accurate method, the classical central space finite difference discretization may not be the best method to employ as, being non-monotone,

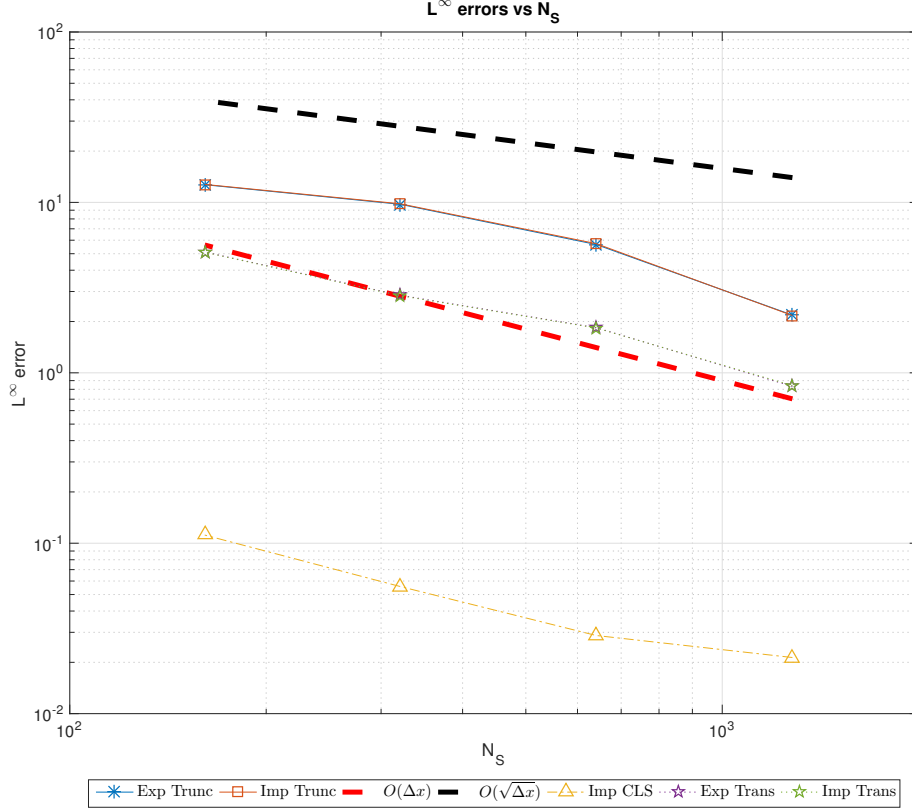


Figure 2.3.4: Errors in the  $L^\infty$  norm for the interval  $(S_1, S_2) \in [0, 200]^2$  for the price of a call option on the maximum of two assets. We denote by *Exp* and *Imp* whether the time-stepping is explicit or implicit. *Trunc* refers to the LISL method with truncated stencil introduced in Section 2.2. *CLS* refers to the classical central space finite difference discretization as in [76]. *Trans* refers to the LISL scheme in [26] applied to the transformed equation (2.3.8). The lines labelled  $\mathcal{O}(\Delta x)$  and  $\mathcal{O}(\sqrt{\Delta x})$  are straight lines with slopes equal to -1 and  $-\frac{1}{2}$  and shown for readability of the plots.

some of the approximate prices may be negative.

### 2.3.4 Extension to the non-linear case

In this section we show that the smooth transformation of the domain and the rescaling of the value function can also be applied to the non-linear case, where the solution is given in the viscosity sense. The results are somewhat obvious from the fact that both transformations involve smooth functions.

**Proposition 2.3.1.** *Let  $u$  be a viscosity solution of*

$$H(t, x, u, u_t, Du, D^2u) = 0, \quad \text{for } (t, x) \in (0, T) \times \Omega,$$

where  $\Omega \subseteq \mathbb{R}^d$  and  $\phi : \Omega \rightarrow \Omega_\phi$  is a smooth and invertible function with inverse  $\phi^{-1}$ .

Then,  $v(t, y) := u(t, \phi(x))$  is a viscosity solution of

$$H(t, \phi^{-1}(y), v, v_t, D(v \circ \phi), D^2(v \circ \phi)) = 0 \quad \text{for } (t, y) \in (0, T) \times \Omega_\phi.$$

*Proof.* By definition  $u$  is a viscosity subsolution of  $H$ . Then pick  $y_0 \in \Omega_\phi$  and a test function  $\varphi$  such that

$$\varphi(y_0) = v(y_0), \quad \text{and} \quad \varphi(y) - v(y), \text{ has local minimum at } y = y_0.$$

Then, by properties of  $\phi$ , there exists a point  $x_0$  such that  $y_0 = \phi(x_0)$  then

$$\varphi(\phi(x_0)) = v(\phi(x_0)) = u(x_0), \quad \text{and} \quad \varphi(\phi(x)) - v(\phi(x)) \text{ has local min at } x_0.$$

Therefore, as  $u$  is a subsolution of the HJB equation  $H$  we have that

$$H(t, \phi^{-1}(y_0), \varphi \circ \phi(x_0), D(\varphi \circ \phi)(x_0), D^2(\varphi \circ \phi)(x_0)) \leq 0.$$

The supersolution case follows similarly so we skip the details.  $\square$

Using similar arguments we can prove the following for the rescaling of the value function

**Proposition 2.3.2.** *Let  $u$  be a viscosity solution of*

$$H(t, x, u, u_t, Du, D^2u) = 0, \quad \text{for } (t, x) \in (0, T) \times \Omega,$$

where  $\Omega \subseteq \mathbb{R}^d$  and  $\phi \in C^2(\Omega)$  be a strictly positive function. Then,  $\bar{u}(t, x) := \frac{u(t, x)}{\phi(x)}$  is a viscosity solution of

$$H(t, x, u/\phi, (u/\phi)_t, D(u/\phi), D^2(u/\phi)) = 0 \quad \text{for } (t, x) \in (0, T) \times \Omega.$$

For instance, using these two propositions we could replicate the analysis in Section 2.3.2 to price options under the uncertain volatility model proposed in [4, 55].

## 2.4 Conclusion

In this chapter we have studied the known issue of stencil overstepping. In Section 2.2 we have shown that it is possible to truncate the semi-Lagrangian stencil in a monotone and consistent way in the presence of Dirichlet boundary conditions. The proposed truncation reduces the local consistency error by half an order and worsens the CFL condition by at least half an order. In Section 2.3 we have shown that, under certain conditions in the coefficients and the domain, it is possible to avoid the truncation of the stencil by means of adequate domain transformations.

The truncation of the stencil alters the scheme significantly for regions near the boundary. This modification renders part of the error analysis in [26] not applicable for the truncated scheme. The purpose of the next chapter is, therefore, to seek an

alternative method to quantify the error convergence rates for the LISL scheme with truncation.

Requiring a rather restrictive CFL condition makes time-explicit schemes computationally inefficient with respect to the consistency error. Using this as motivation we consider the use of implicit time stepping schemes. Thus, in Chapter 4, we analyse how to solve algebraic system of equations for LISL discretization matrices efficiently.

# Chapter 3

## Error bounds for monotone schemes for the Cauchy-Dirichlet problem for HJB equations

### 3.1 Introduction

We consider error estimates for monotone finite difference schemes for the numerical approximation of solutions to the Hamilton-Jacobi-Bellman (HJB) equation

$$u_t + \sup_{\alpha \in \mathcal{A}} \mathcal{L}^\alpha(t, x, u, Du, D^2u) = 0, \quad \text{in } Q_T, \quad (3.1.1)$$

$$u(0, x) = \Psi_0(x), \quad \text{for } x \in \bar{\Omega}, \quad (3.1.2)$$

$$u(t, x) = \Psi_1(t, x), \quad \text{for } (t, x) \in (0, T] \times \partial\Omega, \quad (3.1.3)$$

where  $\Omega$  is an open and bounded subset of  $\mathbb{R}^d$ ,  $Q_T := (0, T] \times \Omega$ ,  $\bar{\Omega} := \Omega \cup \partial\Omega \subset \mathbb{R}^d$ ,  $\mathcal{A}$  is a compact metric space,  $\mathcal{L}^\alpha : (0, T] \times \Omega \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^{d \times d} \rightarrow \mathbb{R}$  defined as

$$\mathcal{L}^\alpha(t, x, r, q, X) = -\text{tr}[a^\alpha(t, x)X] - b^\alpha(t, x)q - c^\alpha(t, x)r - f^\alpha(t, x), \quad (3.1.4)$$

is a second order differential operator,  $\Psi_0$  and  $\Psi_1$  are the initial and boundary conditions respectively. The coefficients  $a^\alpha$ ,  $b^\alpha$ ,  $c^\alpha$  and  $f^\alpha$  take values in  $\mathcal{S}^d$ , the space of  $d \times d$  real symmetric matrices,  $\mathbb{R}^d$ ,  $\mathbb{R}$  and  $\mathbb{R}$ ,  $\partial\Omega$  is the boundary of  $\Omega$  and we denote by  $\partial^*Q_T$  the parabolic boundary of  $Q_T$ , i.e.  $\partial^*Q_T := (\{0\} \times \overline{\Omega}) \cup ((0, T] \times \partial\Omega)$ . For compactness we define

$$F(t, x, u, Du, D^2u) := \sup_{\alpha \in \mathcal{A}} \mathcal{L}^\alpha(t, x, u, Du, D^2u). \quad (3.1.5)$$

where the operator  $\mathcal{L}^\alpha$  is defined in (3.1.4).

The aim of the error analysis is to estimate the difference between the viscosity solution of the HJB equation and an approximate solution computed by means of a numerical scheme. Let  $\mathcal{G}_h \subset \overline{Q_T}$  be a discrete grid with refinement parameter  $h$ , then numerical schemes for (3.1.1)–(3.1.3) can be written as

$$\begin{aligned} S(h, t, x, u_h(t, x), [u_h]_{t,x}) &= 0 & \text{in } \mathcal{G}_h^+ &:= \mathcal{G}_h \setminus (\{t = 0\} \cup \partial\Omega), \\ u_h(0, x) &= \Psi_{h,0}(x) & \text{in } \mathcal{G}_h^0 &:= \mathcal{G}_h \cap \{t = 0\}, \\ u_h(t, x) &= \Psi_{h,1}(t, x) & \text{in } \mathcal{G}_h^1 &:= \mathcal{G}_h \cap ((0, T] \cap \partial\Omega), \end{aligned} \quad (3.1.6)$$

where  $S$  is a consistent, monotone and uniformly continuous approximation of the equation (3.1.1) on the grid  $\mathcal{G}_h^+$ . By analogy to the continuous case we denote by  $\partial^*\mathcal{G}_h := \mathcal{G}_h^0 \cup \mathcal{G}_h^1$ . Following the notation in [8], we say that the numerical solution  $u_h : \mathcal{G}_h \rightarrow \mathbb{R}$  is a grid function and, if finite, belongs to  $C_b(\mathcal{G}_h)$ , the space of bounded and continuous grid functions. We are interested in discrete  $\mathcal{G}_h$ , therefore as noted in [8], any grid function on  $\mathcal{G}_h$  is continuous.

The objective is thus to find upper and lower bounds for the difference  $u - u_h$ . The cornerstone of the analysis is the use of Krylov’s “shaking coefficients” method (see [50, 51]) to find perturbed equations from which to construct smooth approximations to  $u$  and, under certain regularity of the numerical scheme, to  $u_h$ . This regularization



relies on the mollification of certain functions. For doing this we will take convolutions with the following family of mollifiers in time and space

$$\rho_\varepsilon(t, x) := \frac{1}{\varepsilon^{d+2}} \rho\left(\frac{t}{\varepsilon^2}, \frac{x}{\varepsilon}\right) \quad (3.1.7)$$

where  $\varepsilon > 0$ , and

$$\rho \in C^\infty(\mathbb{R}^{d+1}), \quad \rho \geq 0, \quad \text{supp } \rho = (0, 1) \times \{|x| < 1\}, \quad \int_{\text{supp } \rho} \rho(e) de = 1.$$

Equivalently, we define a family of mollifiers in space by

$$\rho_\varepsilon(x) := \frac{1}{\varepsilon^d} \rho\left(\frac{x}{\varepsilon}\right) \quad (3.1.8)$$

where  $\varepsilon > 0$ , and

$$\rho \in C^\infty(\mathbb{R}^d), \quad \rho \geq 0, \quad \text{supp } \rho = \{|x| < 1\}, \quad \int_{\text{supp } \rho} \rho(e) de = 1.$$

Key to the approach is the convexity (or concavity) of the equation under analysis. Convexity of (3.1.1) is used to prove that mollified subsolutions of (3.1.1) are also subsolutions of (3.1.1), see Lemma 2.7 in [7]. Without the convexity (or concavity) of the equation the error analysis yields weaker results. For instance, [18] proves the existence of an algebraic rate of convergence for the finite-difference approximation of

$$F(D^2u) = f(x),$$

on a regular domain with Dirichlet boundary data. This result is further extended to the Isaacs equation in [52]. However, neither of these articles provide an explicit way to calculate such rates. Furthermore, this rate may depend on the constant of ellipticity, see [52].

Convexity (concavity) is also the reason why we can build smooth subsolutions (supersolutions) by “shaking the coefficients” of the equation, but, for the other bound, we need an alternative approach as we cannot construct smooth supersolutions (sub-solutions). Two different approaches have been developed for the error analysis of monotone finite difference schemes. The first approach, constructs smooth subsolutions to both the HJB equation (3.1.1), and the scheme (3.1.6). For instance, this is the approach followed in [26] to derive error bounds for semi-Lagrangian schemes when  $\Omega = \mathbb{R}^d$ . The second approach, also derives one of the bounds by “shaking the coefficients” of the equation, but for the other bound, it relies on approximations to the equation (3.1.1).

More specifically, the first approach treats the equation and the scheme analogously by constructing a smooth subsolutions to each of them. For this approach to work, we need continuous dependence estimates on the boundary data and the coefficients for both (3.1.1) and the scheme (3.1.6). Such results have been proved for Lipschitz solutions of equation (3.1.1) under general conditions on the coefficients and the data for  $\Omega = \mathbb{R}^d$ , see [34, 84] for a derivation based on probabilistic arguments or [47] for a derivation based on analytic methods. For semi-Lagrangian schemes continuous dependence results were proved in [26]. The proof replicates the one for the equation. A key step is to extend the definition of the scheme to the continuous spatial domain, see Section 7 in [26]. This extension avoids the effects of interpolation when proving the continuous dependence result, splitting the lower bound into two, see Corollary 7.3 in [26]. Similar ideas were used in [19, Proposition 2.1] to prove continuous regularity of the numerical solution. Having constant finite difference weights was also used at key steps in the proofs.

For the truncated LISL scheme presented in Section 2.2, it is not clear how to obtain the continuous dependence result. This is due to variable finite difference weights depending on both the equation coefficients and the shape of the boundary.

Additionally, the proof of continuous dependence relies on the numerical solution being Lipschitz when the scheme is extended to the whole domain and it is unclear how to derive this for the schemes in Section 2.2.

As a result, we follow the second approach, as in [8], and obtain the lower bound by using a switching system approximation to the HJB equation. The authors of [8] have compared both of the approaches and found that this alternative approach is more general but results in lower rates. For the semi-Lagrangian schemes as described in [26], a comparison of the error rates obtained by both of these approaches can be found in [27]. Both of the approaches give fractional convergence rates of  $\mathcal{O}(\Delta x^{1/5})$  and  $\mathcal{O}(\Delta x^{1/10})$  for the first and second approach respectively.

The analysis in this chapter closely follows the logic in [8], where the main difference is that we consider a regular and bounded domain  $\Omega$  with Dirichlet conditions. As shown in [15] for a semi-infinite domain with an oblique derivative condition, the application of Krylov's regularization to problems posed on spatial domains with boundaries requires some domain perturbation or extension of the spatial domain.

**Remark 3.1.1.** Assuming that the solution to (3.3.1)–(3.3.3) satisfies the boundary data pointwise is a restrictive requirement, but this assumption is a key assumption for the analysis. As shown in the example below, constructing a scheme that enforces the boundary conditions for a solution that ignores them impacts the convergence of the numerical scheme.

**Example 3.1.1.** Consider the following second order parabolic PDE

$$u_t - \frac{1}{2}x^2(1-x)^2u_{xx} + u = 0, \quad \text{for } (t, x) \in (0, T] \times (0, 1), \quad (3.1.9)$$

$$u(t, x) = 1, \quad \text{for } (t, x) \in (\{0\} \times [0, 1]) \cup ((0, T] \times \{0, 1\}). \quad (3.1.10)$$

By the works of [9, 31, 66], we know that the boundary points  $x \in \{0, 1\}$  are not regular and that the equation holds up to the boundary. We observe that  $u(t, x) =$

$e^{-t}v(t, x)$ , where  $v(t, x)$  is the solution of

$$\begin{aligned} v_t - \frac{1}{2}x^2(1-x)^2v_{xx} &= 0 & \text{for } (t, x) \in (0, T] \times (0, 1), \\ v(t, x) &= 1, & \text{for } (t, x) \in (\{0\} \times [0, 1]). \end{aligned}$$

Hence, for all  $x \in [0, 1]$  we have that  $\lim_{T \rightarrow \infty} u(T, x) = 0$ .

Let  $\Delta t, \Delta x \geq 0$ ,  $N := T/\Delta t$  and  $J := 1/\Delta x$ , then a possible numerical scheme for the approximation of (3.1.9)–(3.1.10) is the following explicit scheme

$$S(h, t_n, x_j, U_j^n, [U]_{n,j}) = \frac{U_j^n - U_j^{n-1}}{\Delta t} - \frac{1}{2}j^2(1-j\Delta x)^2(U_{j+1}^{n-1} - 2U_j^{n-1} + U_{j-1}^{n-1}) + U_j^{n-1},$$

where  $h = (\Delta t, \Delta x)$ ,  $n \in [1, N]$ ,  $j \in [1, J-1]$ ,  $t_n = n\Delta t$ ,  $x_j = j\Delta x$ , and  $U_j^n \equiv U(t_n, x_j)$ . The scheme enforces the initial and boundary conditions. It is straightforward to prove that the scheme is monotone and  $L^\infty$ -stable provided that  $\Delta t \sim \mathcal{O}(\Delta x^2)$ .

Focusing on the node with  $j = 1$ , we observe that, considering viscosity solutions satisfying the boundary conditions in the weak sense for the equation and in the strong sense for the scheme, does not lead to uniform convergence. In particular, from the limits below, the numerical solution at the node with  $j = 1$  converges to a constant different from 0:

$$\begin{aligned} U_1^{n+1} &= \frac{\Delta t}{2}(1-\Delta x)^2[U_2^n + 1] + (1-\Delta t - \Delta t(1-\Delta x)^2)U_1^n \\ &\geq \frac{\Delta t}{2} + (1-2\Delta t)U_1^n \end{aligned} \tag{3.1.11}$$

$$\geq \sum_{m=0}^n (1-2\Delta t)^m \frac{\Delta t}{2} + (1-2\Delta t)^{n+1} \xrightarrow{n \rightarrow \infty} \frac{1+3e^{-2T}}{4} > \frac{1}{4}, \tag{3.1.12}$$

where the inequality in (3.1.11) is obtained from the fact that  $U_1^n, U_2^n \geq 0$ .

Therefore, from (3.1.12) we deduce that for  $T \gg \ln(4)$  the scheme cannot converge

uniformly.

The rest of the chapter is organised as follows. Section 3.2 contains definitions and results for viscosity solution used throughout the rest of the chapter. Section 3.3 contains some theoretical results on switching systems with Dirichlet boundary conditions used in the error analysis. Section 3.4 estimates the convergence rate of some switching system to a related HJB equation. Section 3.5 contains the derivation of the error bounds. Section 3.6 calculates the error bounds for some example schemes. Section 3.7 concludes.

## 3.2 Definitions and results on viscosity solutions

This section contains definitions and well-known results for HJB equations used throughout the rest of the chapter.

We start with a basic property for the differential operators considered.

**Definition 3.2.1** (Proper operator). An operator  $F$  is said to be proper if

$$F(t, x, r, q, X) - F(t, x, s, q, X) \geq \gamma_R(r - s),$$

if  $(t, x) \in \overline{Q}_T$ ,  $R \geq r \geq s \geq -R$ ,  $q \in \mathbb{R}^d$ ,  $X \in \mathcal{S}^d$ , for some  $\gamma_R > -\infty$ , for all  $R < \infty$ , and

$$F(t, x, r, q, X) \leq F(t, x, r, q, Y), \quad \text{whenever } Y \leq X,$$

for all  $(t, x) \in \overline{Q}_T$ ,  $r \in \mathbb{R}$ ,  $q \in \mathbb{R}^d$ ,  $X, Y \in \mathcal{S}^d$ .

The differential operator defined in (3.1.5) is proper, see [46, Section IV.2].

In the following, for a domain  $Q_T$ , we denote by  $\overline{Q}_T$  its closure and by  $\partial^*Q_T$  the parabolic boundary, i.e.  $\partial^*Q_T := (\{0\} \times \overline{\Omega}) \cup ((0, T] \times \partial\Omega)$ . We denote by  $\leq$  the component by component ordering in  $\mathbb{R}^d$  and the ordering in the sense of positive semi-definite matrices in  $\mathcal{S}^d$ .

Let  $\phi : Q \rightarrow \mathbb{R}^d$  be a bounded function from some set  $Q$  into  $\mathbb{R}^d$  with  $d \geq 1$ , then the following function norms are used

$$|\phi|_0 := \sup_{(t,y) \in Q} |\phi(t,y)|,$$

and for any  $\delta \in (0, 1]$ ,

$$[\phi]_\delta := \sup_{(t,x) \neq (s,y)} \frac{|\phi(t,x) - \phi(s,y)|}{(|x-y| + |t-s|^{1/2})^\delta}, \quad \text{and} \quad |\phi|_\delta = |\phi|_0 + [\phi]_\delta.$$

As usual, we denote by  $C^n(Q)$  the space of continuous and  $n$ -times differentiable functions, where  $n = 0$  is used for the space of bounded continuous functions in  $Q$ . Additionally,  $\mathcal{C}^{0,\delta}$  denotes the subset of  $C^0$  with finite  $|\cdot|_\delta$  norm.

Let  $Q$  be an open set and  $d \in \mathbb{N}$ , for a locally bounded function  $\phi : Q \rightarrow \mathbb{R}^d$  we define its upper-semicontinuous envelope

$$\phi^*(x) = \limsup_{\substack{y \rightarrow x \\ y \in Q}} \phi(y),$$

and its lower-semicontinuous envelope

$$\phi_*(x) = \liminf_{\substack{y \rightarrow x \\ y \in Q}} \phi(y).$$

Then,  $\text{USC}(Q; \mathbb{R}^d)$ ,  $\text{LSC}(Q; \mathbb{R}^d)$  are the usual spaces of upper and lower semicontinuous functions, respectively.

The relevant notion of solutions for this type of non-linear equations is that of viscosity solutions, see [23] for a detailed overview. It is known that to account for degenerate operators the boundary conditions need to be considered in the “weak sense”, that is the solution at the boundary can satisfy the equation or the boundary condition, see Section 7 in [23]. However, for the analysis we will assume that the

solutions considered satisfy the boundary conditions pointwise. This is ensured by assumptions on the spatial boundary  $\partial\Omega$  and the equation coefficients according to the analysis in [9, 20], as discussed in Section 2.1.

In the next definition we recall the notion of solution for (3.1.1)–(3.1.3) when the boundary and initial conditions are satisfied in the “strong sense”.

**Definition 3.2.2.** A function  $\bar{u} \in \text{USC}([0, T] \times \bar{\Omega}; \mathbb{R})$  is a viscosity subsolution, if for each function  $\varphi \in C^{1,2}([0, T] \times \bar{\Omega})$ , at each maximum point  $(t, x)$  of  $\bar{u} - \varphi$  we have that

$$\begin{aligned} \varphi_t + F(t, x, \bar{u}, D\varphi, D^2\varphi) &\leq 0, & (t, x) &\in (0, T] \times \Omega, \\ \varphi - \Psi_0 &\leq 0, & (t, x) &\in \{0\} \times \bar{\Omega}, \\ \varphi - \Psi_1 &\leq 0, & (t, x) &\in (0, T] \times \partial\Omega. \end{aligned}$$

Similarly, a function  $\underline{u} \in \text{LSC}([0, T] \times \bar{\Omega}; \mathbb{R})$  is a viscosity supersolution, if for each function  $\varphi \in C^{1,2}([0, T] \times \bar{\Omega})$ , at each minimum point  $(t, x)$  of  $\underline{u} - \varphi$  we have that

$$\begin{aligned} \varphi_t + F(t, x, \underline{u}, D\varphi, D^2\varphi) &\geq 0, & (t, x) &\in (0, T] \times \Omega, \\ \varphi - \Psi_0 &\geq 0, & (t, x) &\in \{0\} \times \bar{\Omega}, \\ \varphi - \Psi_1 &\geq 0, & (t, x) &\in (0, T] \times \partial\Omega. \end{aligned}$$

Finally, a continuous function  $u$  is a viscosity solution of (3.1.1)–(3.1.3) if it is both a subsolution and a supersolution.

This definition of viscosity solutions is done in terms of smooth test functions  $\varphi$ . As noted in Section 2 in [23], it will be useful to find an equivalent definition of viscosity solutions that would allow to represent “ $(Du, D^2u)$ ” for non-differentiable functions  $u$ <sup>1</sup>. This is achieved by means of semijets, see Section 8 in [23].

---

<sup>1</sup>This requirement is linked to the maximum principle for semi-continuous functions, also known

**Definition 3.2.3** (Definition 2.1 in [47]). For a function  $u$  belonging to  $\text{USC}([0, T] \times \bar{\Omega}; \mathbb{R})$  ( $\text{LSC}([0, T] \times \bar{\Omega}; \mathbb{R})$ ) that is locally bounded, the second-order parabolic superjet (subjet) of  $u$  at  $(t, x) \in Q_T$ , denoted by  $\mathcal{P}^{2,+(-)}u(t, x)$ , is defined as the set of triples  $(a, p, X) \in \mathbb{R} \times \mathbb{R}^d \times \mathcal{S}^d$  such that

$$\begin{aligned} u(s, y) \leq (\geq) & u(t, x) + a(s - t) + \langle p, y - x \rangle + \frac{1}{2} \langle X(y - x), y - x \rangle \\ & + o(|s - t| + |x - y|^2), \end{aligned}$$

as  $Q_T \ni (s, y) \rightarrow (t, x)$ . We define the closure  $\overline{\mathcal{P}}^{2,+(-)}u(t, x)$  as the set of  $(a, p, X) \in \mathbb{R} \times \mathbb{R}^d \times \mathcal{S}^d$  for which there exists  $(t_n, x_n, u(t_n, x_n), a_n, p_n, X_n) \in Q_T \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d \times \mathcal{S}^d$  such that  $(t_n, x_n, u(t_n, x_n), a_n, p_n, X_n) \rightarrow (t, x, u(t, x), a, p, X)$  as  $n \rightarrow \infty$  and  $(a_n, p_n, X_n) \in \mathcal{P}^{2,+(-)}u(t_n, x_n)$  for all  $n$ .

It is straightforward to rephrase Definition 3.2.2 in terms of these semijets, using the superjets (subjets) to define subsolutions (supersolutions).

The following theorem allows the comparison between sub- and supersolutions.

**Theorem 3.2.1** (Theorem 8.2 in [23]). *Let  $\Omega \subset \mathbb{R}^d$  be open and bounded. Let  $F \in C([0, T] \times \bar{\Omega} \times \mathbb{R} \times \mathbb{R}^d \times \mathcal{S}^d)$  be continuous, proper, and such that there is a function  $\omega : [0, \infty] \rightarrow [0, \infty]$  for which  $\omega(0+) = 0$  such that,*

$$F(t, y, r, \beta(x - y), Y) - F(t, x, r, \beta(x - y), X) \leq \omega(\beta|x - y|^2 + |x - y|) \quad (3.2.1)$$

for every  $(t, x), (t, y) \in (0, T) \times \Omega$ ,  $\beta > 0$ ,  $r \in \mathbb{R}$  and symmetric matrices  $X, Y$  satisfying

$$-3\beta \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \leq \begin{pmatrix} X & 0 \\ 0 & -Y \end{pmatrix} \leq 3\beta \begin{pmatrix} I & -I \\ -I & I \end{pmatrix}. \quad (3.2.2)$$

---

as the Crandall-Ishii lemma, see Theorem 8.3 in [23] or Theorem 3.2.3.



If  $\bar{u}$  is a subsolution of (3.1.1)–(3.1.3) and  $\underline{v}$  is a supersolution of (3.1.1)–(3.1.3), then  $\bar{u} \leq \underline{v}$  on  $[0, T] \times \bar{\Omega}$ .

Under the structural assumption (3.2.1) uniqueness of the solution  $u$  is proved in Section V.8 of [34] as a corollary of the following theorem.

**Theorem 3.2.2** (Theorem V.8.1 in [34]). *Suppose that  $F$  is continuous and satisfies (3.2.1). Let  $\bar{u} \in \text{USC}(\bar{Q}_T; \mathbb{R})$  be a viscosity subsolution of (3.1.1) in  $Q_T$ , and  $\underline{u} \in \text{LSC}(\bar{Q}_T; \mathbb{R})$  be a viscosity supersolution of (3.1.1) in  $Q_T$ . Then*

$$\sup_{\bar{Q}_T} (\bar{u} - \underline{u}) = \sup_{\partial^* Q_T} (\bar{u} - \underline{u}). \quad (3.2.3)$$

The main theoretical tool for the proof of Theorem 3.2.1 is the Crandall-Ishii lemma [22] which we report here for completeness.

**Theorem 3.2.3** (Theorem 8.3 in [23]). *Let  $u_1$  and  $-u_2$  belong to  $\text{USC}(\bar{Q}_T; \mathbb{R})$ . Let  $\phi(t, x, y) \in C^{1,2,2}((0, T] \times \bar{\Omega} \times \bar{\Omega})$ , i.e. once continuously differentiable in  $t \in (0, T]$  and twice continuously differentiable in  $(x, y) \in \bar{\Omega} \times \bar{\Omega}$ . Suppose  $(t_\phi, x_\phi, y_\phi) \in (0, T] \times \Omega \times \Omega$  is a local maximum of the function*

$$(t, x, y) \rightarrow u_1(t, x) - u_2(t, y) - \phi(t, x, y).$$

Suppose that there is an  $r > 0$  such that for every  $M > 0$  there is a constant  $C$  such that

$$\left\{ \begin{array}{l} a \leq C \text{ whenever } (a, p, X) \in \mathcal{P}^{2,+} u_1(t, x), \\ |x - x_\phi| + |t - t_\phi| \leq r, |u_1(t, x)| + |p| + \|X\| \leq M, \\ b \geq C \text{ whenever } (b, q, Y) \in \mathcal{P}^{2,-} u_2(t, y), \\ |y - y_\phi| + |t - t_\phi| \leq r, |u_2(t, y)| + |q| + \|Y\| \leq M. \end{array} \right.$$

Then for any  $\kappa > 0$  there exist two numbers  $a, b \in \mathbb{R}$  and two matrices  $X, Y \in \mathcal{S}^d$

such that

$$\begin{aligned}(a, D_x \phi(t_\phi, x_\phi, y_\phi), X) &\in \overline{\mathcal{P}}^{2,+} u_1(t_\phi, x_\phi), \\ (b, -D_y \phi(t_\phi, x_\phi, y_\phi), Y) &\in \overline{\mathcal{P}}^{2,-} u_2(t_\phi, y_\phi),\end{aligned}$$

$$-\left(\frac{1}{\kappa} + \|A\|\right) \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \leq \begin{pmatrix} X & 0 \\ 0 & -Y \end{pmatrix} \leq A + \kappa A^2, \quad (3.2.4)$$

where  $A = D^2 \phi(t_\phi, x_\phi, y_\phi)$ ,  $\|A\| = \sup_{|\xi| \leq 1} \{\xi^\top A \xi\}$  and  $a - b = \partial_t \phi(t_\phi, x_\phi, y_\phi)$ .

### 3.3 Background results for switching systems

We will study the following switching system

$$F_i(t, x, u, \partial_t u_i, Du_i, D^2 u_i) = 0, \quad \text{in } Q_T, \quad (3.3.1)$$

$$u_i(0, x) = \Psi_0(x), \quad \text{for } x \in \overline{\Omega}, \quad (3.3.2)$$

$$u_i(t, x) = \Psi_1(t, x), \quad \text{for } (t, x) \in (0, T] \times \partial\Omega, \quad (3.3.3)$$

for all  $i \in \mathcal{I} := \{1, \dots, M\}$  with

$$F_i(t, x, r, p_t, p_x, X) = \max \left\{ p_t + \sup_{\alpha \in \mathcal{A}_i} \mathcal{L}_i^\alpha(t, x, r_i, p_x, X); r_i - \mathcal{M}_i r \right\}, \quad (3.3.4)$$

$$\mathcal{L}_i^\alpha(t, x, s, q, X) = -\text{tr}[a_i^\alpha(t, x)X] - b_i^\alpha(t, x)q - c_i^\alpha(t, x)s - f_i^\alpha(t, x), \quad (3.3.5)$$

$$\mathcal{M}_i r = \min_{j \neq i} \{r_j + k\}, \quad (3.3.6)$$

and  $\Psi_0 \in \mathcal{C}^{0,1}(\overline{\Omega})$ ,  $\Psi_1$  is a bounded function Lipschitz in  $x$ , and Hölder continuous in time with exponent  $\frac{1}{2}$ ,  $r \in \mathbb{R}^M$ , and  $k > 0$  is a constant representing the switching cost.

We work under the following assumption related to the coefficients and the boundary data.

**(A1)** For any  $i \in \mathcal{I}$ ,  $\mathcal{A}_i$  is a compact separable metric space. For any  $i \in \mathcal{I}$  and  $\alpha \in \mathcal{A}_i$ , let  $a_i^\alpha = \frac{1}{2}\sigma_i^\alpha \sigma_i^{\alpha,T}$  for some  $d \times P$  matrix  $\sigma_i^\alpha$ . Furthermore, there is a constant  $C$  independent of  $i, \alpha, t$ , such that

$$|\Psi_0|_1 + |\Psi_1|_1 + |\sigma_i^\alpha(t, \cdot)|_1 + |b_i^\alpha(t, \cdot)|_1 + |c_i^\alpha(t, \cdot)|_1 + |f_i^\alpha(t, \cdot)|_1 \leq \bar{C}.$$

Regarding the regularity of the boundary, we adapt the ideas in [9, 20] (reproduced in Section 2.1) to obtain sufficient conditions under which a switching system satisfies the same boundary condition for all the regimes. We only consider that the same boundary condition is applied to all the regimes as we are interested in switching systems as approximations to HJB equations.

We start by adapting the definition of  $\Gamma_{out}$  from Section 2.1

$$\Gamma_{out}(t_0, i) := \left\{ x \in \partial\Omega \left| \begin{array}{l} \exists \zeta \in Z(x) \text{ such that} \\ \forall \alpha \in \mathcal{A}_i, \sigma_i^{\alpha,T}(t_0, x) D\zeta(x) \neq 0 \text{ or} \\ \text{tr}[a_i^\alpha(t_0, x) D^2\zeta(x)] - b_i^\alpha(t_0, x) D\zeta(x) < 0 \end{array} \right. \right\}, \quad (3.3.7)$$

where  $(t_0, i) \in (0, T] \times \mathcal{I}$  and the set  $Z(x)$  is defined in (2.1.2). The next assumption requires that the set  $Z(x)$  is non-empty and that the boundary conditions are satisfied pointwise for the  $M$  regimes.

**(A2)** For any  $x \in \partial\Omega$ , the set  $Z(x)$  defined in (2.1.2) is nonempty. Furthermore, for all  $(t_0, i) \in (0, T] \times \mathcal{I}$  we have that  $\Gamma_{out}(t_0, i) = \partial\Omega$ .

As discussed in [20] (and reproduced in Section 2.1), the set  $Z(x)$  is not empty if  $\Omega$  satisfies an exterior sphere condition. Using assumption (A2), we can then derive the equivalent of Proposition 4.1 in [9] using the extension to non-smooth domains proposed in [20].

**Proposition 3.3.1.** *Assume that (A1) and (A2) hold. Let  $u \in USC(\bar{Q}_T; \mathbb{R}^M)$  be a subsolution of (3.3.1)–(3.3.3) bounded above and  $v \in LSC(\bar{Q}_T; \mathbb{R}^M)$  be a supersolution of (3.3.1)–(3.3.3) bounded below. Then, for all  $(t_0, i) \in (0, T] \times \mathcal{I}$*

$$u_i(t_0, x_0) \leq \Psi_1(t_0, x_0) \leq v_i(t_0, x_0) \text{ for all } x_0 \in \Gamma_{out}(t_0, i).$$

*Sketch of Proof.* By assumption (A2) and using the control perspective as in [9], for all  $(t_0, i) \in (0, T] \times \mathcal{I}$  we have that  $\Gamma_{out}(t_0, i) = \partial\Omega$ . This implies that for any  $x_0 \in \partial\Omega$ , all the trajectories starting at  $x_0 \in \partial\Omega$  are expected to exit the domain for any of the  $M$  regimes. Furthermore, as we are prescribing the same boundary condition  $\Psi_1$  on all the regimes, there is no switching at  $x_0$ .  $\square$

We seek to establish comparison, existence, uniqueness, and  $L^\infty$  bounds on the solution and its gradient. Before stating the results, and for the sake of readability, we state two useful properties of operators (3.3.4) and (3.3.5) and recall some classic results. The first result contains a key technical lemma due to [45] allowing to ignore the switching part of the equation  $u_i - \mathcal{M}_i u$  when doubling variables and seeking to apply the maximum principle for semicontinuous functions in Theorem 3.2.3. The second one contains a useful structural property of (3.3.5) key for the proof of the comparison principle.

**Lemma 3.3.2** (Lemma A.2 in [8]). *Let  $u \in USC(Q_T; \mathbb{R}^M)$  be a bounded above subsolution of (3.3.1) and  $\bar{u} \in LSC(Q_T; \mathbb{R}^M)$  be a bounded below supersolution of another equation (3.3.1) where the functions  $\mathcal{L}_i^\alpha$  are replaced by functions  $\bar{\mathcal{L}}_i^\alpha$  satisfying the same assumptions. Let  $\phi : (0, T) \times \Omega \times \Omega \rightarrow \mathbb{R}$  be a smooth function bounded from below. We denote*

$$\psi_i(t, x, y) := u_i(t, x) - \bar{u}_i(t, y) - \phi(t, x, y),$$

and  $\bar{M} := \sup_{i,t,x,y} \psi_i(t,x,y)$ . If there exists a maximum point for  $\bar{M}$ , i.e. a point  $(i', t_0, x_0, y_0) \in \mathcal{I} \times (0, T) \times \Omega \times \Omega$  such that  $\bar{M} = \psi_{i'}(t_0, x_0, y_0)$ , then there exists  $i_0 \in \mathcal{I}$  such that  $(i_0, t_0, x_0, y_0)$  is also a maximum point for  $\bar{M}$ , and, in addition  $\bar{u}_{i_0}(t_0, y_0) < \mathcal{M}_{i_0} \bar{u}(t_0, y_0)$ .

For the next result we also reproduce the proof as the treatment of the second order terms will be employed later on.

**Lemma 3.3.3** (Lemma V.7.1 in [34]). *Let  $\mathcal{L}_i^\alpha$  be as in (3.3.5). Assume (A1). Then, there exists a continuous function  $\omega : [0, \infty) \rightarrow [0, \infty)$ , independent of  $i$ , satisfying  $\omega(0^+) = 0$  such that*

$$\sup_{\alpha \in \mathcal{A}_i} \mathcal{L}_i^\alpha(t, y, r, \beta(x-y), Y) - \sup_{\alpha \in \mathcal{A}_i} \mathcal{L}_i^\alpha(t, x, r, \beta(x-y), X) \leq \omega(\beta|x-y|^2 + |x-y|), \quad (3.3.8)$$

for every  $(t, x), (t, y) \in Q_T$ ,  $\beta > 0$ , and symmetric matrices  $X, Y \in \mathcal{S}^d$  satisfying (3.2.2).

*Proof.* Set  $p_\beta = \beta(x-y)$ ,  $D = \sigma_i^\alpha(t, x)$ ,  $C = \sigma_i^\alpha(t, y)$  so that

$$\begin{aligned} \sup_{\alpha \in \mathcal{A}_i} \mathcal{L}_i^\alpha(t, y, r, \beta(x-y), Y) - \sup_{\alpha \in \mathcal{A}_i} \mathcal{L}_i^\alpha(t, x, r, \beta(x-y), X) &\leq \sup_{\alpha \in \mathcal{A}_i} \left\{ \frac{1}{2} \text{tr}[DD^T X - CC^T Y] \right\} \\ &+ \sup_{\alpha \in \mathcal{A}_i} \{ |b_i^\alpha(t, x) - b_i^\alpha(t, y)| |p_\beta| + |c_i^\alpha(t, x) - c_i^\alpha(t, y)| |r| + |f_i^\alpha(t, x) - f_i^\alpha(t, y)| \}. \end{aligned}$$

By (A1)

$$\begin{aligned} |b_i^\alpha(t, x) - b_i^\alpha(t, y)| |p_\beta| &\leq C|x-y| |p_\beta| \leq \bar{C}\beta|x-y|^2, \\ |c_i^\alpha(t, x) - c_i^\alpha(t, y)| |r| + |f_i^\alpha(t, x) - f_i^\alpha(t, y)| &\leq \bar{C}|x-y|. \end{aligned}$$

Now use (3.2.2) to obtain

$$\begin{aligned}
\text{tr}[DD^T X - CC^T Y] &= \text{tr} \left( \begin{bmatrix} DD^T & DC^T \\ CD^T & CC^T \end{bmatrix} \begin{bmatrix} X & 0 \\ 0 & -Y \end{bmatrix} \right) \\
&\leq 3\beta \text{tr} \left( \begin{bmatrix} DD^T & DC^T \\ CD^T & CC^T \end{bmatrix} \begin{bmatrix} I & -I \\ -I & I \end{bmatrix} \right) \\
&= 3\beta \text{tr}(DD^T - DC^T - CD^T + CC^T) \\
&= 3\beta \text{tr}([D - C][D - C]^T) \\
&= 3\beta \|D - C\|^2 = 3\beta \|\sigma_i^\alpha(t, x) - \sigma_i^\alpha(t, y)\|^2 \\
&\leq \bar{C}\beta |x - y|^2.
\end{aligned}$$

□

The following classical results regarding the extension of Lipschitz continuous functions will be used to prove some of the results, see [57] for further details.

**Lemma 3.3.4** (Lemma 6.3 in [42]). *If  $\{f_\zeta\}_{\zeta \in Z}$  is a family of real valued functions on a metric space  $X$  with modulus of continuity*

$$\bar{\omega}(\delta, f_\zeta) := \sup_{d_X(x, y) \leq \delta} |f_\zeta(x) - f_\zeta(y)|,$$

*such that  $\bar{\omega}(\delta, f_\zeta) \leq h(\delta)$  for some continuous subadditive function  $h$ . Then,  $f := \inf_{\zeta \in Z} f_\zeta$  and  $F := \sup_{\zeta \in Z} f_\zeta$  are also continuous with modulus of continuity bounded by  $h$ , provided their respective infima and suprema are finite at one point.*

**Theorem 3.3.5** (Theorem 2 in [57]). *If  $f(x)$  is a real function defined on a subset  $E$  of a metric space  $S$ , and  $f(x)$  has a modulus of continuity  $\bar{\omega}(\delta)$  which is concave for  $\delta \geq 0$  and which approaches zero with  $\delta$ , then  $f(x)$  can be extended to  $S$  preserving the modulus of continuity  $\bar{\omega}(\delta)$ .*

**Corollary 3.3.6** (McShane's Theorem. Corollary 1 in [57]). *If  $f(x)$  is a real function defined on a subset  $E$  of a metric space  $S$ , and  $f(x)$  satisfies on  $E$  a Lipschitz or Hölder condition*

$$|f(x_1) - f(x_2)| \leq L d_E(x_1, x_2)^\alpha,$$

*where  $\alpha \in (0, 1]$ , then  $f(x)$  can be extended to  $S$  preserving the Lipschitz or Hölder condition with the same constant  $L$ .*

Using these results we can now state and prove a modified version of Theorem A.1 in [8]. This theorem is different from the original in that the equations are defined on a bounded spatial domain with Dirichlet boundary conditions satisfied in the classical sense. More specifically, in [8] the spatial domain is  $\mathbb{R}^d$  and no boundary conditions are imposed. As a result of the change in the domain, the proof of the Lipschitz regularity in space requires an extra assumption equivalent to that in Theorem VII.1 in [46].

**Theorem 3.3.7.** *Assume (A1) and (A2) hold.*

(i) *If  $u \in USC(\bar{Q}_T; \mathbb{R}^M)$  is a subsolution of (3.3.1)–(3.3.3) bounded above and  $v \in LSC(\bar{Q}_T; \mathbb{R}^M)$  a supersolution of (3.3.1)–(3.3.3) bounded below satisfying  $u \leq v$  in  $\partial^* Q_T$ , then  $u \leq v$  in  $\bar{Q}_T$ .*

(ii) *There exists a unique bounded continuous solution  $u$  of (3.3.1)–(3.3.3).*

(iii) *If for any  $t \in [0, T]$  the solution  $u$  of (3.3.1)–(3.3.3) satisfies*

$$|u(t, x) - u(t, y)| \leq L|x - y|, \quad \forall (x, y) \in \partial(\Omega \times \Omega), \quad (3.3.9)$$

*then it belongs to  $C^{0,1}(\bar{Q}_T)$ , i.e. the space of bounded continuous functions with finite  $|\cdot|_1$  norm, and satisfies for all  $t, s \in [0, T]$*

$$e^{-\lambda t} \max_i |u_i(t, \cdot)|_0 \leq \max_{j \in \{0,1\}} |\Psi_j|_0 + t \sup_{i,\alpha} |f_i^\alpha|_0,$$

where  $\lambda := \sup_{i,\alpha} |c_i^{\alpha+}|_0$ ,

$$e^{-\lambda_0 t} \max_i [u_i(t, \cdot)]_1 \leq \max_{j \in \{0,1\}} [\Psi_j]_1 + L + t \sup_{i,\alpha,s} \left\{ |u_i|_0 [c_i^\alpha(s, \cdot)]_1 + [f_i^\alpha(s, \cdot)]_1 \right\},$$

where  $\lambda_0 := \sup_{i,\alpha,s} \{|c_i^{\alpha+}(s, \cdot)|_0 + [\sigma_i^\alpha(s, \cdot)]_1^2 + [b_i^\alpha(s, \cdot)]_1\}$ , and

$$\max_i |u_i(t, x) - u_i(s, x)| \leq C |t - s|^{1/2},$$

where  $C \leq 8\bar{M}\bar{C} + \sqrt{T}\bar{C}(2\bar{M} + 1)$  and  $\bar{M} := \sup_{i,t} |u_i(t, \cdot)|_1$ .

*Proof.* To prove (i) for the parabolic case with strong boundary conditions, we adapt the proof of Theorem 8.2 in [23] using the parabolic version of the maximum principle for semicontinuous functions as in Theorem 3.2.3. We start by noticing that for  $\rho > 0$ ,  $u^\rho = u - \rho/(T - t)$  is also a subsolution of (3.3.1). We assume by contradiction that  $u_i(s, z) - v_i(s, z) > 0$  for some  $(i, s, z) \in \mathcal{I} \times Q_T$ . For some  $\rho, \gamma, \beta > 0$  consider the auxiliary function

$$\Phi(i, t, x, y) = u_i^\rho(t, x) - v_i(t, y) - \beta|x - y|^2 - \gamma t, \quad t \in [0, T], x, y \in \bar{\Omega},$$

where  $\rho, \gamma, \beta$  are chosen such that

$$\bar{M} := \sup_{\mathcal{I} \times [0, T] \times \bar{\Omega} \times \bar{\Omega}} \Phi(i, t, x, y) > 0.$$

Assume that the maximum is attained for  $(\tilde{i}, \hat{t}, \hat{x}, \hat{y})$ . Then, by standard arguments for viscosity solutions, for  $\beta$  big enough we have that the maximum of  $\Phi$  is attained for  $(\hat{t}, \hat{x}, \hat{y}) \in (0, T) \times \Omega \times \Omega$ , see Lemma 3.1 in [23], and  $\beta|\hat{x} - \hat{y}|^2 \rightarrow 0$ . Therefore,  $\lim_{\beta \rightarrow \infty} \beta|\hat{x} - \hat{y}| = 0$  and as  $u_i^\rho(t, z) - v_i(t, z) \leq 0$  for any  $z \in \partial\Omega$  we have that for  $\beta$  big enough  $(\hat{x}, \hat{y}) \in \Omega \times \Omega$ .<sup>2</sup>

---

<sup>2</sup>In [23]  $(\tilde{i}, \hat{t}, \hat{x}, \hat{y})$  are denoted as  $(\tilde{i}_\beta, \hat{t}_\beta, \hat{x}_\beta, \hat{y}_\beta)$  to make explicit their dependence in  $\beta$ .



Moreover, by Lemma 3.3.2 there exists  $\hat{i} \in \mathcal{I}$  such that  $(\hat{i}, \hat{t}, \hat{x}, \hat{y})$  is also a maximum point for  $\Phi$ , and, in addition  $v_{\hat{i}}(\hat{t}, \hat{y}) < \mathcal{M}_{\hat{i}}v(\hat{t}, \hat{y})$ . Now we can make use of Theorem 3.2.1 and Theorem 3.2.3 with  $u_{\hat{i}}^\rho$ ,  $v_{\hat{i}}$  and

$$\phi(t, x, y) = \beta|x - y|^2 + \gamma t,$$

to infer that there are numbers  $a, b$  and symmetric matrices  $X, Y \in \mathcal{S}^d$  such that

$$(a, \beta(\hat{x} - \hat{y}), X) \in \overline{\mathcal{P}}^{2,+} u_{\hat{i}}^\rho(\hat{t}, \hat{x}), \quad (b, \beta(\hat{x} - \hat{y}), Y) \in \overline{\mathcal{P}}^{2,-} v_{\hat{i}}(\hat{t}, \hat{y}),$$

satisfying

$$a - b = \gamma, \quad \text{and} \quad -3\beta \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \leq \begin{pmatrix} X & 0 \\ 0 & -Y \end{pmatrix} \leq 3\beta \begin{pmatrix} I & -I \\ -I & I \end{pmatrix}.$$

By sub- and supersolution properties of  $u^\rho$  and  $v$  we have that

$$a + \sup_{\alpha \in \mathcal{A}_{\hat{i}}} \mathcal{L}_{\hat{i}}^\alpha(\hat{t}, \hat{x}, u^\rho(\hat{t}, \hat{x}), \beta(\hat{x} - \hat{y}), X) \leq 0,$$

$$b + \sup_{\alpha \in \mathcal{A}_{\hat{i}}} \mathcal{L}_{\hat{i}}^\alpha(\hat{t}, \hat{y}, v(\hat{t}, \hat{y}), \beta(\hat{x} - \hat{y}), Y) \geq 0.$$

Subtracting these two inequalities and using Lemma 3.3.3 we have that

$$\begin{aligned} \gamma = a - b &\leq \sup_{\alpha \in \mathcal{A}_{\hat{i}}} \mathcal{L}_{\hat{i}}^\alpha(\hat{t}, \hat{y}, v(\hat{t}, \hat{y}), \beta(\hat{x} - \hat{y}), Y) - \sup_{\alpha \in \mathcal{A}_{\hat{i}}} \mathcal{L}_{\hat{i}}^\alpha(\hat{t}, \hat{x}, u^\rho(\hat{t}, \hat{x}), \beta(\hat{x} - \hat{y}), X) \\ &\leq \omega(\beta|\hat{x} - \hat{y}|^2 + |\hat{x} - \hat{y}|) \rightarrow 0 \end{aligned}$$

which leads to a contradiction as  $\beta \rightarrow \infty$ , because  $0 < \gamma \leq 0$ , and concludes the proof of the comparison principle in (i).

Uniqueness and existence results for viscosity solutions of switching systems sat-

isfying Dirichlet boundary conditions in the weak sense are proved in [45, 49] for the elliptic case. As the operator in [45, 49] is allowed to be degenerate, the parabolic case can be inferred from these, whereas the satisfaction of the boundary conditions in the strong sense is guaranteed by (A2) following [9, 20].

We focus now on proving (iii) starting with the boundedness of the solution in the  $L^\infty$ -norm. Let

$$w(t) := e^{\lambda t} \left( \max_{j \in \{0,1\}} |\Psi_j|_0 + t \sup_{i,\alpha} |f_i^\alpha|_0 \right),$$

to prove the bound on  $|u|_0$  we start by proving that  $w$  (or  $-w$ ) is a supersolution (or subsolution) of (3.3.1)–(3.3.3). Here we only check the supersolution property as the subsolution property is established similarly. The function  $w$  is a classical supersolution of (3.3.1) as for any  $i \in \mathcal{I}$

$$\begin{aligned} F_i(t, x, w, \partial_t w, 0, 0) &= \lambda w(t) + e^{\lambda t} \sup_{i,\alpha} |f_i^\alpha|_0 + \sup_{\alpha} [-c_i^\alpha(t, x)w - f_i^\alpha(t, x)] \\ &\geq w(t)(\lambda - c_i^a(t, x)) + e^{\lambda t} \sup_{i,\alpha} |f_i^\alpha|_0 - f_i^a(t, x) \\ &\geq 0, \end{aligned}$$

for any  $a \in \mathcal{A}$ . This function also satisfies that

$$\begin{aligned} w(0) &\geq \Psi_0(x) \text{ for all } x \in \overline{\Omega}, \\ w(t) &\geq \Psi_1(t, x) \text{ for all } (t, x) \in (0, T] \times \partial\Omega. \end{aligned}$$

Hence by the comparison principle  $u_i(t, x) \leq w(t)$  for all  $(i, t, x) \in \mathcal{I} \times [0, T] \times \overline{\Omega}$ . Proceeding similarly with  $-w$  we obtain the bound on  $|u|_0$ .

To establish the Lipschitz regularity of the solution  $u$  we define

$$m := \sup_{i,t,x,y} \{u_i(t,x) - u_i(t,y) - \bar{w}(t)|x - y|\},$$

where

$$\bar{w}(t) := e^{\lambda_0 t} \left\{ \max_{j \in \{0,1\}} [\Psi_j]_1 + L + t \sup_{i,\alpha,s} \left\{ |u_i|_0 [c_i^\alpha(s, \cdot)]_1 + [f_i^\alpha(s, \cdot)]_1 \right\} \right\},$$

and  $L$  is the constant in the assumption (3.3.9). Following [8], the result follows if we can prove that  $m \leq 0$ . We proceed by contradiction assuming that  $m > 0$  and that the maximum is attained for  $\bar{i}, \bar{t}, \bar{x}, \bar{y}$ . First of all, we notice if  $m > 0$  then there exists a  $\eta > 0$  such that

$$u_{\bar{i}}(\bar{t}, \bar{x}) - u_{\bar{i}}(\bar{t}, \bar{y}) - \bar{w}(\bar{t})|\bar{x} - \bar{y}| - \bar{t}e^{\lambda_0 \bar{t}}\eta > 0.$$

Thus, we define an auxiliary function  $\psi_i(t, x, y) := u_i(t, x) - u_i(t, y) - \bar{w}(t)|x - y| - te^{\lambda_0 t}\eta$ , then  $\psi$  also has maximum  $\bar{M}$  at some point  $(\tilde{i}, \tilde{t}, \tilde{x}, \tilde{y})$ . By construction of  $\psi_i$  (choice of  $\eta$ ), the maximum is also strictly positive, i.e.  $\bar{M} > 0$ . Thus, by definition of  $\bar{w}(t)$  we infer that  $\tilde{t} > 0$ ,  $\tilde{x} \neq \tilde{y}$  and  $(\tilde{x}, \tilde{y}) \notin \partial(\Omega \times \Omega)$ , by the assumption in (3.3.9).

Now we check whether the maximum's location could be on the interior of the domain, that is if  $(\tilde{i}, \tilde{t}, \tilde{x}, \tilde{y}) \in (0, T) \times \Omega \times \Omega$ . As noted in [8],  $\bar{w}(t)|x - y| + te^{\lambda_0 t}\eta$  is a smooth function at  $(\tilde{t}, \tilde{x}, \tilde{y})$ , therefore, proceeding as in the proof of the comparison principle in (i), we can use Theorem 3.2.3 and Lemma 3.3.2 to ignore the switching part for the supersolution, to obtain that  $\eta \leq 0$ . This is a contradiction and hence  $m \leq 0$ .

Regarding the time regularity result, by assumption the solution  $u$  satisfies the boundary conditions pointwise, therefore for all  $(i, x) \in \mathcal{I} \times \partial\Omega$ ,  $u_i$  is Hölder continuous in time with exponent  $1/2$  if  $\Psi_1$  is. For  $x \in \Omega$  assume that  $s < t$  and let  $u^\varepsilon$  with

$\varepsilon > 0$  be the solution of the following system

$$F_i(t, x, u^\varepsilon, \partial_t u_i^\varepsilon, Du_i^\varepsilon, D^2 u_i^\varepsilon) = 0, \quad \text{for } (t, x) \in (s, T] \times \Omega, \quad (3.3.10)$$

$$u_i^\varepsilon(s, x) = \Psi_{0,i}^\varepsilon(x), \quad \text{for } x \in \overline{\Omega}, \quad (3.3.11)$$

$$u_i^\varepsilon(t, x) = \Psi_{1,i}^\varepsilon(t, x), \quad \text{for } (t, x) \in (s, T] \times \partial\Omega. \quad (3.3.12)$$

where the initial and boundary conditions are set by first extending each element of the solution  $u_i$  of the switching system (3.3.1)–(3.3.3) according to McShane's Theorem and then mollifying the resulting function in space using the mollifiers defined in (3.1.8). However, we need to construct the boundary conditions such that it is not optimal to switch at the boundary.

To do so, let  $L_u^*(t) := \max_{i \in \mathcal{I}} [u_i(t, \cdot)]_1$  for any  $t \in [s, T]$ . Then, without loss of generality, fixing  $i = 1$  for any  $(r, z) \in [s, T] \times \mathbb{R}^d$  define

$$y_*(r, z) = \operatorname{argmin}_{y \in \overline{\Omega}} (u_1(r, y) + L_u^*(r)|z - y|).$$

Based on the functions  $L_u^*$  and  $y_*$ , for each  $i \in \mathcal{I}$  and any  $(r, z) \in [s, T] \times \mathbb{R}^d$  define

$$\tilde{U}_i(r, z) := u_i(r, y_*(r, z)) + L_u^*(r)|z - y_*(r, z)|,$$

now we set

$$\Psi_{0,i}^\varepsilon(x) := \tilde{U}_i(s, \cdot) * \rho_\varepsilon(x), \quad \text{for } (i, x) \in \mathcal{I} \times \overline{\Omega}, \quad (3.3.13)$$

$$\Psi_{1,i}^\varepsilon(t, x) := \tilde{U}_i(t, \cdot) * \rho_\varepsilon(x), \quad \text{for } (i, t, x) \in \mathcal{I} \times (s, T] \times \partial\Omega. \quad (3.3.14)$$

By definition of  $\Psi_1$ ,  $u_i(t, x)$  is Hölder continuous in time with exponent  $1/2$  for  $(i, x) \in \mathcal{I} \times \partial\Omega$ . By construction  $\tilde{U}_i(t, x)$  is also Hölder continuous in time with exponent  $1/2$  for  $(i, x) \in \mathcal{I} \times \partial\Omega$ . Hence, for  $x \in \partial\Omega$ , let  $y \in \mathbb{R}^d$  be such that

$|y - x| \leq \varepsilon$ , then for any  $t, r \in [s, T]$

$$\begin{aligned} |\tilde{U}_i(t, y) - \tilde{U}_i(r, y)| &\leq |\tilde{U}_i(t, y) - u_i(t, x)| + |u_i(t, x) - u_i(r, x)| + |u_i(r, x) - \tilde{U}_i(r, y)| \\ &\leq \left( \max_{\tau \in \{t, r\}} L_u^*(\tau) \wedge [\Psi_1]_1 \right) (\sqrt{|t - r|} + 2\varepsilon). \end{aligned}$$

The existence of such points  $y$  is a consequence of (A2). Recall that this is because the set  $Z(x)$  in (A2) is not empty under the exterior ball condition, see Section 2.1 or the discussion of assumption (H4) in [20] for the original reference.

We can use the previous estimate to derive the next bound for  $\Psi_{1,i}^\varepsilon$  as defined in (3.3.14). For any  $(t, r, x) \in (s, T]^2 \times \partial\Omega$  we have that

$$|\Psi_{1,i}^\varepsilon(t, x) - \Psi_{1,i}^\varepsilon(r, x)| \leq K(\sqrt{|t - r|} + 2\varepsilon), \quad (3.3.15)$$

where  $K > 0$  is a constant independent of  $t, r$  or  $\varepsilon$ .

Given that both  $u$  and  $u^\varepsilon$  are solutions of (3.3.1) in  $Q_T^s := (s, T] \times \Omega$ , by the comparison principle we have that for all  $(i, t) \in \mathcal{I} \times [s, T]$

$$u_i - u_i^\varepsilon \leq e^{\lambda t} \max_i \sup_{\partial^* Q_T^s} (u_i - u_i^\varepsilon)^+, \quad \text{and} \quad u_i^\varepsilon - u_i \leq e^{\lambda t} \max_i \sup_{\partial^* Q_T^s} (u_i^\varepsilon - u_i)^+.$$

We can estimate the difference between  $u_i^\varepsilon$  and  $u_i$  at the parabolic boundary of  $Q_T^s$  using Lipschitz regularity of  $u$  and mollifiers' properties. Thus, for  $(i, x) \in \mathcal{I} \times \overline{\Omega}$

$$|u_i(s, x) - u_i^\varepsilon(s, x)| = \left| u_i(s, x) - \int_{|e| < \varepsilon} \tilde{U}_i(s, x - e) \rho_\varepsilon(e) \, de \right| \leq L_u^*(s) \varepsilon,$$

and similarly for  $(i, t, x) \in \mathcal{I} \times (s, T] \times \partial\Omega$

$$|u_i(t, x) - u_i^\varepsilon(t, x)| = \left| u_i(t, x) - \int_{|e| < \varepsilon} \tilde{U}_i(t, x - e) \rho_\varepsilon(e) \, de \right| \leq L_u^*(t) \varepsilon.$$

Therefore,

$$u_i - u_i^\varepsilon \leq \sup_{t \in [s, T]} e^{\lambda t} L_u^*(t) \varepsilon.$$

To obtain the time regularity result we will make use of this estimate in the following way

$$\begin{aligned} |u_i(t, x) - u_i(s, x)| &\leq |u_i(t, x) - u_i^\varepsilon(t, x)| + |u_i^\varepsilon(t, x) - \Psi_{0,i}^\varepsilon(x)| + |\Psi_{0,i}^\varepsilon(x) - u_i(s, x)| \\ &\leq 2 \sup_{t \in [s, T]} e^{\lambda t} L_u^*(t) \varepsilon + |u_i^\varepsilon(t, x) - \Psi_{0,i}^\varepsilon(x)|, \end{aligned}$$

where  $i \in \mathcal{I}$ .

To bound the last term consider the smooth function

$$w_{\varepsilon,i}^\pm(t, x) := \Psi_{0,i}^\varepsilon(x) \pm e^{\lambda(t-s)} \left[ (t-s)C_\varepsilon + K(\sqrt{|t-s|} + 2\varepsilon) \right],$$

it is straightforward to show that  $w_\varepsilon^+$  is a (classical) supersolution and  $w_\varepsilon^-$  is a (classical) subsolution of (3.3.10)–(3.3.12) given the growth estimate of the boundary condition in (3.3.15) and provided that we set

$$C_\varepsilon = \bar{C}^2 |D^2 \Psi_0^\varepsilon|_0 + \bar{C} (|D \Psi_0^\varepsilon|_0 + |\Psi_0^\varepsilon|_0 + 1), \quad \text{and} \quad \lambda = \sup_{i, \alpha} |c_i^{\alpha,+}|_0.$$

We only check the supersolution property as checking the subsolution property is analogous. For  $t = s$  (initial time)

$$u^\varepsilon(s, x) = \Psi_0^\varepsilon(x) \leq w_\varepsilon^+(s, x).$$

Next, for  $x \in \partial\Omega$  the solution  $u^\varepsilon$  satisfies the spatial boundary condition, therefore

by definition of  $\Psi_1^\varepsilon$  and (3.3.15), for any  $(i, t, x) \in \mathcal{I} \times (s, T] \times \partial\Omega$

$$\begin{aligned} u_i^\varepsilon(t, x) - w_{\varepsilon, i}^+(t, x) &= \Psi_{1, i}^\varepsilon(t, x) - \Psi_{1, i}^\varepsilon(s, x) - e^{\lambda(t-s)} \left[ (t-s)C_\varepsilon + K(\sqrt{|t-s|} + 2\epsilon) \right] \\ &\leq K(\sqrt{|t-s|} + 2\epsilon) \left[ 1 - e^{\lambda(t-s)} \right] - e^{\lambda(t-s)}(t-s)C_\varepsilon \leq 0. \end{aligned}$$

Finally, it is easy to check that

$$F_i(t, x, w_\varepsilon^+, \partial_t w_\varepsilon^+, Dw_\varepsilon^+, D^2 w_\varepsilon^+) \geq 0,$$

as

$$\partial_t w_\varepsilon^+ + \sup_{\alpha \in \mathcal{A}_i} \mathcal{L}_i^\alpha(t, x, w_\varepsilon^+, D\Psi_0^\varepsilon, D^2\Psi_0^\varepsilon) \geq 0.$$

Hence, applying the comparison principle we obtain that

$$w_\varepsilon^- \leq u^\varepsilon \leq w_\varepsilon^+ \quad \text{in} \quad [s, T] \times \bar{\Omega}.$$

The result now follows from

$$|u_i(t, x) - u_i(s, x)| \leq 2 \sup_{t \in [s, T]} e^{\lambda t} L_u^*(t) \varepsilon + e^{\lambda(t-s)} |t-s| C_\varepsilon + e^{\lambda(t-s)} K(\sqrt{|t-s|} + 2\epsilon),$$

and a minimization with respect to  $\varepsilon$  of the right-hand side after noting that  $C_\varepsilon \leq C(\varepsilon^{-1} + 1)$ .  $\square$

**Theorem 3.3.8** (Theorem A.3 in [8]). *Let  $u$  and  $\bar{u}$  be solutions of (3.3.1)–(3.3.3) with coefficients  $\sigma, b, c, f$  and  $\bar{\sigma}, \bar{b}, \bar{c}, \bar{f}$  respectively. If both sets of coefficients and the domain satisfy (A1) and (A2), and  $|u|_0 + |\bar{u}|_0 + [u(t, \cdot)]_1 + [\bar{u}(t, \cdot)]_1 \leq \bar{M} < \infty$  for*

$t \in [0, T]$ , then

$$\begin{aligned} e^{-\lambda t} \max_i |u_i(t, \cdot) - \bar{u}_i(t, \cdot)|_0 &\leq \max_i \sup_{\partial^* Q_T} |u_i - \bar{u}_i| + t^{1/2} K \sup_{i, \alpha} |\sigma^\alpha - \bar{\sigma}^\alpha|_0 \\ &\quad + t \sup_{i, \alpha} \left\{ 2\bar{M} |b^\alpha - \bar{b}^\alpha|_0 + \bar{M} |c^\alpha - \bar{c}^\alpha|_0 + |f^\alpha - \bar{f}^\alpha|_0 \right\}, \end{aligned}$$

where  $\lambda := \sup_{i, \alpha} |c^-|_0$  and

$$\begin{aligned} K^2 &\leq 8\bar{M}^2 + 8\bar{M}T \sup_{i, \alpha} \left\{ 2\bar{M} [\sigma^\alpha]_1^2 \wedge [\bar{\sigma}^\alpha]_1^2 \right. \\ &\quad \left. + 2\bar{M} [b^\alpha]_1 \wedge [\bar{b}^\alpha]_1 + \bar{M} [c^\alpha]_1 \vee [\bar{c}^\alpha]_1 + [f^\alpha]_1 \wedge [\bar{f}^\alpha]_1 \right\}. \end{aligned}$$

*Sketch of Proof.* As done in the proof of Theorem A.3 in [8], without loss of generality we assume that  $\lambda = 0$ . We start by defining some auxiliary functions

$$\begin{aligned} \psi^i(t, x, y) &:= u_i(t, x) - \bar{u}_i(t, y) - \frac{1}{\delta} |x - y|^2, \\ m &:= \sup_{i, t, x, y} \psi^i(t, x, y) - \sup_{\mathcal{I} \times Q^*} (\psi^i(t, x, y))^+, \\ \bar{m} &:= \sup_{i, t, x, y} \left\{ \psi^i(t, x, y) - \frac{\eta m t}{T} \right\}, \end{aligned}$$

where  $\eta \in (0, 1)$  and  $Q^* := \{(t, x, y) \in (\{0\} \times \bar{\Omega} \times \bar{\Omega}) \cup ((0, T] \times \partial(\Omega \times \Omega))\}$ . The aim is to obtain an upper bound for  $m$  using the fact that  $u$  and  $\bar{u}$  are viscosity solutions (and therefore sub- and supersolution to the corresponding equation). Let  $m \leq 0$  and assume that the supremum in the second term is attained for  $(\bar{t}, \bar{x}, \bar{y}) \in (0, T] \times \partial\Omega \times \Omega$ , then by Lipschitz regularity of  $u_i$

$$\begin{aligned} \sup_{\mathcal{I} \times Q^*} (\psi^i(t, x, y))^+ &\leq \sup_{(t, x) \in \partial^* Q_T} |u_i(t, x) - \bar{u}_i(t, x)| + [u_i(t, \cdot)]_1 |\bar{x} - \bar{y}| - \frac{1}{\delta} |\bar{x} - \bar{y}|^2 \\ &\leq \sup_{(t, x) \in \partial^* Q_T} |u_i(t, x) - \bar{u}_i(t, x)| + \frac{\delta}{4} ([u_i(t, \cdot)]_1)^2, \end{aligned}$$

similar bounds can be obtained using the Lipschitz regularity in  $\bar{\Omega}$  of the boundary



and initial conditions for different combinations of  $(\bar{t}, \bar{x}, \bar{y}) \in Q^{*3}$ .

Let  $m > 0$  and consider that the supremum for  $\bar{m}$  is attained at some point  $(i_0, t_0, x_0, y_0)$ . Since  $m > 0$ , arguing by contradiction, it follows that  $(t_0, x_0, y_0) \notin Q^*$ ,  $\bar{m} > 0$  and by Lemma 3.3.2, the index  $i_0$  may be chosen so that  $\bar{u}_{i_0}(t_0, y_0) < \mathcal{M}_{i_0} \bar{u}_{i_0}(t_0, y_0)$ .

The rest is equal to the proof of the original theorem, so we refer to Theorem A.3 in [8] for further details. In a nutshell, as  $(t_0, x_0, y_0) \notin Q^*$  and  $i_0$  is chosen such that the equation applies at the maximum point for  $\bar{m}$ , we can apply the maximum principle as in Theorem 3.2.3 to  $\bar{m}$  and use the resulting inequalities to obtain an upper bound for  $m$ . Then, switching the roles of  $u$  and  $\bar{u}$  as super- and subsolution we can obtain the lower bound.  $\square$

### 3.4 Convergence rate for a switching system

Based on the regularity results from the previous section, we derive the convergence of the switching system to the HJB equation. To do so, we introduce three different second order non-linear parabolic equations and their relations.

We consider the following type of switching systems,

$$\begin{aligned} F_i(t, x, v, \partial_t v_i, Dv_i, D^2 v_i) &= 0 & \text{in } Q_T, \quad i \in \mathcal{I} := \{1, \dots, M\}, \\ v(0, x) &= v_0(x) & \text{in } \bar{\Omega}, \\ v(t, x) &= v_0(t, x) & \text{in } (0, T] \times \partial\Omega, \end{aligned} \tag{3.4.1}$$

where the solution is  $v = (v_1, \dots, v_M)$ , and for  $i \in \mathcal{I}$ ,  $(t, x) \in Q_T$ ,  $r = (r_1, \dots, r_M) \in$

---

<sup>3</sup>We highlight that assuming  $[u(t, \cdot)]_1 + [\bar{u}(t, \cdot)]_1 \leq \bar{M}$ , from the statement of Theorem 3.3.7 implies that both viscosity solutions  $u$  and  $\bar{u}$  satisfy (3.3.9).

$\mathbb{R}^M$ ,  $p_t \in \mathbb{R}$ ,  $p_x \in \mathbb{R}^d$ , and  $X \in \mathcal{S}^d$ ,  $F_i$  is given by

$$F_i(t, x, r, p_t, p_x, X) = \max \left\{ p_t + \sup_{\alpha \in \mathcal{A}_i} \mathcal{L}^\alpha(t, x, r_i, p_x, X); r_i - \mathcal{M}_i r \right\},$$

$\mathcal{A}_i$  is a subset of  $\mathcal{A}$ ,  $\mathcal{L}^\alpha$  is defined in (3.1.4) and  $\mathcal{M}_i r$  in (3.3.6).

Our objective is to obtain a convergence rate for (3.4.1), as  $k \rightarrow 0$  (the switching cost in (3.3.6)), to the following HJB equation

$$\begin{aligned} u_t + \sup_{\alpha \in \tilde{\mathcal{A}}} \mathcal{L}^\alpha(t, x, u, Du, D^2u) &= 0 & \text{in } Q_T, \\ u(0, x) &= \Psi_0(x) & \text{in } \bar{\Omega}, \\ u(t, x) &= \Psi_1(t, x) & \text{in } (0, T] \times \partial\Omega, \end{aligned} \tag{3.4.2}$$

where  $\tilde{\mathcal{A}} = \cup_i \mathcal{A}_i$  and  $\lim_{t \downarrow 0} \Psi_1(t, x) = \Psi_0(x)$  for all  $x \in \partial\Omega$ . Therefore, for obvious reasons we set the initial and boundary data of the switching system as follows:  $v_0 = (\Psi_0, \dots, \Psi_0)$  and  $v_1 = (\Psi_1, \dots, \Psi_1)$ .

Similarly to Proposition 2.1 in [8], under the assumptions in the previous section, the following proposition is a corollary of Theorem 3.3.7.

**Proposition 3.4.1.** *Assume (A1) and (A2) hold. Let  $v$  and  $u$  be the unique viscosity solutions of (3.4.1) and (3.4.2) respectively. If both satisfy that for any  $t \in [0, T]$*

$$|v(t, x) - v(t, y)| + |u(t, x) - u(t, y)| \leq L|x - y|, \quad \forall (x, y) \in \partial(\Omega \times \Omega),$$

then

$$|v|_1 + |u|_1 \leq C,$$

where the constant  $C$  only depends on  $T$ ,  $L$  and  $K$  appearing in (A1). Additionally, if  $w_1$  and  $w_2$  are sub- and supersolutions of (3.4.1) or (3.4.2) satisfying  $w_1(s, z) \leq w_2(s, z)$  for  $(s, z) \in \partial^* Q_T$ , then  $w_1 \leq w_2$ .

To obtain the convergence rate we will use a regularization approach introduced by Krylov [51] which requires the definition of an auxiliary system

$$\begin{aligned} F_i^\varepsilon(t, x, v^\varepsilon, \partial_t v_i^\varepsilon, Dv_i^\varepsilon, D^2 v_i^\varepsilon) &= 0, & \text{in } (0, T + \varepsilon^2] \times \Omega, \quad i \in \mathcal{I}, \\ v_i^\varepsilon(0, x) &= \Psi_0(x), & \text{in } \bar{\Omega}, \\ v_i^\varepsilon(t, x) &= \Psi_1^\varepsilon(t, x), & \text{in } (0, T + \varepsilon^2] \times \partial\Omega, \end{aligned} \quad (3.4.3)$$

where  $v^\varepsilon = (v_1^\varepsilon, \dots, v_M^\varepsilon)$ ,

$$F_i^\varepsilon(t, x, r, p_t, p_x, X) = \max \left\{ p_t + \sup_{\substack{\alpha \in \mathcal{A}_i \\ 0 \leq \eta \leq \varepsilon^2, |\xi| \leq \varepsilon}} \mathcal{L}^\alpha \left( t + \eta, (1 + \kappa\varepsilon)(x + \xi), r_i, \frac{p_x}{1 + \kappa\varepsilon}, \frac{X}{(1 + \kappa\varepsilon)^2} \right); r_i - \mathcal{M}_i r \right\},$$

and the coefficients in the definition of  $\mathcal{L}^\alpha$  in (3.1.4) are extended to the relevant domain according to McShane's Theorem.

We would like to compare  $v$  and  $v^\varepsilon$ , hence we set the boundary conditions to

$$\Psi_1^\varepsilon(t, x) := \begin{cases} u(t, x), & \text{if } (t, x) \in (0, T] \times \partial\Omega, \\ u(T, x) + [u]_1 \sqrt{t - T}, & \text{if } (t, x) \in (T, T + \varepsilon^2] \times \partial\Omega. \end{cases} \quad (3.4.4)$$

The proof of the main result in this section relies on the Lipschitz continuity in space of the solution to the family of switching systems with parameter  $\varepsilon > 0$  in (3.4.3). As in Theorem 3.3.7 to prove the Lipschitz continuity in space we need the following assumption to hold:

**(A3) (Lipschitz regularity)** There exists  $L > 0$  not depending on  $\varepsilon$ , such that for any  $t \in (0, T + \varepsilon^2]$ ,  $v^\varepsilon$ , the solution of (3.4.3), satisfies

$$|v^\varepsilon(t, x) - v^\varepsilon(t, y)| \leq L|x - y|, \quad \forall (x, y) \in \partial(\Omega \times \Omega).$$

From Assumption (A3) together with Theorems 3.3.7 and 3.3.8 we infer the following result.

**Proposition 3.4.2.** *Assume (A1), (A2) and (A3). Let  $v^\varepsilon : [0, T + \varepsilon^2] \times \overline{\Omega} \rightarrow \mathbb{R}^M$  be the unique viscosity solution of (3.4.3). If  $v^\varepsilon$  satisfies (A3), then for all  $i \in \mathcal{I}$*

$$|v_i^\varepsilon|_1 \leq C_1 \text{ in } [0, T + \varepsilon^2] \times \overline{\Omega}, \quad \text{and} \quad \frac{1}{\varepsilon} |v_i^\varepsilon - v_i|_0 \leq C_2 \text{ in } \overline{Q}_T,$$

where  $v$  solves (3.4.1) and the constants  $C_1$  and  $C_2$  only depend on  $T$ ,  $\sup_{x \in \Omega} |x|$ ,  $K$  from (A1) and the constant from (A3).

Furthermore, if  $w_1$  and  $w_2$  are sub- and supersolutions of (3.4.3) satisfying  $w_1(s, z) \leq w_2(s, z)$  for  $(s, z) \in (\{0\} \times \overline{\Omega}) \cup ((0, T + \varepsilon^2] \times \partial\Omega)$ , then  $w_1 \leq w_2$ .

*Sketch of Proof.* The bound on  $|v_i^\varepsilon|_1$  is a consequence of Theorem 3.3.7. The second claim follows from the continuous dependence estimates in Theorem 3.3.8. In particular, for the coefficients setting  $\phi = b, \sigma, c, f$  we have that

$$\left| \phi^\alpha(t, x) - \frac{\phi^\alpha(t + \eta, (1 + \kappa\varepsilon)(x + \xi))}{1 + \kappa\varepsilon} \right| \leq \kappa\varepsilon |\phi^\alpha|_0 + (2\varepsilon + \kappa\varepsilon|x|) [\phi^\alpha]_1.$$

□

Krylov's regularization procedure shows a way to construct smooth subsolutions on  $Q_T$  by mollification of the solution to a system with “shaken coefficients”. For bounded domains, if applied directly, this requires to define such a solution at points lying outside  $\Omega$ . For this we would need to assume (A3) in the enlarged domain  $\Omega_\varepsilon$ , see (A4) below for the definition of  $\Omega_\varepsilon$ . This assumption seems quite restrictive to us. An alternative is to stretch the domain  $\Omega$  by an affine transformation and to define the operator  $F^\varepsilon$  as in (3.4.3). Combining these two gives the desired solutions of the HJB equation in the enlarged domain.

**Remark 3.4.1.** Under certain restrictions in the shape of the domain, it seems possible to generalise the enlargement of the domain to smooth transformations. Here we limit the scope to affine transformations to have explicit and simple computations. Regarding the shape of the domain, the current transformation is suited for star shaped domains, if we assume, without loss of generality, that the center of the star is located at the origin of the coordinate system.

The next assumption states the conditions on  $\Omega$  so that after a transformation, the mollification of functions in the transformed domain have  $\Omega$  as support.

**(A4) (Domain stretching)** Let  $\varepsilon \in (0, \varepsilon_0]$  and define  $\Omega_\varepsilon := \bigcup_{x \in \overline{\Omega}} \mathcal{B}(x, \varepsilon)$ , where  $\mathcal{B}(y, \delta)$  is the open ball of center  $y$  and radius  $\delta$ . Then, there exist  $\kappa > 0$ , such that for  $\tilde{\Omega} := \{x \in \mathbb{R}^d : \frac{x}{1+\kappa\varepsilon} \in \Omega\}$  we have that  $\Omega_\varepsilon \subset \tilde{\Omega}$ .

Following (A4) we define,

$$\tilde{v}^\varepsilon(t, x) := v^\varepsilon\left(t, \frac{x}{1+\kappa\varepsilon}\right), \quad \forall (t, x) \in [0, T + \varepsilon^2] \times \overline{\tilde{\Omega}}, \quad (3.4.5)$$

for  $\varepsilon, \kappa > 0$ . By construction, we have that  $\tilde{v}^\varepsilon(t, x)$  is the unique viscosity solution of

$$\begin{aligned} \tilde{F}_i^\varepsilon(t, x, \tilde{v}^\varepsilon, \partial_t \tilde{v}_i^\varepsilon, D\tilde{v}_i^\varepsilon, D^2\tilde{v}_i^\varepsilon) &= 0 & \text{in } (0, T + \varepsilon^2] \times \tilde{\Omega}, \quad i \in \mathcal{I}, \quad (3.4.6) \\ \tilde{v}_i^\varepsilon(0, x) &= \Psi_0\left(\frac{x}{1+\kappa\varepsilon}\right) & \text{in } \overline{\tilde{\Omega}}, \\ \tilde{v}_i^\varepsilon(t, x) &= \Psi_1\left(t, \frac{x}{1+\kappa\varepsilon}\right) & \text{in } (0, T + \varepsilon^2] \times \partial\tilde{\Omega}, \end{aligned}$$

where

$$\tilde{F}_i^\varepsilon(t, x, r, p_t, p_x, X) := F_i^\varepsilon\left(t, \frac{x}{1+\kappa\varepsilon}, r, p_t, (1+\kappa\varepsilon)p_x, (1+\kappa\varepsilon)^2 X\right), \quad (3.4.7)$$

and  $\tilde{\Omega} := \{x \in \mathbb{R}^d : \frac{x}{1+\kappa\varepsilon} \in \Omega\}$ . Thus, we have the equivalent result to Proposition 3.4.2 applied to  $\tilde{v}^\varepsilon$ .

**Proposition 3.4.3.** *Assume (A1), (A2), (A3) and (A4) for  $v^\varepsilon(t, x)$ . Let  $\tilde{v}^\varepsilon(t, x)$  defined in (3.4.5) be the unique viscosity solution of (3.4.6).*

*Then, for all  $i \in \mathcal{I}$*

$$|\tilde{v}_i^\varepsilon|_1 \leq C_1 \text{ in } [0, T + \varepsilon^2] \times \tilde{\Omega}, \quad \text{and} \quad \frac{1}{\varepsilon} |\tilde{v}_i^\varepsilon - v_i|_0 \leq C_2 \text{ in } \overline{Q}_T,$$

*where  $v$  solves (3.4.1) and the constants  $C_1$  and  $C_2$  only depend on  $T$ ,  $\sup_{x \in \Omega} |x|$ ,  $K$  from (A1) and  $|v^\varepsilon|_1$ .*

*Sketch of Proof.* The proof follows from the same arguments as in Proposition 3.4.2, together with the regularity of the initial and boundary conditions. Indeed,

$$\left| \Psi_0(x) - \Psi_0\left(\frac{x}{1 + \kappa\varepsilon}\right) \right| \leq [\Psi_0]_1 \kappa\varepsilon |x|,$$

and similarly for the boundary conditions. □

Using these results we have all the necessary ingredients to state and prove the analogue of Theorem 2.3 in [8] in the case of bounded spatial domains with Dirichlet boundary conditions. The proof is almost identical to the one in Theorem 2.3 in [8], the only difference is the way to estimate the bound obtained from the comparison principle. However, for completeness we reproduce the whole argument here and indicate where the two differ.

**Theorem 3.4.4.** *Assume (A1), (A2), (A3), (A4) and  $v_0 = (u_0, \dots, u_0)$ . If  $u$  and  $v$  are the solutions of (3.4.2) and (3.4.1) respectively, then for  $k$  small enough,*

$$0 \leq v_i - u \leq Ck^{1/3} \quad \text{in } Q_T, \quad i \in \mathcal{I},$$

*where  $C$  only depends on  $T$ ,  $K$  from (A1).*

*Proof.* For the lower bound consider  $w = (u, \dots, u) \in \mathbb{R}^M$ . Then, it is easy to check

that  $w$  is a subsolution of (3.4.1). Indeed,

$$\begin{aligned} w_i - \mathcal{M}_i w &= \max_{j \neq i} \{w_i - w_j\} - k = -k \leq 0, \\ u_t + \sup_{\alpha \in \mathcal{A}_i} \mathcal{L}^\alpha(t, x, u, Du, D^2 u) &\leq u_t + \sup_{\alpha \in \tilde{\mathcal{A}}} \mathcal{L}^\alpha(t, x, u, Du, D^2 u) \leq 0, \end{aligned}$$

then given that  $w = v_i$  on  $\partial^* Q_T$  for  $i \in \mathcal{I}$ , by comparison for (3.4.1) (Proposition 3.4.1) yields  $u \leq v_i$  for  $i \in \mathcal{I}$ .

For the upper bound we use the regularization procedure of Krylov [51]. Consider the system (3.4.6). Then, by definition for every fixed pair of controls  $(\eta, \xi) = (s, \frac{e}{1+\kappa\varepsilon}) \in [0, \varepsilon^2] \times \overline{\mathcal{B}(0, \varepsilon)}$ ,

$$\partial_t \tilde{v}_i^\varepsilon + \sup_{\alpha \in \mathcal{A}_i} \mathcal{L}^\alpha(t + s, x + e, \tilde{v}_i^\varepsilon(t, x), D\tilde{v}_i^\varepsilon, D^2 \tilde{v}_i^\varepsilon) \leq 0 \quad \text{in} \quad (0, T + \varepsilon^2] \times \tilde{\Omega}, \quad i \in \mathcal{I}.$$

By the same arguments as in the proof of Theorem 2.3 in [8], shifting the variables preserves the subsolution property, in particular,  $\tilde{v}^\varepsilon(t - s, x - e)$  is a subsolution of the following system of independent equations

$$\partial_t w_i + \sup_{\alpha \in \mathcal{A}_i} \mathcal{L}^\alpha(t, x, w_i, Dw_i, D^2 w_i) = 0 \quad \text{in} \quad Q_T^\varepsilon, \quad i \in \mathcal{I}, \quad (3.4.8)$$

where  $Q_T^\varepsilon := (\varepsilon^2, T] \times \Omega$ . Next, following the arguments in [8], we define  $\tilde{v}_\varepsilon := \tilde{v}^\varepsilon * \rho_\varepsilon$  where  $\{\rho_\varepsilon\}_\varepsilon$  is the sequence of mollifiers defined in (3.1.7) and conclude that  $\tilde{v}_\varepsilon$  is also a subsolution of equation (3.4.8). This is a consequence of using a Riemann-sum approximation to the mollification and using Lemma 2.7 in [7].

Recall that  $\tilde{v}^\varepsilon$  is also a continuous subsolution of (3.4.6), hence from the switching part

$$\tilde{v}_i^\varepsilon \leq \min_{j \neq i} \tilde{v}_j^\varepsilon + k \quad \text{in} \quad (0, T + \varepsilon^2] \times \tilde{\Omega}, \quad i \in \mathcal{I}.$$

As the right-hand side does not depend on  $i$ , taking the maximum we conclude that

$$|\tilde{v}_i^\varepsilon - \tilde{v}_j^\varepsilon|_0 \leq k, \quad i, j \in \mathcal{I}.$$

By properties of mollifiers and the previous bound, using integration by parts we obtain the same bounds as in [8]

$$|\partial_t \tilde{v}_{\varepsilon i} - \partial_t \tilde{v}_{\varepsilon j}|_0 \leq C \frac{k}{\varepsilon^2}, \quad |D^n \tilde{v}_{\varepsilon i} - D^n \tilde{v}_{\varepsilon j}|_0 \leq C \frac{k}{\varepsilon^n}, \quad n \in \mathbb{N}, \quad i, j \in \mathcal{I},$$

where  $C$  depends only on  $\rho$  and the uniform bounds on  $v_{\varepsilon i}$  and its gradient. As a result, for  $\varepsilon < 1$ ,

$$\left| \partial_t \tilde{v}_{\varepsilon j} + \sup_{\alpha \in \mathcal{A}_i} \mathcal{L}^\alpha[\tilde{v}_{\varepsilon j}] - \partial_t \tilde{v}_{\varepsilon i} - \sup_{\alpha \in \mathcal{A}_i} \mathcal{L}^\alpha[\tilde{v}_{\varepsilon i}] \right| \leq C \frac{k}{\varepsilon^2} \quad \text{in } Q_T^\varepsilon, \quad i, j \in \mathcal{I},$$

where  $C$  is a constant as above and for compactness we write  $\mathcal{L}^\alpha[u] \equiv \mathcal{L}^\alpha(t, x, u, Du, D^2u)$ .

Using the fact that  $\tilde{v}_\varepsilon$  is a subsolution of (3.4.8), we obtain that

$$\partial_t \tilde{v}_{\varepsilon j} + \sup_{\alpha \in \mathcal{A}_i} \mathcal{L}^\alpha(t, x, \tilde{v}_{\varepsilon j}, D\tilde{v}_{\varepsilon j}, D^2\tilde{v}_{\varepsilon j}) \leq C \frac{k}{\varepsilon^2} \quad \text{in } Q_T^\varepsilon, \quad j \in \mathcal{I},$$

then taking the maximum over all  $i \in \mathcal{I}$  results in

$$\partial_t \tilde{v}_{\varepsilon j} + \sup_{\alpha \in \bar{\mathcal{A}}} \mathcal{L}^\alpha(t, x, \tilde{v}_{\varepsilon j}, D\tilde{v}_{\varepsilon j}, D^2\tilde{v}_{\varepsilon j}) \leq C \frac{k}{\varepsilon^2} \quad \text{in } Q_T^\varepsilon, \quad j \in \mathcal{I}.$$

From Assumption (A1) and the standard arguments using the comparison principle, we see that  $\tilde{v}_{\varepsilon i} - te^{Kt}C\frac{k}{\varepsilon^2}$  is a smooth subsolution of equation (3.4.2) restricted to  $Q_T^\varepsilon$ , as noted in [8].

By the comparison for (3.4.2) in Proposition 3.4.1 we have

$$\tilde{v}_{\varepsilon i} - u \leq e^{Kt} \left( \sup_{\partial^* Q_T^\varepsilon} |\tilde{v}_{\varepsilon i}(t, x) - u(t, x)| + Ct \frac{k}{\varepsilon^2} \right) \quad \text{in } Q_T^\varepsilon, \quad i \in \mathcal{I},$$



where as defined previously  $\partial^* Q_T^\varepsilon$  denotes the parabolic envelope of  $Q_T^\varepsilon$ , i.e.  $\partial^* Q_T^\varepsilon = (\{\varepsilon^2\} \times \overline{\Omega}) \cup ((\varepsilon^2, T] \times \partial\Omega)$ . Decomposing the supremum into the initial condition and boundary terms,

$$\sup_{\partial^* Q_T^\varepsilon} |\tilde{v}_{\varepsilon i}(t, x) - u(t, x)| = \max \left( |\tilde{v}_{\varepsilon i}(\varepsilon^2, \cdot) - u(\varepsilon^2, \cdot)|_0, \sup_{(\varepsilon^2, T] \times \partial\Omega} |\tilde{v}_{\varepsilon i}(t, x) - u(t, x)| \right),$$

we can bound these two terms as follows

$$|\tilde{v}_{\varepsilon i}(\varepsilon^2, \cdot) - u(\varepsilon^2, \cdot)|_0 \leq |\tilde{v}_{\varepsilon i}(\varepsilon^2, \cdot) - \tilde{v}_i^\varepsilon(\varepsilon^2, \cdot)|_0 + |\tilde{v}_i^\varepsilon(\varepsilon^2, \cdot) - u(\varepsilon^2, \cdot)|_0 \leq C\varepsilon,$$

where the bound follows from mollifier properties and Hölder continuity in time of  $v_i^\varepsilon$  and  $u$  together with the fact that the difference of the initial conditions is  $\mathcal{O}(\varepsilon)$ . For the boundary of the spatial domain, under the assumption that the solutions to these systems satisfy the boundary data pointwise, we have

$$\begin{aligned} \sup_{(\varepsilon^2, T] \times \partial\Omega} |\tilde{v}_{\varepsilon i}(t, x) - u(t, x)| &\leq \sup_{(\varepsilon^2, T] \times \partial\Omega} |\tilde{v}_i^\varepsilon(t, x) - \Psi_1(t, x)| + \sup_{(\varepsilon^2, T] \times \partial\Omega} |\tilde{v}_{\varepsilon i} - \tilde{v}_i^\varepsilon| \\ &\leq \sup_{(s, x) \in (\varepsilon^2, T] \times \partial\Omega} [v_i^\varepsilon(s, \cdot)]_1 \kappa \varepsilon |x| + C\varepsilon, \end{aligned}$$

where we have applied the definition of  $\tilde{v}^\varepsilon$  and the Lipschitz regularity of  $v_i^\varepsilon$ .

To extend the comparison to the rest of the time domain, we use the same logic as in [8]. Recall that the time regularity of  $u$  and  $v_i$  (Proposition 3.4.1) implies that

$$|u(t, \cdot) - v_i(t, \cdot)|_0 \leq ([u]_1 + [v_i]_1)\varepsilon \quad \text{in } [0, \varepsilon^2].$$

Then by Propositions 3.4.3, regularity of  $u$  and  $\tilde{v}_i^\varepsilon$ , and properties of mollifiers, we have

$$v_i - u \leq v_i - \tilde{v}_{\varepsilon i} + \tilde{v}_{\varepsilon i} - u \leq C\left(\varepsilon + \frac{k}{\varepsilon^2}\right) \quad \text{in } Q_T^\varepsilon, \quad i \in \mathcal{I}.$$

The result follows by minimizing with respect to  $\varepsilon$ . □

### 3.5 Error bounds for discretizations of the Cauchy-Dirichlet problem on bounded domains

We start by listing the assumptions in [8]. For the HJB equation, in addition to (A1) we have:

**(A5)** The control set  $\mathcal{A}$  is a separable metric space and the coefficients  $\sigma^\alpha, b^\alpha, c^\alpha, f^\alpha$  are continuous in  $\alpha$  for all  $t, x$ .

For the scheme (3.1.6) the following conditions need to be fulfilled.

**(S1) (Monotonicity)** There exists  $\lambda, \mu \geq 0, h_0 > 0$  such that if  $|h| \leq h_0, u \leq v$  are functions in  $C_b(\mathcal{G}_h)$ , and  $\phi(t) = e^{\mu t}(a + bt) + c$  for  $a, b, c \geq 0$ , then

$$S(h, t, x, r + \phi(t), [u + \phi]_{t,x}) \geq S(h, t, x, r, [v]_{t,x}) + b/2 - \lambda c \quad \text{in } \mathcal{G}_h^+.$$

**(S2) (Regularity)** For every  $h$  and  $\phi \in C_b(\mathcal{G}_h)$ , the function  $(t, x) \mapsto S(h, t, x, \phi(t, x), [\phi]_{t,x})$  is bounded and continuous in  $\mathcal{G}_h^+$  and the function  $r \mapsto S(h, t, x, r, [\phi]_{t,x})$  is uniformly continuous for bounded  $r$ , uniformly in  $(t, x) \in \mathcal{G}_h^+$ .

**(S3) (Consistency)** There exists a function  $E(\tilde{K}, h, \varepsilon)$  such that for every  $h = (\Delta t, \Delta x) > 0, (t, x) \in \mathcal{G}_h^+$ , and for any sequence  $\{\phi_\varepsilon\}_{\varepsilon>0}$  of smooth functions satisfying

$$|\partial_t^{\beta_0} D^{\beta'} \phi_\varepsilon(x, t)| \leq \tilde{K} \varepsilon^{1-2\beta_0-|\beta'|} \quad \text{in } \overline{Q}_T, \quad \text{for any } \beta_0 \in \mathbb{N}_0, \beta' = (\beta'_i)_i \in \mathbb{N}_0^d,$$

where  $|\beta'| = \sum_{i=1}^d \beta'_i$ , the following estimate holds:

$$|\partial_t \phi_\varepsilon + F(t, x, \phi_\varepsilon, D\phi_\varepsilon, D^2\phi_\varepsilon) - S(h, t, x, \phi_\varepsilon(t, x), [\phi_\varepsilon]_{t,x})| \leq E(\tilde{K}, h, \varepsilon).$$

Before stating the main result and its proof, we state a comparison result for bounded continuous sub- and supersolutions of the numerical scheme (3.1.6) implied

by assumptions (S1) and (S2). This result is a slight modification of Lemma 3.2 in [8].

**Lemma 3.5.1.** *Assume (S1), (S2), and that  $u, v \in C_b(\mathcal{G}_h)$  satisfy*

$$S(h, t, x, u(t, x), [u]_{t,x}) \leq g_1 \quad \text{in } \mathcal{G}_h^+,$$

$$S(h, t, x, v(t, x), [v]_{t,x}) \geq g_2 \quad \text{in } \mathcal{G}_h^+,$$

where  $g_1, g_2 \in C_b(\mathcal{G}_h)$ . Then

$$u - v \leq e^{\mu t} \sup_{(t,x) \in \partial^* \mathcal{G}_h} |(u(t, x) - v(t, x))^+|_0 + 2te^{\mu t} |(g_1 - g_2)^+|_0,$$

where  $\lambda$  and  $\mu$  are given by (S1).

*Proof.* The proof is identical to that of Lemma 3.2 in [8] once we have accounted for the difference of  $u$  and  $v$  at the parabolic boundary.  $\square$

### 3.5.1 Upper bound by Krylov regularization

In this section we prove an upper bound for the difference between the solution of (3.1.1)–(3.1.3) and the numerical solution for the scheme (3.1.6).

Before stating the result, we introduce the equations involved in the proof. We start by considering the solution  $u^\varepsilon$  to a shaken and perturbed equation

$$u_t^\varepsilon + \sup_{0 \leq \eta \leq \varepsilon^2, |\xi| \leq \varepsilon} F^\varepsilon(t + \eta - \varepsilon^2, x + \xi, u^\varepsilon(t, x), Du^\varepsilon, D^2u^\varepsilon) = 0 \quad \text{in } (0, T + \varepsilon^2] \times \Omega, \quad (3.5.1)$$

$$u^\varepsilon(0, x) = \Psi_0(x) \quad \text{in } \overline{\Omega}, \quad (3.5.2)$$

$$u^\varepsilon(t, x) = \Psi_1^\varepsilon(t, x) \quad \text{in } (0, T + \varepsilon^2] \times \partial\Omega, \quad (3.5.3)$$

where  $\Psi_1^\varepsilon$  is defined as in (3.4.4) and  $F^\varepsilon$  is defined in terms of operator  $F$  in (3.1.5) as follows

$$F^\varepsilon(t, x, r, p, X) = F\left(t, (1 + \kappa\varepsilon)x, r, \frac{p}{1 + \kappa\varepsilon}, \frac{X}{(1 + \kappa\varepsilon)^2}\right), \quad (3.5.4)$$

and the coefficients in the definition of  $F$  are extended appropriately according to McShane's Theorem.

The proof of the bound relies on the regularity of  $u^\varepsilon$ . This can be proved under Assumption (A3), as (3.5.1) is a particular case of (3.4.3).

The next step in Krylov's approach is to construct a smooth subsolution to (3.1.1) by fixing  $(\eta, \xi)$  in (3.5.1) and then mollifying the resulting subsolution. This mollification requires points outside the domain  $\bar{\Omega}$ , as a consequence we introduce  $\tilde{u}^\varepsilon$  such that

$$\tilde{u}^\varepsilon(t, x) := u^\varepsilon\left(t, \frac{x}{1 + \kappa\varepsilon}\right), \quad (t, x) \in [0, T + \varepsilon^2] \times \bar{\tilde{\Omega}}, \quad (3.5.5)$$

for  $\varepsilon, \kappa > 0$  and  $\Omega$  satisfying assumption (A4). By construction, we have that  $\tilde{u}^\varepsilon(t, x)$  is the unique viscosity solution of

$$\tilde{u}_t^\varepsilon + \sup_{0 \leq \eta \leq \varepsilon^2, |\xi| \leq \varepsilon} \tilde{F}^\varepsilon(t + \eta - \varepsilon^2, x + \xi, \tilde{u}^\varepsilon(t, x), D\tilde{u}^\varepsilon, D^2\tilde{u}^\varepsilon) = 0, \quad \text{in } (0, T + \varepsilon^2] \times \tilde{\Omega}, \quad (3.5.6)$$

$$\begin{aligned} \tilde{u}^\varepsilon(0, x) &= \Psi_0\left(\frac{x}{1 + \kappa\varepsilon}\right), \quad \text{in } \bar{\tilde{\Omega}}, \\ \tilde{u}^\varepsilon(t, x) &= \Psi_1^\varepsilon\left(t, \frac{x}{1 + \kappa\varepsilon}\right), \quad \text{in } (0, T + \varepsilon^2] \times \partial\tilde{\Omega}, \end{aligned}$$

where

$$\tilde{F}^\varepsilon(t, x, r, p, X) := F^\varepsilon\left(t, \frac{x}{1 + \kappa\varepsilon}, r, (1 + \kappa\varepsilon)p, (1 + \kappa\varepsilon)^2 X\right), \quad (3.5.7)$$

and  $\tilde{\Omega} := \{x \in \mathbb{R}^d : \frac{x}{1+\kappa\varepsilon} \in \Omega\}$ .

**Theorem 3.5.2.** *Assume (A1), (A2), (A3), (A4), (S1), (S2), (S3) and that (3.1.6) has a unique solution  $u_h \in C_b(\mathcal{G}_h)$ . Let  $u$  denote the solution of (3.1.1) satisfying (3.1.2) and (3.1.3) in the strong sense, and let  $h$  be sufficiently small.*

**(Upper bound)** *There exists a constant  $C$  depending only  $\mu, K$  in (S1), (A1) such that*

$$u - u_h \leq e^{\mu T} \sup_{(t,x) \in \partial^* \mathcal{G}_h} |(u - u_h)^+|_0 + C \min_{\varepsilon > 0} \left( \varepsilon + E(\tilde{K}, h, \varepsilon) \right) \quad \text{in } \mathcal{G}_h, \quad (3.5.8)$$

where  $\tilde{K} = |u|_1$ .

*Proof.* The proof is based on Krylov's regularization procedure together with a domain transformation to ensure that the mollification occurs on the interior of the original spatial domain. We start by considering  $u^\varepsilon$ , the unique viscosity solution to (3.5.1)–(3.5.3). From the assumption (A3) and Theorem 3.3.7, we have that  $u^\varepsilon$  is Lipschitz in space and Hölder continuous with exponent  $\frac{1}{2}$  in time. Next, consider  $\tilde{u}^\varepsilon$  defined in (3.5.5), it is straightforward to verify that

$$\begin{aligned} |\tilde{u}^\varepsilon(t, x) - \tilde{u}^\varepsilon(t, y)| &\leq \frac{[u^\varepsilon(t, \cdot)]_1}{1 + \kappa\varepsilon} |x - y|, \quad \forall (x, y) \in \tilde{\Omega} \times \tilde{\Omega}, \\ |u^\varepsilon(t, x) - \tilde{u}^\varepsilon(t, x)| &\leq [u^\varepsilon(t, \cdot)]_1 \frac{\kappa\varepsilon}{1 + \kappa\varepsilon} |x|, \quad \forall x \in \tilde{\Omega}. \end{aligned}$$

Let  $\bar{u}^\varepsilon(t, x) := \tilde{u}^\varepsilon(t + \varepsilon^2, x)$ , then for every fixed  $\eta = s$  and  $\xi = e$ , with  $0 \leq s \leq \varepsilon^2$  and  $|e| \leq \varepsilon$ ,  $\bar{u}^\varepsilon(t - s, x - e)$  is a subsolution of

$$w_t + F(t, x, w(t, x), Dw, D^2w) = 0 \quad \text{in } (0, T) \times \Omega.$$

Now let

$$\tilde{u}_\varepsilon := \bar{u}^\varepsilon * \rho_\varepsilon = \int_{0 \leq s \leq \varepsilon^2} \int_{|e| \leq \varepsilon} \tilde{u}^\varepsilon(t + \varepsilon^2 - s, x - e) \rho_\varepsilon(s, e) \, de \, ds,$$

where  $\{\rho_\varepsilon\}_\varepsilon$  is the sequence of mollifiers defined in (3.1.7). Realizing that  $\tilde{u}_\varepsilon$  is a convex combination of viscosity subsolutions and by stability results of viscosity solutions, see Lemma 2.7 in [7], we conclude that  $\tilde{u}_\varepsilon$  is a classical subsolution of (3.1.1).

To obtain the upper bound we will make use of this classical subsolution. First, we seek to bound the difference between  $u$  and  $\tilde{u}_\varepsilon$  using  $\tilde{u}^\varepsilon$ . Indeed, for all  $(t, x) \in \overline{Q}_T$

$$|u(t, x) - \tilde{u}_\varepsilon(t, x)| \leq |u(t, x) - \tilde{u}^\varepsilon(t, x)| + |\tilde{u}^\varepsilon(t, x) - \tilde{u}_\varepsilon(t, x)|,$$

where we can bound the second term using Lipschitz regularity and properties of the mollifier. To estimate the first term we employ Theorem 3.3.8 and the fact that both functions are viscosity solutions of their corresponding equations, therefore for any  $(t, x) \in \overline{Q}_T$

$$e^{-\lambda t} |u(t, x) - \tilde{u}^\varepsilon(t, x)| \leq \sup_{\partial^* Q_T} |u - \tilde{u}^\varepsilon| + C_1 \varepsilon,$$

where  $C_1 > 0$  depends on  $T$  and the  $|\cdot|_1$  norm of the coefficients, but not on  $\varepsilon$ . To bound the difference on  $\partial^* Q_T$  we can proceed as follows

$$\begin{aligned} \sup_{x \in \overline{\Omega}} \left| \Psi_0(x) - \Psi_0 \left( \frac{x}{1 + \kappa \varepsilon} \right) \right| &\leq [\Psi_0]_1 \frac{\kappa \varepsilon}{1 + \kappa \varepsilon} \sup_{x \in \overline{\Omega}} |x|, \\ \sup_{(0, T] \times \partial \Omega} |u - \tilde{u}^\varepsilon| &= \sup_{(t, x) \in (0, T] \times \partial \Omega} \left| \Psi_1(t, x) - u^\varepsilon \left( t, \frac{x}{1 + \kappa \varepsilon} \right) \right| \\ &\leq \sup_{(t, x) \in (0, T] \times \partial \Omega} [u^\varepsilon(t, \cdot)]_1 \frac{\kappa \varepsilon}{1 + \kappa \varepsilon} |x|, \end{aligned}$$

which are finite by boundedness of  $\overline{\Omega}$ .

Next, by the consistency property (S3) and the fact that  $\tilde{u}_\varepsilon$  is a smooth subsolu-

tion, we have that

$$S(h, t, x, \tilde{u}_\varepsilon(t, x), [\tilde{u}_\varepsilon]_{t,x}) \leq E(\tilde{K}, h, \varepsilon) \quad \text{in } \mathcal{G}_h^+,$$

where  $\tilde{K} \geq 0$  is the finite bound of  $\tilde{u}_\varepsilon$ . Finally, we compare  $u_h$  and  $\tilde{u}_\varepsilon$  using the scheme's comparison principle formulated in Lemma 3.5.1, and use it to establish the upper bound as

$$\begin{aligned} u - u_h &\leq e^{\mu T} \sup_{(t,x) \in \partial^* \mathcal{G}_h} |(\tilde{u}_\varepsilon - u_h)^+|_0 + C \min_{\varepsilon > 0} \left( \varepsilon + E(\tilde{K}, h, \varepsilon) \right) \\ &\leq e^{\mu T} \sup_{(t,x) \in \partial^* \mathcal{G}_h} |(u - u_h)^+|_0 + C \min_{\varepsilon > 0} \left( \varepsilon + E(\tilde{K}, h, \varepsilon) \right). \end{aligned}$$

□

### 3.5.2 Lower bound by switching system approximation

We continue by stating the theorem for the lower bound. For the proof we follow the approach in [8] and use of a switching system approximation to build “almost smooth” supersolutions to (3.1.1). There are two main steps in the proof. Firstly, we consider the case of a finite control set  $\mathcal{A}$ . Secondly, the result is extended to the general case using Assumption (A5). It is in the first step that the proof needs to be adapted for the bounded domain case with Dirichlet conditions. The second part is identical to the original proof in [8].

The next set of lemmas contain some key results regarding the solutions of the auxiliary switching system below and its relation to the solution of (3.1.1)–(3.1.3). The purpose of this auxiliary system is to ensure that the “almost smooth superso-

lution" is defined for the whole  $Q_T$ .

$$F_i^\varepsilon(t, x, v^\varepsilon, \partial_t v_i^\varepsilon, Dv_i^\varepsilon, D^2 v_i^\varepsilon) = 0 \quad \text{in } (0, T + 2\varepsilon^2] \times \Omega, \quad i \in \mathcal{I} = \{1, \dots, M\}, \quad (3.5.9)$$

$$\begin{aligned} v^\varepsilon(0, x) &= \Psi_0(x) \quad \text{in } \bar{\Omega}, \\ v^\varepsilon(t, x) &= \Psi_1^\varepsilon(t, x) \quad \text{in } (0, T + 2\varepsilon^2] \times \partial\Omega, \end{aligned}$$

where  $\Psi_1^\varepsilon$  is defined as in (3.4.4) but for  $(0, T + 2\varepsilon^2] \times \partial\Omega$ ,

$$\begin{aligned} F_i^\varepsilon(t, x, r, p_t, p_x, X) = \\ \max \left\{ p_t + \min_{0 \leq s \leq \varepsilon^2, |e| \leq \varepsilon} \mathcal{L}^{\alpha_i} \left( t + \eta, (1 + \kappa\varepsilon)(x + \xi), r_i, \frac{p_x}{1 + \kappa\varepsilon}, \frac{X}{(1 + \kappa\varepsilon)^2} \right); r_i - \mathcal{M}_i r \right\}, \end{aligned} \quad (3.5.10)$$

for any  $\alpha_i \in \mathcal{A}$ ,  $\mathcal{L}_i^\alpha$  is defined in (3.1.4) and  $\mathcal{M}_i r$  in (3.3.6).

**Lemma 3.5.3** (Lemma 3.3 in [8]). *Assume (A1), (A2), and (A3), then the solution  $v^\varepsilon$  of (3.5.9) satisfies*

$$|v^\varepsilon|_1 \leq \bar{K}, \quad |v_i^\varepsilon - v_j^\varepsilon|_0 \leq k, \quad \text{and, for small } k, \quad \max_{i \in \mathcal{I}} |u - v_i^\varepsilon|_0 \leq C(\varepsilon + k^{1/3}),$$

where  $u$  solves (3.1.1)–(3.1.3),  $i, j \in \mathcal{I}$ , and  $\bar{K}, C$  only depend on  $T$  and  $K$  from (A1).

*Proof.* The regularity of  $v^\varepsilon$  follows from Theorem 3.3.7. The bound on the elements of  $v^\varepsilon$  follows by the subsolution property, see proof of Theorem 3.4.4. Finally, the convergence rate to  $u$  from Theorem 3.4.4 and the continuous dependence result for viscosity solutions in Theorem 3.3.8.  $\square$

As in the previous section we expand the domain  $\Omega$  using the transformation



considered in (A4). We define  $\tilde{v}^\varepsilon$  as

$$\tilde{v}_i^\varepsilon(t, x) := v_i^\varepsilon\left(t, \frac{x}{1 + \kappa\varepsilon}\right), \quad (i, t, x) \in \mathcal{I} \times [0, T + \varepsilon^2] \times \bar{\tilde{\Omega}}, \quad (3.5.11)$$

for  $\varepsilon, \kappa > 0$ . By construction, we have that  $\tilde{v}^\varepsilon(t, x)$  is the unique viscosity solution of

$$\tilde{F}_i^\varepsilon(t, x, \tilde{v}^\varepsilon, \partial_t \tilde{v}_i^\varepsilon, D\tilde{v}_i^\varepsilon, D^2\tilde{v}_i^\varepsilon) = 0 \quad \text{in} \quad (0, T + 2\varepsilon^2] \times \tilde{\Omega}, \quad i \in \mathcal{I} = \{1, \dots, M\}, \quad (3.5.12)$$

$$\begin{aligned} \tilde{v}^\varepsilon(0, x) &= \Psi_0\left(\frac{x}{1 + \kappa\varepsilon}\right) && \text{in} \quad \bar{\tilde{\Omega}}, \\ \tilde{v}^\varepsilon(t, x) &= \Psi_1^\varepsilon\left(t, \frac{x}{1 + \kappa\varepsilon}\right) && \text{in} \quad (0, T + 2\varepsilon^2] \times \partial\tilde{\Omega}, \end{aligned} \quad (3.5.13)$$

where

$$\tilde{F}_i^\varepsilon(t, x, r, p_t, p_x, X) := F_i^\varepsilon\left(t, \frac{x}{1 + \kappa\varepsilon}, r, p_t, (1 + \kappa\varepsilon)p_x, (1 + \kappa\varepsilon)^2 X\right), \quad (3.5.14)$$

and  $\tilde{\Omega} := \{x \in \mathbb{R}^d : \frac{x}{1 + \kappa\varepsilon} \in \Omega\}$ . The equivalent of Lemma 3.5.3 also holds for  $\tilde{v}^\varepsilon$ .

**Lemma 3.5.4.** *Assume (A1), (A2), and that the solution of (3.5.9) satisfies (A3), then  $\tilde{v}_i^\varepsilon$  given by (3.5.11) satisfies*

$$|\tilde{v}^\varepsilon|_1 \leq \bar{K}, \quad |\tilde{v}_i^\varepsilon - \tilde{v}_j^\varepsilon|_0 \leq k, \quad \text{and, for small } k, \quad \max_{i \in \mathcal{I}} |u - \tilde{v}_i^\varepsilon|_0 \leq C(\varepsilon + k^{1/3}),$$

where  $u$  solves (3.1.1)–(3.1.3),  $i, j \in \mathcal{I}$ , and the constants  $\bar{K}$  and  $C$  only depend on  $T$  and  $K$  from (A1).

The next two lemmas are necessary for the proof and concern certain properties of the mollification of the solution  $\tilde{v}^\varepsilon$ . We reproduce them from [8] for readability. The lemmas being concerned with points in the interior of the domain, their proofs also hold in the case of bounded domains with Dirichlet boundary conditions.

**Lemma 3.5.5** (Lemma 3.4 in [8]). *Assume (A1) and  $\varepsilon \leq (8 \sup_i [\tilde{v}_i^\varepsilon]_1)^{-1}k$  where  $\tilde{v}^\varepsilon$  is defined in (3.5.11). Let*

$$\tilde{v}_{\varepsilon i}(t, x) := \tilde{v}_i^\varepsilon(t + \varepsilon^2, x) * \rho_\varepsilon(t, x) \quad \text{for } i \in \mathcal{I},$$

*then for every  $(t, x) \in Q_T$ , if  $j := \operatorname{argmin}_{i \in \mathcal{I}} \tilde{v}_{\varepsilon i}(t, x)$ , we have that*

$$\partial_t \tilde{v}_{\varepsilon j}(t, x) + \mathcal{L}^{\alpha_j}(t, x, \tilde{v}_{\varepsilon j}(t, x), D\tilde{v}_{\varepsilon j}(t, x), D^2\tilde{v}_{\varepsilon j}(t, x)) \geq 0.$$

**Lemma 3.5.6** (Lemma 3.5 in [8]). *Assume (A1) and  $\varepsilon \leq (8 \sup_i [\tilde{v}_i^\varepsilon]_1)^{-1}k$  where  $\tilde{v}^\varepsilon$  is defined in (3.5.11). Let*

$$\tilde{v}_{\varepsilon i}(t, x) := \tilde{v}_i^\varepsilon(t + \varepsilon^2, x) * \rho_\varepsilon(t, x) \quad \text{for } i \in \mathcal{I}.$$

*Then the function  $w := \min_{i \in \mathcal{I}} \tilde{v}_{\varepsilon i}$  is an approximate supersolution of the scheme (3.1.6) in the sense that*

$$S(h, t, x, w(t, x), [w]_{t,x}) \geq -E(\bar{K}, h, \varepsilon) \quad \text{in } \mathcal{G}_h^+,$$

*where  $\bar{K}$  comes from Lemma 3.5.4.*

**Theorem 3.5.7** (Theorem 3.1b in [8]). *Assume (A1), (A2), (A3), (A4), (A5), (S1), (S2), (S3) and that (3.1.6) has a unique solution  $u_h$  in  $C_b(\mathcal{G}_h)$ . Let  $u$  denote the solution of (3.1.1) satisfying (3.1.2) and (3.1.3) in the strong sense, and let  $h$  be sufficiently small.*

**(Lower bound)** *There exists a constant  $C$  depending only on  $\mu, K$  in (S1), (A1), such that*

$$u - u_h \geq -e^{\mu T} \sup_{(t,x) \in \partial^* \mathcal{G}_h} |(u - u_h)^-|_0 - C \min_{\varepsilon > 0} \left( \varepsilon^{1/3} + E(\tilde{K}, h, \varepsilon) \right) \quad \text{in } \mathcal{G}_h, \quad (3.5.15)$$

where  $\tilde{K} = |u|_1$ .

*Proof.* We proceed as in [8] and we consider the case of finite control set  $\mathcal{A} = \{\alpha_1, \dots, \alpha_M\}$  and approximate the original problem (3.1.1)–(3.1.3) by the solution of the switching system approximation (3.5.12). By Lemma 3.5.4 we know that the solution of this switching system is expected to be close to that of (3.1.1) when  $k$  and  $\varepsilon$  are small.

We intend to construct approximate smooth supersolutions of (3.1.1) and then use an analogue argument to that in Theorem 3.5.2 to derive the lower bound. For this purpose, we mollify the following function

$$\bar{v}_i^\varepsilon(t, x) := \tilde{v}_i^\varepsilon(t + \varepsilon^2, x),$$

where the variable change is done to ensure that we have initial data at  $t = 0$ . This function  $\bar{v}_i^\varepsilon$  is defined on  $Q_T^\varepsilon := (-\varepsilon^2, T + \varepsilon^2] \times \tilde{\Omega}$  and by Hölder continuity of the solution in time with exponent  $1/2$ ,

$$|\bar{v}_i^\varepsilon - \tilde{v}_i^\varepsilon|_0 \leq \bar{K}\varepsilon \quad \text{for } i \in \mathcal{I}. \quad (3.5.16)$$

Lemma 3.5.5 applied to  $\tilde{v}_i^\varepsilon$  gives the basis to construct a smooth supersolution to (3.1.1). Indeed, defining

$$\tilde{v}_{\varepsilon i}(t, x) := \tilde{v}_i^\varepsilon(t + \varepsilon^2, x) * \rho_\varepsilon(t, x) \quad \text{for } i \in \mathcal{I},$$

when  $\varepsilon$  is small compared to the switching cost  $k$ , we have that  $w$  given by

$$w := \min_{i \in \mathcal{I}} \tilde{v}_{\varepsilon i} \quad \text{for } i \in \mathcal{I},$$

is a supersolution of (3.1.1) in  $Q_T$ .

As a consequence, let  $k = 8 \sup_i [\tilde{v}_i^\varepsilon]_1 \varepsilon$  and use Lemma 3.5.6 together with Lemma 3.5.1 to compare  $u_h$  and  $w$ , obtaining

$$u_h - w \leq e^{\mu t} \sup_{(t,x) \in \partial^* \mathcal{G}_h} |(u_h(t, x) - w(t, x))^+|_0 + 2te^{\mu t} E(\bar{K}, h, \varepsilon) \quad \text{in } \mathcal{G}_h.$$

However, by previous results on the convergence to the solution of switching systems to HJB equations we have

$$|w - u|_0 \leq C(\varepsilon + k + k^{1/3}),$$

and therefore

$$u_h - u \leq e^{\mu t} \sup_{(t,x) \in \partial^* \mathcal{G}_h} |(u_h(t, x) - u(t, x))^+|_0 + 2te^{\mu t} E(\bar{K}, h, \varepsilon) + C(\varepsilon + k + k^{1/3}) \quad \text{in } \mathcal{G}_h,$$

for some constant  $C$ . We conclude the first step of the proof by minimizing w.r.t  $\varepsilon$ .

The extension of the proof to the general case  $\mathcal{A}$ , is identical to the one in [8]. By (A5)  $\mathcal{A}$  is a separable metric space and therefore has a countable dense subset  $\mathcal{A}_\infty$ . Moreover, given the continuity of the coefficients in  $\alpha$

$$\sup_{\mathcal{A}} \mathcal{L}^\alpha(t, x, r, p, X) = \sup_{\mathcal{A}_\infty} \mathcal{L}^\alpha(t, x, r, p, X).$$

Define a strictly increasing sequence of subsets  $\{\mathcal{A}_M\}_{M=1}^\infty \subset \mathcal{A}_\infty$  such that

$$\mathcal{A}_M \subset \mathcal{A}_{M+1} \text{ for } M \in \mathbb{N} \quad \text{and} \quad \cup_{M=1}^\infty \mathcal{A}_M = \mathcal{A}_\infty,$$

and let  $u^M$  be the solution of (3.1.1) when  $\mathcal{A}$  is replaced by  $\mathcal{A}_M$ , and similarly for  $u_h^M$ .

The rest of the proof consists of the following steps:

1. Use Arzela-Ascoli's theorem to state that there is a subsequence of  $\{u_M\}_M$  that

converges locally uniformly to  $u$ .

2. Note that the supersolution candidate  $w$  is also a supersolution of (3.1.1) for general  $\mathcal{A}$  and that the proof of Lemma 3.5.6 also holds when replacing  $u_h^M$  by  $u_h$ .
3. Finalise the proof by realising that the bound we obtained does not depend on  $M$  and hence we can send  $M \rightarrow \infty$  on the left-hand side and the results still holds.

□

## 3.6 Error bounds for some monotone finite difference schemes

In this section we employ Theorems 3.5.2 and 3.5.7 to derive error bounds for some monotone finite difference schemes approximating (3.1.1)–(3.1.3).

### 3.6.1 The scheme by Kushner and Dupuis

This scheme is a conditionally monotone finite difference method using a node's neighbours to approximate the second order operator (3.1.4), see [53]. The locality of the scheme prevents it from overstepping the domain, but it only yields monotone discretizations if and only if  $a^\alpha$  is diagonally dominant.

The error bounds for this scheme were also analysed in Section 4 of [8] for  $\Omega = \mathbb{R}^d$  and no boundary conditions. As the stencil of the scheme only relies on the immediate neighbours, the scheme does not overstep and can be used to approximate problems on bounded domains without modification. For completeness, we reproduce the scheme from [53] and report the error bounds.

For any function  $\phi$  we denote by  $\phi^+ = \max(\phi, 0)$  to its positive part,  $\phi^- = \max(-\phi, 0)$  to its negative part. We denote by  $\{e_i\}_{i=1}^d$  the canonical basis in  $\mathbb{R}^d$ , and define

$$\begin{aligned}
\Delta_t^+ \phi(t, x) &= \frac{1}{\Delta t} (\phi(t + \Delta t, x) - \phi(t, x)), \\
\Delta_i^\pm \phi(t, x) &= \pm \frac{1}{\Delta x} (\phi(t, x \pm e_i \Delta x) - \phi(t, x)), \\
\Delta_{ii} \phi(t, x) &= \frac{1}{\Delta x^2} (\phi(t, x + e_i \Delta x) - 2\phi(t, x) + \phi(t, x - e_i \Delta x)), \\
\Delta_{ij}^+ \phi(t, x) &= \frac{1}{2\Delta x^2} (2\phi(t, x) + \phi(t, x + e_i \Delta x + e_j \Delta x) + \phi(t, x - e_i \Delta x - e_j \Delta x)) \\
&\quad - \frac{1}{2\Delta x^2} (\phi(t, x + e_i \Delta x) + \phi(t, x - e_i \Delta x) + \phi(t, x + e_j \Delta x) + \phi(t, x - e_j \Delta x)), \\
\Delta_{ij}^- \phi(t, x) &= -\frac{1}{2\Delta x^2} (2\phi(t, x) + \phi(t, x + e_i \Delta x - e_j \Delta x) + \phi(t, x - e_i \Delta x + e_j \Delta x)) \\
&\quad + \frac{1}{2\Delta x^2} (\phi(t, x + e_i \Delta x) + \phi(t, x - e_i \Delta x) + \phi(t, x + e_j \Delta x) + \phi(t, x - e_j \Delta x)),
\end{aligned}$$

where  $\phi$  is a smooth function.

We combine the finite difference operators above to obtain the discretization of the operator (3.1.4) proposed by Kushner and Dupuis

$$\begin{aligned}
L_h^\alpha \phi(t, x) &= \sum_{i=1}^d \left( a_{ii}^\alpha(t, x) \Delta_{ii} \phi(t, x) + \sum_{i \neq j} [a_{ij}^{\alpha,+}(t, x) \Delta_{ij}^+ \phi(t, x) - a_{ij}^{\alpha,-}(t, x) \Delta_{ij}^- \phi(t, x)] \right) \\
&\quad + \sum_{i=1}^d [b_i^{\alpha,+}(t, x) \Delta_i^+ \phi(t, x) - b_i^{\alpha,-}(t, x) \Delta_i^- \phi(t, x)]. \tag{3.6.1}
\end{aligned}$$

By construction of the scheme, the numerical solution is exact at the parabolic boundary. Therefore, the error bounds for the Kushner-Dupuis scheme applied to the Cauchy-Dirichlet problem are identical to the ones obtained in Theorem 4.1 in [8].

**Theorem 3.6.1.** *If  $u_h \in C_b(\mathcal{G}_h)$  is the solution of the Kushner and Dupuis scheme*

and  $u$  is the solution of (3.1.1)–(3.1.3), then there is a  $C > 0$  such that in  $\mathcal{G}_h$

$$-e^{\mu t} \sup_{(t,x) \in \partial^* \mathcal{G}_h} |(u - u_h)^-|_0 - Ch_\ell \leq u_h - u \leq e^{\mu t} \sup_{(t,x) \in \partial^* \mathcal{G}_h} |(u - u_h)^+|_0 + Ch_u,$$

where  $h_\ell$  and  $h_u$  are defined as follows

$$h_\ell = \max(\Delta t^{1/10}, \Delta x^{1/5}), \text{ and } h_u = \max(\Delta t^{1/4}, \Delta x^{1/2}).$$

### 3.6.2 The truncated semi-Lagrangian scheme

We consider the error bounds for the truncated semi-Lagrangian scheme. The scheme being monotone, it is of positive type as in Definition 2.2.1, therefore it can be written according to (2.2.2) as follows

$$S^\alpha(h, t_n, x_j, r, [U]_{t_n, x_j}) = \max_{\alpha \in \mathcal{A}} \{S^\alpha(h, t_n, x_j, r, [U]_{t_n, x_j})\}, \quad (3.6.2)$$

where

$$S^\alpha(h, t_n, x_j, r, [U]_{t_n, x_j}) := \mathcal{W}_{j,j}^{\alpha,n,n} r - \sum_{i \neq j} \mathcal{W}_{j,i}^{\alpha,n,n} U_i^n - \sum_{i=1}^N \mathcal{W}_{j,i}^{\alpha,n,n-1} U_i^{n-1} - F_j^{\alpha,n-1+\theta}$$

and the coefficients  $\mathcal{W}$  are given by

$$\begin{aligned} \mathcal{W}_{j,j}^{\alpha,n,n} &= \frac{1}{\Delta t_n} + \theta \left( \sum_{p=1}^M \frac{A_p^{\alpha,n} + B_p^{\alpha,n}}{2\Delta x} - \hat{l}_{j,j}^{\alpha,n} - c_j^{\alpha,n-1+\theta} \right), \\ \mathcal{W}_{j,j}^{\alpha,n,n-1} &= \frac{1}{\Delta t_n} - (1 - \theta) \left( \sum_{p=1}^M \frac{A_p^{\alpha,n-1} + B_p^{\alpha,n-1}}{2\Delta x} - \hat{l}_{j,j}^{\alpha,n-1} - c_j^{\alpha,n-1+\theta} \right), \\ \mathcal{W}_{j,i}^{\alpha,n,n} &= \theta \hat{l}_{j,i}^{\alpha,n}, \quad \mathcal{W}_{j,i}^{\alpha,n,n-1} = (1 - \theta) \hat{l}_{j,i}^{\alpha,n-1}, \end{aligned}$$

with  $(n, j) \in \{1, \dots, N_t\} \times \{1, \dots, N_x\}$  with  $N_t$  and  $N_x$  being the number of the temporal and spatial nodes in  $\mathcal{G}_h^+$  respectively.  $U_i^n$  is the numerical solution at node

$(t_n, x_i)$ ,  $[U]_{t_n, x_j}$  are the values of  $U_i^m$  for  $m \neq n$  and  $i \neq j$ ,  $c_j^{\alpha, n-1+\theta} = c^\alpha(t_{n-1} + \theta\Delta t, x_j)$  and  $A_p^{\alpha, n}$ ,  $B_p^{\alpha, n}$  are continuous functions depending on  $\sigma_p^\alpha$ ,  $b^\alpha$  and  $\partial\Omega$ .

Recall that the scheme is of positive type according to Definition 2.2.1 if the conditions in Proposition 2.2.4 hold.

**Proposition 3.6.2.** *Under the conditions (2.2.16), the scheme (3.6.2) satisfies the assumptions (S1), (S2) and (S3) with*

$$E(C, h, \varepsilon) := C(|1 - 2\theta|\Delta t\varepsilon^{-3} + \Delta t^2\varepsilon^{-5} + \sqrt{\Delta x}\varepsilon^{-2} + \Delta x\varepsilon^{-1}), \quad (3.6.3)$$

where  $C > 0$  is a constant. Moreover, the solution of the scheme is unique.

*Proof.* To check (S1) we recall that the positivity of the coefficients  $\mathcal{W}$  we see that the scheme (3.6.2) is monotone, in the sense that it is increasing in  $r$  and decreasing in  $U$ , and that by properties of linear monotone interpolation

$$\sum_{i=1}^{N_x} \hat{l}_{j,i}^{\alpha, n} = \sum_{p=1}^M \frac{A_p^{\alpha, n-1} + B_p^{\alpha, n-1}}{2\Delta x}.$$

Let  $\phi(t) = \tilde{\phi}(t) + d$  with  $\tilde{\phi}(t) = e^{\mu t}(a + bt)$  for  $a, b, d \geq 0$ , then

$$\begin{aligned} & S(h, t_n, x_j, r, [u + \phi]_{t_n, x_j}) \\ &= \max_{\alpha \in \mathcal{A}} \left\{ S^\alpha(h, t_n, x_j, r, [u]_{t_n, x_j}) + \left( \mathcal{W}_{j,j}^{\alpha, n, n} - \sum_{i \neq j} \mathcal{W}_{j,i}^{\alpha, n, n} \right) \phi(t_n) - \sum_{i=1}^N \mathcal{W}_{j,i}^{\alpha, n, n} \phi(t_{n-1}) \right\} \\ &= \max_{\alpha \in \mathcal{A}} \left\{ S^\alpha(h, t_n, x_j, r, [u]_{t_n, x_j}) - c_j^{\alpha, n-1+\theta} d + \frac{\tilde{\phi}(t_n) - \tilde{\phi}(t_{n-1})}{\Delta t_n} \right. \\ & \quad \left. - c_j^{\alpha, n-1+\theta} [\tilde{\phi}(t_{n-1}) + \theta(\tilde{\phi}(t_n) - \tilde{\phi}(t_{n-1}))] \right\}, \end{aligned}$$

then it is straightforward to see that we should set  $\lambda = \sup_\alpha |c^{\alpha+}|_0$ , as in Theorem



3.3.7. With regards to  $\mu$ , by the mean value theorem  $\exists \xi \in [t_{n-1}, t_n]$  such that

$$\frac{\tilde{\phi}(t_n) - \tilde{\phi}(t_{n-1})}{\Delta t_n} - c_j^{\alpha, n-1+\theta} [\tilde{\phi}(t_{n-1}) + \theta(\tilde{\phi}(t_n) - \tilde{\phi}(t_{n-1}))] = \tilde{\phi}'(\xi) - c_j^{\alpha, n-1+\theta} [\tilde{\phi}(t_{n-1}) + \theta \Delta t_n \tilde{\phi}'(\xi)]$$

choosing  $\mu = \lambda + 1$  and using the fact that  $\tilde{\phi}'(t) = \mu \tilde{\phi}(t) + be^{\mu t}$  we obtain the following inequality

$$\tilde{\phi}'(\xi) - c_j^{\alpha, n-1+\theta} [\tilde{\phi}(t_{n-1}) + \theta \Delta t_n \tilde{\phi}'(\xi)] \geq \tilde{\phi}(\xi) + be^{\mu \xi} - c_j^{\alpha, n-1+\theta} \theta \Delta t_n \tilde{\phi}'(\xi) \geq \frac{b}{2},$$

thus setting  $|h_0|^{-1} = 2\lambda \tilde{\phi}'(T)$  ensures that the scheme satisfies (S1).

(S2) follows from the continuity and the  $L^\infty$ -stability of the scheme.

Regarding the consistency of the scheme required in (S3), the scheme is consistent by construction and for any smooth function  $\phi$  we have

$$\frac{|1 - 2\theta|}{2} |\phi_{tt}|_0 \Delta t + \kappa \left( \Delta t^2 |\phi_{tt}|_0 + \Delta x |D^2 \phi|_0 + \sqrt{\Delta x} |D^3 \phi|_0 \right)$$

with  $\kappa > 0$ , see Corollary 2.2.3 for further details on this calculation. The final expression is obtained assuming that  $\phi$  is constructed by mollification and that  $\partial_t^{k_1} D_x^{k_2} \phi = \mathcal{O}(\varepsilon^{1-2k_1-k_2})$  for any  $k_1, k_2 \in \mathbb{N}_0$ .

The existence and uniqueness of the solution can be proved in two ways. First, by adapting the arguments in Theorem 4.2 in [26] to (3.6.2). This theorem expresses the scheme as a fixed point iteration and shows that it is a contraction in the Banach space of bounded grid functions. However, a constructive proof can be given using the arguments in [14]. Schemes of positive type yield discretization matrices in the class of M-matrices with non-negative row sums. We can therefore use policy iteration to solve the resulting non-linear system of equations. By Theorem 2.1 in [14] there exists a uniqueness solution to the system and the approximate solution computed using policy iteration converges to it.  $\square$

**Corollary 3.6.3.** *Let  $u_h$  be the solution of the truncated LISL discretization of (3.1.1)–(3.1.3) and  $u$  be the solution of (3.1.1)–(3.1.3), then there is  $C > 0$  such that in  $\mathcal{G}_h$*

$$-e^{\mu t} \sup_{(t,x) \in \partial^* \mathcal{G}_h} |(u - u_h)^-|_0 - Ch_\ell \leq u_h - u \leq e^{\mu t} \sup_{(t,x) \in \partial^* \mathcal{G}_h} |(u - u_h)^+|_0 + Ch_u,$$

where  $h_\ell$  and  $h_u$  are defined as follows

$$h_\ell = \max(\Delta t^{1/10}, \Delta x^{1/14}), \text{ and } h_u = \max(\Delta t^{1/4}, \Delta x^{1/6}).$$

*Proof.* The result follows by combining the bounds from Theorem 3.5.2 and Theorem 3.5.7 together with the consistency error given by (3.6.3). Hence, the result follows by minimizing w.r.t.  $\varepsilon$  the following functions:

1. For the upper bound by (3.5.8) and (3.6.3)

$$\varepsilon + C(\Delta t \varepsilon^{-3} + \Delta t^2 \varepsilon^{-5} + \Delta x \varepsilon^{-1} + \sqrt{\Delta x} \varepsilon^{-2}),$$

2. For the lower bound by (3.5.15) and (3.6.3)

$$\varepsilon^{1/3} + C(\Delta t \varepsilon^{-3} + \Delta t^2 \varepsilon^{-5} + \Delta x \varepsilon^{-1} + \sqrt{\Delta x} \varepsilon^{-2}).$$

By minimizing separately in  $\Delta t$  and  $\Delta x$ , one finds that  $\varepsilon$  has to be of order  $\Delta t^{1/4}$  and  $\Delta x^{1/6}$  in the first case, and that  $\varepsilon^{1/3}$  has to be of order  $\Delta t^{1/10}$  and  $\Delta x^{1/14}$  in the second case. The result follows by taking  $\varepsilon = \max(\Delta t^{1/4}, \Delta x^{1/6})$  in the first case and  $\varepsilon^{1/3} = \max(\Delta t^{1/10}, \Delta x^{1/14})$  in the second case.  $\square$

### 3.6.3 The truncated semi-Lagrangian scheme with local refinement

We consider a refinement of the mesh on boundary layers of width  $\mathcal{O}(\Delta x)$  where the semi-Lagrangian scheme oversteps. The objective is to improve the consistency error of the truncated stencil from  $\mathcal{O}(\sqrt{\Delta x})$  to  $\mathcal{O}(\Delta x)$ . For this purpose, we combine a local refinement of the mesh with appropriate changes to the stencil step in the region near the boundary. The stencil step is an extra parameter in the schemes studied in [26] denoted by  $k$ . The expression of the scheme in terms of  $k$  is as follows

$$\begin{aligned} \text{tr}[a^\alpha(t, x)D^2\phi] + b^\alpha(t, x)D\phi &\approx \frac{\mathcal{I}_{\Delta x}\phi(x + k^2b^\alpha(t, x)) - \mathcal{I}_{\Delta x}\phi(x)}{k^2} \\ &+ \sum_{j=1}^P \frac{\mathcal{I}_{\Delta x}\phi(x + k\sigma_j^\alpha(t, x)) - 2\mathcal{I}_{\Delta x}\phi(x) + \mathcal{I}_{\Delta x}\phi(x - k\sigma_j^\alpha(t, x))}{2k^2}, \end{aligned}$$

where  $\phi \in C^\infty(\mathbb{R}^d)$ ,  $(t, x) \in Q_T$ ,  $\mathcal{I}_{\Delta x}$  is the multi-linear interpolation operator and  $P$  is the number of columns of the matrix  $\sigma^\alpha$ . As shown in Corollary 5.1 in [26], it is optimal, with respect to the local consistency error, to choose  $k = \mathcal{O}(\sqrt{\Delta x})$  in the absence of overstepping.

For simplicity, consider first a one dimensional spatial domain, i.e.  $\Omega \subset \mathbb{R}$  and construct a grid  $\Omega_{\Delta x} \subset \Omega$  with refinement parameter  $\Delta x > 0$  and  $N_x$  nodes, and assume that  $\sigma^\alpha(t, x) = \sigma(x)$ . For any  $y \in \Omega$ , we denote by  $\mathcal{N}(y) \subset \Omega_{\Delta x}$  the set of grid nodes used in the interpolation to compute  $\mathcal{I}_{\Delta x}\phi(y)$  for some function  $\phi : \Omega \rightarrow \mathbb{R}$ .

Since it is the overstepping of the diffusion stencil that reduces the local truncation error, we split the nodes in  $\Omega_{\Delta x}$  according to the overstepping into three subsets as in the next definition:

**Definition 3.6.1.** We define

- $\Omega_{\Delta x}^{(1)} := \{x \in \Omega_{\Delta x} \mid x + \sqrt{\Delta x}\sigma(x) \notin \bar{\Omega} \text{ or } x - \sqrt{\Delta x}\sigma(x) \notin \bar{\Omega}\}$ , that is, the set of nodes whose stencil oversteps  $\bar{\Omega}$ .

- $\Omega_{\Delta x}^{(2)} := \left( \bigcup_{j=1}^{|\Omega_{\Delta x}^{(1)}|} (\mathcal{N}(x_j + \sqrt{\Delta x} \sigma(x_j)) \cup \mathcal{N}(x_j - \sqrt{\Delta x} \sigma(x_j))) \right) \cap (\Omega_{\Delta x} \setminus \Omega_{\Delta x}^{(1)})$ . This is the set of nodes not belonging to  $\Omega_{\Delta x}^{(1)}$  but that belong to the stencil of a node from  $\Omega_{\Delta x}^{(1)}$ .
- $\Omega_{\Delta x}^{(3)} := \Omega_{\Delta x} \setminus (\Omega_{\Delta x}^{(1)} \cup \Omega_{\Delta x}^{(2)})$ , the set of nodes whose stencil does not overstep the domain and that are not part of the stencil of nodes in  $\Omega_{\Delta x}^{(1)}$ .

Therefore,  $\Omega_{\Delta x} = \Omega_{\Delta x}^{(1)} \cup \Omega_{\Delta x}^{(2)} \cup \Omega_{\Delta x}^{(3)}$  with  $\Omega_{\Delta x}^{(i)} \cap \Omega_{\Delta x}^{(j)} = \emptyset$  for  $i \neq j$  and  $i, j \in \{1, 2, 3\}$ .

Given that the width of the stencil is of length  $\mathcal{O}(\sqrt{\Delta x})$ , the cardinality of  $\Omega_{\Delta x}^{(1)} \cup \Omega_{\Delta x}^{(2)}$  is  $|\Omega_{\Delta x}^{(1)} \cup \Omega_{\Delta x}^{(2)}| \sim \mathcal{O}(\sqrt{N_x})$ . Next, we refine the mesh elements in  $\Omega_{\Delta x}^{(1)}$  and  $\Omega_{\Delta x}^{(2)}$  with mesh-refinement parameter proportional to  $\Delta x^\gamma$  where  $\gamma > 1$  is a parameter to be determined so that the resulting local consistency error is at least  $\mathcal{O}(\Delta x)$ . We also allow for each regions to have a different stencil step  $k_i$  where  $i \in \{1, 2, 3\}$ .

We proceed in reverse order. As the local consistency error for nodes in  $\Omega_{\Delta x}^{(3)}$  was already first order accurate we do not modify the stencil step, i.e.  $k_3 \sim \mathcal{O}(\sqrt{\Delta x})$ . From Corollary 5.2 in [26], the local consistency error for nodes in the refined  $\Omega_{\Delta x}^{(2)}$  is

$$\mathcal{O} \left( k_2^2 + \frac{\Delta x^2}{k_2^2} + \frac{\Delta x^{2\gamma}}{k_2^2} \right),$$

where the first term corresponds to the consistency error of the finite difference approximation of the second order derivative and the last two terms to the interpolation error in the original and the refined regions respectively. Choosing  $k_2 \sim \mathcal{O}(\sqrt{\Delta x})$  the local consistency error is  $\mathcal{O}(\Delta x)$ .

Similarly, assuming that the truncated stencil uses the exact value at the boundary the local consistency error for nodes in refined  $\Omega_{\Delta x}^{(1)}$  is

$$\mathcal{O} \left( k_1 + \frac{\Delta x^{2\gamma}}{k_1^2} \right),$$

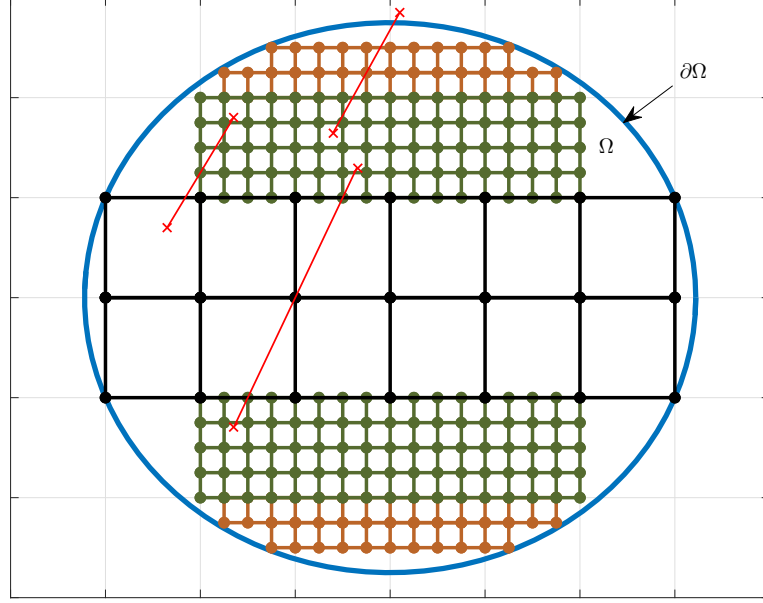


Figure 3.6.1: Locally refined grid according to the diffusion stencil shown in red. The regions in Definition 3.6.1 are shown in different colors. Brown for  $\Omega_{\Delta x}^{(1)}$ , green for  $\Omega_{\Delta x}^{(2)}$  and black for  $\Omega_{\Delta x}^{(3)}$ . In red we plot the stencil for one node per region.

if one side oversteps, and

$$\mathcal{O}(k_1),$$

if both sides overstep. Thus, choosing  $\gamma = \frac{3}{2}$  and  $k_1 \sim \mathcal{O}(\Delta x)$  the local truncation error is  $\mathcal{O}(\Delta x)$ . Moreover, after the refinement the number of nodes in this region of width  $\mathcal{O}(\sqrt{\Delta x})$  is  $\mathcal{O}(N_x)$ .

As shown in Corollary 2.2.5, the other undesirable effect of the truncation of stencil was the worsening of the CFL condition of the scheme. The local refinement of the grid results in a stricter CFL condition compared to the one in Corollary 2.2.5, as shown in Proposition 3.6.4.

We can extend this approach to the multidimensional case, i.e.  $\Omega \subset \mathbb{R}^d$  with  $d > 1$ . Figure 3.6.1 shows a locally refined mesh.

**Remark 3.6.1.** Figure 3.6.1 shows that the local refinement leaves “hanging nodes”

at the interface between  $\Omega_{\Delta x}^{(2)}$  and  $\Omega_{\Delta x}^{(3)}$ . These nodes do not pose a problem for semi-Lagrangian discretizations. For the interpolation, the “hanging nodes” are not used for stencil points with any neighbours belonging to  $\Omega_{\Delta x}^{(3)}$ .

The next proposition summarises these findings.

**Proposition 3.6.4.** *Let  $\Delta t, \Delta x > 0$  be the time and space mesh refinement parameters for  $\mathcal{T}_{\Delta t} \subset [0, T]$  and  $\Omega_{\Delta x} \subset \Omega \subset \mathbb{R}^d$ , with  $N_t := |\mathcal{T}_{\Delta t}|$  and  $N_x^d := |\Omega_{\Delta x}|$ . Then, if for every  $(t_n, p) \in \mathcal{T}_{\Delta t} \times ([1, P] \cap \mathbb{N})$  we split the nodes in  $\Omega_{\Delta x}$  according to Definition 3.6.1 further refine the region contained in  $\Omega_{\Delta x}^{(1)} \cup \Omega_{\Delta x}^{(2)}$  with mesh refinement parameter  $\mathcal{O}(\Delta x^{3/2})$ , the complexity of the method is  $\mathcal{O}(N_t N_x^d)$ . If for the nodes requiring truncation in the refined grid we use  $k \sim \mathcal{O}(\Delta x)$ , then, globally the consistency error of this modified scheme becomes*

$$\frac{|1 - 2\theta|}{2} |\phi_{tt}|_0 \Delta t + C \left( \Delta t^2 |\phi_{tt}|_0 + \Delta x (|D^2 \phi|_0 + |D^3 \phi|_0 + |D^4 \phi|_0) \right). \quad (3.6.4)$$

*Additionally, provided that all the nodes are, at least,  $\mathcal{O}(\Delta x^{3/2})$  away from the boundary of the domain, a scheme with  $\theta < 1$  requires  $\Delta t \sim \mathcal{O}(\Delta x^{5/2})$  if only one side of the stencil oversteps or  $\Delta t \sim \mathcal{O}(\Delta x^3)$  if both sides of the stencil overstep.*

**Remark 3.6.2.** We note that each  $t_n$  in the time grid and column  $\sigma_p^\alpha$  of the diffusion matrix, could yield a different split according to Definition 3.6.1. Therefore, even if the number of spatial nodes after refinement stays proportional to the one for the original problem, computing the split may affect the computational time and complicates the implementation of the method.

**Remark 3.6.3.** Ensuring that all nodes in the mesh are at least  $\mathcal{O}(\Delta x^{3/2})$  away from the boundary of the domain, can be done by removing the outermost layer of the grid inside the domain after refinement, as discussed in Section 2.2.1.

**Corollary 3.6.5.** *Let  $u_h$  be the solution of the truncated LISL discretization in Proposition 3.6.4 for (3.1.1)–(3.1.3) and  $u$  be the solution of (3.1.1)–(3.1.3), then there is  $C > 0$  such that in  $\mathcal{G}_h$*

$$-e^{\mu t} \sup_{(t,x) \in \partial^* \mathcal{G}_h} |(u - u_h)^-|_0 - Ch_\ell \leq u_h - u \leq e^{\mu t} \sup_{(t,x) \in \partial^* \mathcal{G}_h} |(u - u_h)^+|_0 + Ch_u,$$

where  $h_\ell$  and  $h_u$  are defined as follows

$$h_\ell = \max(\Delta t^{1/10}, \Delta x^{1/10}), \text{ and } h_u = \max(\Delta t^{1/4}, \Delta x^{1/4}).$$

*Proof.* To obtain the result combine the bounds from Theorem 3.5.2 and Theorem 3.5.7 together with the consistency error given by (3.6.4). Assume that  $\partial_t^{k_1} D_x^{k_2} \phi = \mathcal{O}(\varepsilon^{1-2k_1-k_2})$  for any  $k_1, k_2 \in \mathbb{N}_0$ , then

1. For the upper bound by (3.5.8) and (3.6.4)

$$\varepsilon + C(\Delta t \varepsilon^{-3} + \Delta t^2 \varepsilon^{-5} + \Delta x(\varepsilon^{-1} + \varepsilon^{-2} + \varepsilon^{-3})),$$

2. For the lower bound by (3.5.15) and (3.6.4)

$$\varepsilon^{1/3} + C(\Delta t \varepsilon^{-3} + \Delta t^2 \varepsilon^{-5} + \Delta x(\varepsilon^{-1} + \varepsilon^{-2} + \varepsilon^{-3})).$$

By minimizing separately in  $\Delta t$  and  $\Delta x$ , one finds that  $\varepsilon$  has to be of order  $\Delta t^{1/4}$  and  $\Delta x^{1/4}$  in the first case, and that  $\varepsilon^{1/3}$  has to be of order  $\Delta t^{1/10}$  and  $\Delta x^{1/10}$  in the second case. The result follows by taking  $\varepsilon = \max(\Delta t^{1/4}, \Delta x^{1/4})$  in the first case and  $\varepsilon^{1/3} = \max(\Delta t^{1/10}, \Delta x^{1/10})$  in the second case.  $\square$

**Remark 3.6.4.** The rates obtained in Corollary 3.6.5 are the same as for the semi-Lagrangian scheme in [26] where  $\Omega = \mathbb{R}^d$  and no boundary conditions are imposed, see Theorem 6.1 in [27].

## 3.7 Conclusion

This chapter has extended the error analysis in [8] to the Cauchy-Dirichlet problem for HJB equations on bounded domains. Using the framework developed, we have analysed the classical Kushner and Dupuis scheme and the truncated semi-Lagrangian scheme proposed in Section 2.2. Finally, we argued that using local mesh refinements we can theoretically obtain global consistency error proportional to  $\mathcal{O}(\Delta x)$  for the truncated stencil scheme. This improves the estimate in Section 2.2 by half an order. However, the refinement increases the number of time steps required for stability of explicit time stepping schemes. This is another argument in favour of using implicit time stepping schemes and studying efficient solvers for algebraic systems of equations for semi-Lagrangian discretization matrices as done in Chapter 4.



# Chapter 4

## Multigrid preconditioning

### 4.1 Introduction

The truncation of the semi-Lagrangian stencil introduced in Section 2.2 modifies the CFL condition of explicit schemes by at least half an order, from  $\Delta t = \mathcal{O}(\Delta x)$  to  $\Delta t = \mathcal{O}(\Delta x^{3/2})$ . However, the CFL condition for implicit schemes remains unaffected by the truncation. As the error observed empirically is  $\mathcal{O}(\Delta t) + \mathcal{O}(\Delta x)$  for fully implicit schemes, the computationally most efficient choice is  $\Delta t \sim \Delta x$ . This choice of  $\Delta t$  is outside the stability region of explicit schemes, which makes implicit schemes attractive. Thus, in this chapter we study the application of multigrid preconditioners together with policy iteration [14] to solve the non-linear system (1.1.11).

There are two main families of multigrid methods in the literature: geometric multigrid and algebraic multigrid. So far, only the geometric approach has been used in the literature to solve non-linear systems arising from discretizations of HJB equations, see [3, 12, 41, 43]. Moreover, all of these works use classical finite difference schemes to approximate the HJB equation. We therefore start by studying the use of geometric multigrid for the LISL scheme.

Geometric multigrid requires us to predefine a grid hierarchy based on the geome-

try of the problem. Defining a grid hierarchy requires good knowledge of the spectral properties of the problem and the scheme. This is difficult for approximations using non-local stencils like the semi-Lagrangian schemes. Indeed, the variability of the width of the LISL stencil within a given grid (variable coefficients) and through the grid hierarchy makes it difficult, even for simple problems, to design an appropriate grid hierarchy and a good smoother. Furthermore, even for simple linear problems, changes in the value of the coefficients affect the performance of standard geometric multigrid cycles significantly, as seen in Section 4.4.

Another aspect to consider is related to the transfer operators. Standard grid interpolations provide approximations using the grid neighbours of a given node, whereas for the LISL stencil, being non-local, the solution at a given node may not be best approximated by its neighbours on the grid but by those on its stencil. These heuristics suggest that the algebraic approach to multigrid, that is, fixing the smoother and building operator dependent intergrid transfer operators, may result in more efficient multigrid preconditioning for LISL discretizations.

Algebraic multigrid (AMG), introduced in [69], constructs “coarse grids” based solely on the matrix coefficients. However, as pointed out in Section 6.2 of the recent review on preconditioning [82], AMG may not achieve an effective reduction in the number of variables from one grid to the next. This is due to a slow reduction in the number of unknowns and the use of the Galerkin principle to build the coarse system matrix, together with intergrid transfer operators using weighted averages. As a result, this method results in preconditioners with high complexity, as defined below.

To measure the complexity the following quantities are commonly used:

**Definition 4.1.1.** The grid complexity  $c_G$  is the total number of variables  $N$ , on all

multigrid levels, divided by the number of variables on the finest level  $N_1$ ,

$$c_G = \frac{1}{N_1} \sum_{\ell=1}^{n_{\text{levels}}} N_\ell.$$

**Definition 4.1.2.** The algebraic complexity  $c_A$  is the total number of non-zero entries, in all matrices  $A_\ell$ , divided by the number of non-zero entries of the finest level operator  $A_1$ ,

$$c_A = \frac{1}{\text{nnz}(A_1)} \sum_{\ell=1}^{n_{\text{levels}}} \text{nnz}(A_\ell).$$

We will find it beneficial for speed of convergence to construct the “coarse grids” algebraically already for simple examples of LISL matrices (Section 4.4, in particular Table 4.4.1), and that algebraic construction of the grid hierarchy deals well with the varying LISL stencils (Section 4.6). However, we see a significant increase in complexity of AMG (see Table 4.4.1) with the size of the mesh. Empirically we observe that this is mainly due to the use of interpolation in the LISL scheme, which renders the method impractical for large problems.

Recent and on-going research on algebraic multigrid [62, 63] shows how one can construct good multigrid cycles using simplified intergrid transfer operators based on aggregation of the unknown variables, using piecewise constant interpolation, thus avoiding the problem of increased complexity in coarser levels. In particular, [63] proves convergence of a simplified two-grid scheme using aggregation for non-singular M-matrices with non-negative row and column sums. We will show that these results apply for LISL discretizations matrices and justify the use of AGMG both theoretically and empirically.

The rest of the chapter is organised as follows. Section 4.2 considers the spectrum of LISL discretization matrices and shows that the eigenvectors associated to small eigenvalues may not be smooth as for local finite difference discretizations. Section 4.3

uses Local Fourier Analysis (LFA) to analyse the smoothing properties of common smoothers applied to LISL discretization matrices. Section 4.4 compares the performance of geometric multigrid against algebraic multigrid methods and concludes that geometric multigrid is not suitable for non-local schemes. It also shows that AMG yields dense grid hierarchies. Section 4.5 checks the conditions for a LISL discretization matrix to have non-negative row and column sum. Section 4.6 concludes by benchmarking the performance of both of algebraic multigrid methods against MATLAB's sparse direct solver based on UMFPACK [25].

## 4.2 On the spectrum of LISL matrices

To assess the suitability of preconditioning based on geometric multigrid, we start by considering the spectrum of LISL matrices for a simplified model. For illustration, we first calculate the eigenvalues and eigenvectors of the LISL discretization of the diffusion operator with constant coefficients, for any function  $u : \mathbb{R}^d \rightarrow \mathbb{R}$

$$-\frac{1}{2}\nabla^T\sigma\sigma^T\nabla u = -\frac{1}{2}\sum_{i=1}^d\sigma_i^2\frac{\partial^2 u}{\partial x_i^2}, \quad (4.2.1)$$

where  $\sigma \in \mathbb{R}^{d \times d}$  is a diagonal matrix with  $(\sigma)_{ii} = \sigma_i$ .

We start by considering the one-dimensional case on an equispaced grid  $\Omega_{\Delta x}$ , where  $\Delta x > 0$  is the distance between two consecutive nodes. For  $\sigma > 0$ , we define

$$m := \left\lfloor \frac{\sigma}{\sqrt{\Delta x}} \right\rfloor, \quad \text{and} \quad \gamma := (m+1) - \frac{\sigma}{\sqrt{\Delta x}}, \quad (4.2.2)$$

where  $m \in \mathbb{N}$  denotes the stencil length and  $\gamma \in [0, 1]$  is the interpolation weight of the one-dimensional linear interpolation operator, such that for any real function  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  the linear interpolation operator on  $\Omega_{\Delta x}$  is  $\mathcal{I}_{\Delta x}(\phi)(x_i + \sqrt{\Delta x}\sigma) = \gamma\phi(x_i + m\Delta x) + (1 - \gamma)\phi(x_i + (m+1)\Delta x)$ . Without loss of generality and for simplicity of

the notation we assume that  $\Delta x = 1$ . Denote by  $L_N$  the following  $N \times N$  Laplacian matrix

$$L_N := \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 2 \end{pmatrix}.$$

Let now  $m = 2$  and  $\Delta x = 1$ , then the LISL discretization matrix is given by

$$L_{SL}^{N,m,\gamma} := \begin{pmatrix} 2 & 0 & -\gamma & -1+\gamma & 0 & \cdots & 0 \\ 0 & 2 & 0 & -\gamma & -1+\gamma & \cdots & 0 \\ -\gamma & 0 & 2 & 0 & -\gamma & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 2 \end{pmatrix}. \quad (4.2.3)$$

Noticing the structure in the diagonals, we re-write  $L_{SL}^{N,m,\gamma}$  as

$$L_{SL}^{N,m,\gamma} = \gamma L_N^m + (1 - \gamma) L_N^{m+1},$$

where  $L_N^m = L_{SL}^{N,m,1}$ .

Using the properties of Kronecker products we can characterize the eigenvalues of the matrices  $L_N^m$  in terms of the eigenvalues of the standard  $L_N$  matrices. Denoting by  $\lambda(L_N) \in \mathbb{R}^N$  and  $V(L_N) \in \mathbb{R}^{N \times N}$  the eigenvalues and eigenvectors of  $L_N$ , respectively,

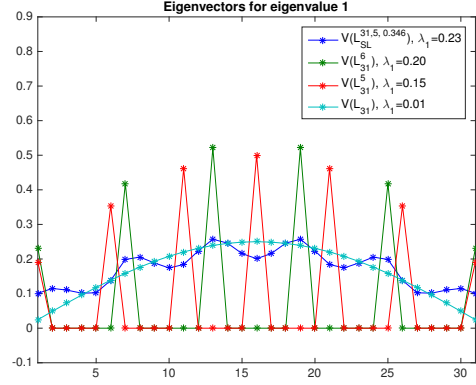
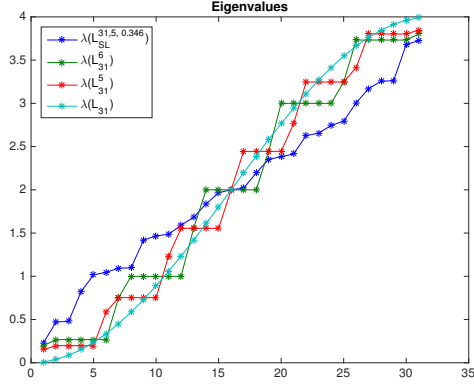
we have that

$$\begin{aligned}\lambda(L_N^m) &= \left[ \lambda \left( L_{\lceil \frac{N}{m} \rceil} \right) \otimes e_1 \right]_N + \left[ \lambda \left( L_{\lfloor \frac{N}{m} \rfloor} \right) \otimes \sum_{i=2}^N e_i \right]_N, \\ V(L_N^m) &= \left[ V \left( L_{\lceil \frac{N}{m} \rceil} \right) \otimes \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{0}_{m-1} \end{pmatrix} \right]_{N \times N} + \left[ V \left( L_{\lfloor \frac{N}{m} \rfloor} \right) \otimes \begin{pmatrix} 0 & 0 \\ 0 & I_{m-1} \end{pmatrix} \right]_{N \times N},\end{aligned}$$

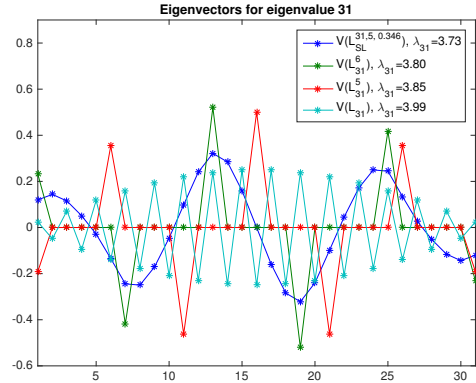
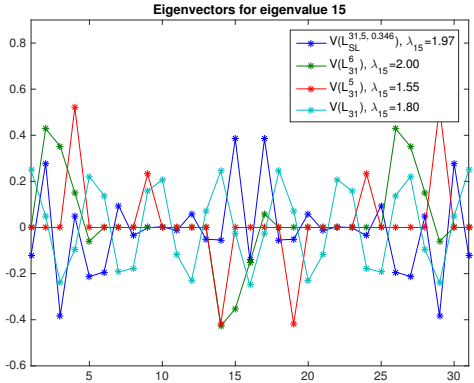
where  $e_i$  is the  $i$ -th canonical basis vector of  $\mathbb{R}^N$ ,  $I_N$  is the  $N \times N$  identity matrix and  $\mathbf{0}_m$  denotes the  $m \times m$  zero matrix. By  $[A]_{N \times N}$  we mean that we select the first  $N$  rows and  $N$  columns of  $A$ , and similar for  $[v]_N$  for a vector  $N$ . This is required as  $N$  will in general not be a multiple of both  $m$  and  $m+1$  so the resulting matrices from the Kronecker product will be of size  $\lceil \frac{N}{m} \rceil m$  and  $\lfloor \frac{N}{m+1} \rfloor (m+1)$  which are greater or equal to  $N$ .

In the presence of interpolation, that is, when  $\gamma \in (0, 1)$ , we are unable to provide any closed formula for the eigenvalues and eigenvectors of  $L_{SL}^{N,m,\gamma} = \gamma L_N^m + (1-\gamma)L_N^{m+1}$ . Figure 4.2.1 contains graphs with the eigenvalues and some eigenvectors of the matrices  $L_{SL}^{N,m,\gamma}$ ,  $L_N^m$ ,  $L_N^{m+1}$  and  $L_N$ . The plots show that for LISL discretization matrices, in contrast to the standard case, small eigenvalues are not necessarily associated to smooth modes. As a result, these components cannot be represented accurately on the coarse mesh.

The spectrum of higher-dimensional constant coefficient Laplacians can be inferred from the spectrum of the one-dimensional matrices by means of Kronecker products. Next, we consider the properties of common smoothers when applied to LISL discretization matrices of the two-dimensional Laplacian and conclude with an example illustrating the impact of the diffusion coefficient on the convergence of geometric multigrid cycles.



(a) Comparison of the eigenvalues of  $L_{SL}^{31,5,0.346}$ ,  $L_{31}^5$ ,  $L_{31}^6$  and  $L_{31}$  in increasing order. (b) Eigenvectors corresponding to the smallest eigenvalue for  $L_{SL}^{31,5,0.346}$ ,  $L_{31}^5$ ,  $L_{31}^6$  and  $L_{31}$ .



(c) Eigenvectors corresponding to the 15-th eigenvalue for  $L_{SL}^{31,5,0.346}$ ,  $L_{31}^5$ ,  $L_{31}^6$  and  $L_{31}$ . (d) Eigenvectors corresponding to the largest eigenvalue for  $L_{SL}^{31,5,0.346}$ ,  $L_{31}^5$ ,  $L_{31}^6$  and  $L_{31}$ .

Figure 4.2.1: Eigenvalues of  $L_{SL}^{N,m,\gamma}$ ,  $L_N^m$ ,  $L_N^{m+1}$  and  $L_N$  with parameter values  $N = 31$ ,  $m = 5$  and  $\gamma = 0.346$  and the eigenvectors corresponding to three eigenvalues of the same matrices.

### 4.3 Local Fourier analysis of the smoothers

We seek to analyse how a varying size stencil affects the properties of the standard Gauss-Seidel smoother. We base the analysis on Local Fourier Analysis (LFA) as described in Chapter 4 of [77] and state the *smoothing factors*  $\mu_{\text{loc}}$  of Gauss-Seidel iterations when applied to wide stencil finite difference discretizations. The key to the analysis is the use of grid functions of the form  $\varphi(\boldsymbol{\theta}, \mathbf{x}) = e^{i\boldsymbol{\theta} \cdot \mathbf{x}}$ , where  $i$  is the imaginary unit,  $\mathbf{x} \in \mathbb{R}^d$ ,  $\boldsymbol{\theta} \in [-\pi, \pi)^d$  and  $\cdot$  is the inner product for vectors in  $\mathbb{R}^d$ . For simplicity we consider equispaced grids  $\Omega_{\Delta x}$  with refinement parameter  $\Delta x > 0$ . Therefore, any  $\mathbf{x} \in \Omega_{\Delta x}$  can be written as  $\mathbf{x} \equiv \mathbf{x}_0 + \kappa \Delta x$  for some fixed  $\mathbf{x}_0 \in \Omega_{\Delta x}$  and  $\kappa \in \mathbb{Z}^d$ . It is thus convenient to rescale the exponent of  $\varphi$  by  $\Delta x^{-1}$ .

The functions  $\varphi$  are important since, as shown in Lemma 4.2.1 of [77], “all grid functions  $\varphi(\boldsymbol{\theta}, \mathbf{x})$  are (formal) eigenfunctions of any discrete operator which can be described by a difference stencil”. This property allows us to associate to each discrete finite difference operator  $L_{\Delta x}$  a so-called symbol  $\tilde{L}_{\Delta x}(\boldsymbol{\theta})$  defined by

$$L_{\Delta x} \varphi(\boldsymbol{\theta}, \mathbf{x}) = \sum_{\kappa \in \mathbb{Z}^d} s_{\kappa} e^{i\boldsymbol{\theta} \cdot \kappa} = \tilde{L}_{\Delta x}(\boldsymbol{\theta}) e^{i\boldsymbol{\theta} \cdot \kappa}, \quad (4.3.1)$$

where  $s_{\kappa} \in \mathbb{R}$  is the finite difference coefficient at the location  $\kappa$  with respect to the node  $\mathbf{x}_0$ .

As in [77], we consider smoothers,  $S_{\Delta x}$ , that can be defined from splittings of the discrete operator with the following general form  $L_{\Delta x} = L_{\Delta x}^+ + L_{\Delta x}^-$ , where the choices of  $L_{\Delta x}^+$  and  $L_{\Delta x}^-$  result in different smoothers. By construction we have that

$$S_{\Delta x} = (L_{\Delta x}^+)^{-1} L_{\Delta x}^-.$$

Examples 4.3.1, 4.3.2 and 4.3.3 show that the smoother based Gauss-Seidel iterations can be analysed within this framework. For this smoother  $L_{\Delta x}^+$  contains the elements



in the main and lower diagonals, and  $L_{\Delta x}^-$  contains the elements in the upper diagonal.

Lemma 4.3.1 in [77] derives the expression for the symbol for the smoother as

$$\tilde{S}_{\Delta x}(\boldsymbol{\theta}) := \frac{\tilde{L}_{\Delta x}^-(\boldsymbol{\theta})}{\tilde{L}_{\Delta x}^+(\boldsymbol{\theta})},$$

where  $\tilde{L}_{\Delta x}^+$  and  $\tilde{L}_{\Delta x}^-$  are defined as for  $L_{\Delta x}$  in (4.3.1).

With multigrid, the objective of the smoother is to dampen error components not reduced by the coarse grid correction. Therefore, assessing the properties of a given smoother requires fixing the coarse grid correction. We limit the study to the simplest coarsening strategy, that is if  $\Omega_{\Delta x}$  is the fine grid then  $\Omega_{2\Delta x}$  is the coarse grid. This leads to the definition of low and high frequencies below.

**Definition 4.3.1** (Definition 4.2.1 in [77]). For the coarsening considered, we define the high and low frequencies as follows:

$$\begin{aligned} \varphi(\boldsymbol{\theta}, \cdot) \text{ low frequency component} &\iff \boldsymbol{\theta} \in T^{\text{low}} := \left[-\frac{\pi}{2}, \frac{\pi}{2}\right)^d; \\ \varphi(\boldsymbol{\theta}, \cdot) \text{ high frequency component} &\iff \boldsymbol{\theta} \in T^{\text{high}} := [-\pi, \pi)^d \setminus \left[-\frac{\pi}{2}, \frac{\pi}{2}\right)^d. \end{aligned}$$

**Definition 4.3.2** (Definition 4.3.1 in [77]). The smoothing factor for standard coarsening is

$$\mu_{\text{loc}} = \mu_{\text{loc}}(S_{\Delta x}) := \sup \left\{ |\tilde{S}_{\Delta x}(\boldsymbol{\theta})| : \boldsymbol{\theta} \in T^{\text{high}} \right\}.$$

We employ these definitions to compare the smoothing factors for the standard two-dimensional Laplacian, setting  $d = 2$ , discretised using standard local finite differences and the LISL discretization.

**Example 4.3.1** (Example 4.3.4 in [77]). The smoothing factor for the Gauss-Seidel smoother for the standard Laplacian discretisation is given by

$$\mu_{\text{loc}} = \sup \left\{ \left| \frac{e^{i\theta_1} + e^{i\theta_2}}{4 - e^{-i\theta_1} - e^{-i\theta_2}} \right| : \boldsymbol{\theta} \in T^{\text{high}} \right\}.$$

Similarly, the smoothing factor for the LISL scheme can be derived. In the present case of pure diffusion, Schemes 1–3 coincide.

**Example 4.3.2.** Proceeding as in [77] for Example 4.3.1, the symbols  $L_{\Delta x}^+$  and  $L_{\Delta x}^-$  for the LISL discretizations are

$$\begin{aligned}\tilde{L}_{\Delta x}^+(\boldsymbol{\theta}) &= \frac{1}{\Delta x} (4 - \gamma_1 e^{-im_1\theta_1} - (1 - \gamma_1) e^{-i(m_1+1)\theta_1} - \gamma_2 e^{-im_2\theta_2} - (1 - \gamma_2) e^{-i(m_2+1)\theta_2}), \\ \tilde{L}_{\Delta x}^-(\boldsymbol{\theta}) &= -\frac{1}{\Delta x} (\gamma_1 e^{im_1\theta_1} + (1 - \gamma_1) e^{i(m_1+1)\theta_1} + \gamma_2 e^{im_2\theta_2} + (1 - \gamma_2) e^{i(m_2+1)\theta_2}),\end{aligned}$$

where  $m_i$  and  $\gamma_i$  are given by (4.2.2) replacing  $\sigma$  by  $\sigma_i$ . For compactness of notation, define

$$g(\theta, \gamma, m) := \gamma e^{im\theta} + (1 - \gamma) e^{i(m+1)\theta},$$

then the smoothing factor for a Gauss-Seidel smoother with standard coarsening and the LISL scheme is given by

$$\mu_{\text{loc}}(S_{\Delta x}^{SL}) = \sup \left\{ \left| \frac{g_1 + g_2}{4 - \bar{g}_1 - \bar{g}_2} \right| : \boldsymbol{\theta} \in T^{\text{high}} \right\}, \quad (4.3.2)$$

for  $\boldsymbol{\theta} \in T^{\text{high}}$ , where  $g_1 \equiv g(\theta_1, \gamma_1, m_1)$ ,  $g_2 \equiv g(\theta_2, \gamma_2, m_2)$ , and  $\bar{c}$  denotes the complex conjugate of the complex number  $c$ .

From (4.3.2) we see that as the non-locality of the discretization grows, i.e.  $m_i \rightarrow \infty$  then the smoothing factor approaches 1 (no smoothing) and so highly oscillatory modes will be transferred to the coarser subspace. Figure 4.3.1 compares the smoothing factor for the fixed stencil 5 point discretization and a specific semi-Lagrangian stencil.

**Example 4.3.3.** We can generalise the results in the previous example to the case of diffusion given by a vector  $(\sigma_1, \sigma_2)^T$  not necessarily parallel to any of the axes. If

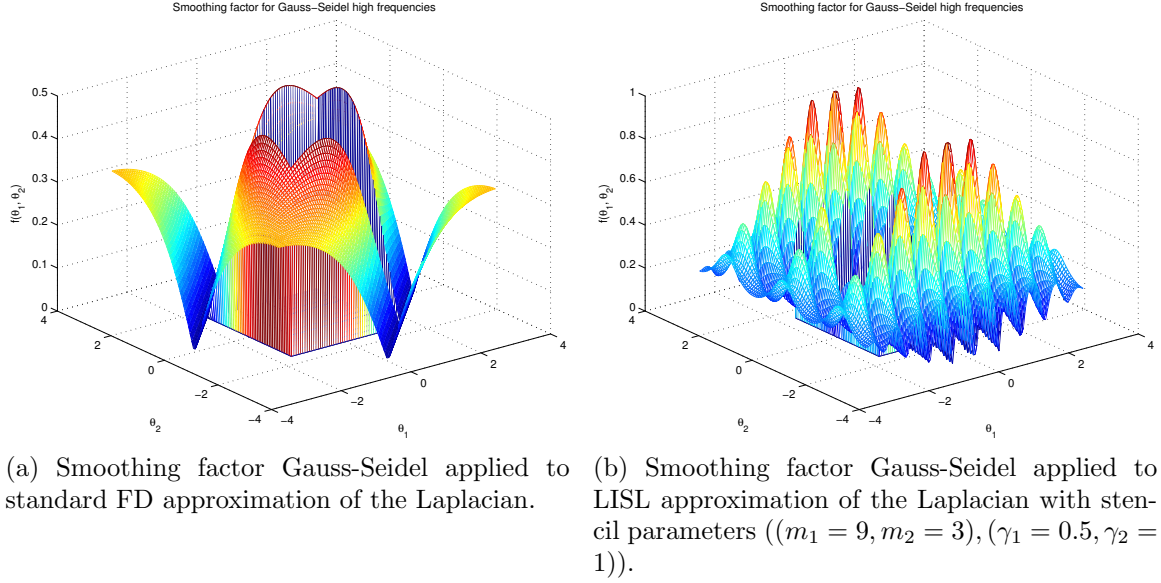


Figure 4.3.1: Representation of the smoothing factor for high frequencies, i.e.  $\theta \in [-\pi, \pi]^2 \setminus [-\pi/2, \pi/2]^2$ , for the Gauss-Seidel iteration for the classical fixed stencil Finite Difference (FD) and the LISL schemes of the two-dimensional Laplacian operator. The maxima calculated numerically are 0.49 (theoretical value is 0.5) for the fixed stencil FD and 0.95 for the LISL scheme (lower is better).

$\sigma_1$  and  $\sigma_2$  have the same sign, then

$$\tilde{L}_{\Delta x}^+(\boldsymbol{\theta}) = \frac{1}{\Delta x}(2 - \bar{g}(\theta_1, \gamma_1, m_1)\bar{g}(\theta_2, \gamma_2, m_2)), \quad \tilde{L}_{\Delta x}^-(\boldsymbol{\theta}) = -\frac{1}{\Delta x}(g(\theta_1, \gamma_1, m_1)g(\theta_2, \gamma_2, m_2)).$$

If, however,  $\sigma_1$  and  $\sigma_2$  have different signs, then

$$\tilde{L}_{\Delta x}^+(\boldsymbol{\theta}) = \frac{1}{\Delta x}(2 - g(\theta_1, \gamma_1, m_1)\bar{g}(\theta_2, \gamma_2, m_2)), \quad \tilde{L}_{\Delta x}^-(\boldsymbol{\theta}) = -\frac{1}{\Delta x}(\bar{g}(\theta_1, \gamma_1, m_1)g(\theta_2, \gamma_2, m_2)).$$

To account for the fact that  $\sigma_i$  can be negative, we re-define  $m_i := \lfloor |\sigma_i|/\sqrt{\Delta x} \rfloor$  and  $\gamma_i := (m_i + 1) - |\sigma_i|/\sqrt{\Delta x}$ . The deterioration of the smoother for large  $m_i$  is present here too.

## 4.4 Performance of geometric multigrid

We conclude the discussion of geometric multigrid by testing its performance against an iterative solver used in [56], i.e. BICGSTAB [78] with and without ILU(0)<sup>1</sup> as preconditioner [70], and algebraic multigrid algorithms, namely, the classical Ruge-Stüben AMG [69] using our own implementation, and AGMG from [63], using the implementation from [61].

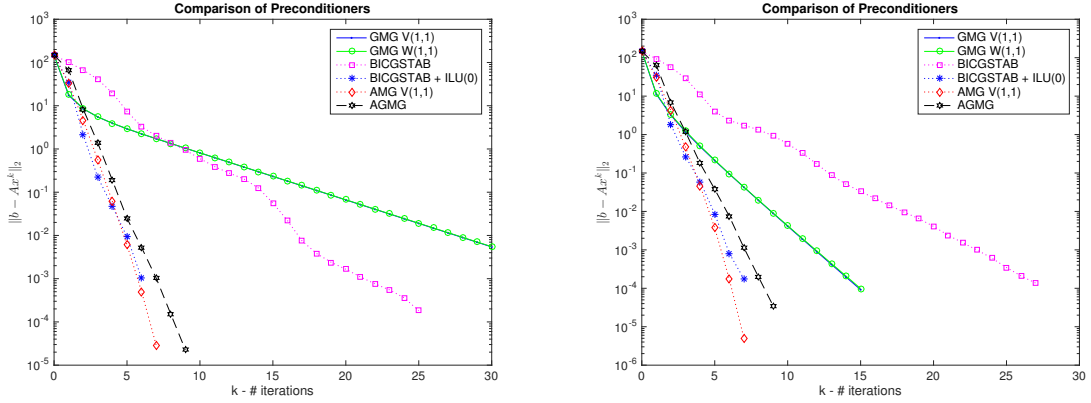
As benchmark examples, we choose a linear system  $Ax = b$  whose coefficient matrix is the LISL discretization of (4.2.1) in the two-dimensional square  $[0, 1]^2$  with Dirichlet boundary conditions, and  $\sigma = 2I_2$  and  $\sigma = \sqrt{5}I_2$ , respectively, where  $I_2$  is the  $2 \times 2$  identity matrix. We use a Cartesian grid with equal number of equispaced nodes in both directions, the smoother is Gauss-Seidel, the prolongation operator bilinear interpolation, the restriction the transpose of the prolongation, and the coarse grid operator is constructed using the Galerkin principle.

Figure 4.4.1 presents the reduction of the residual,  $r_k \equiv \|b - Ax^k\|_2$ , against the number of iterations  $k$  with  $x^0 = 0$ , for a discrete mesh where the distance between two consecutive nodes is  $\Delta x = 2^{-8}$ . The algorithm is stopped whenever the relative residual,  $\|b - Ax^k\|_2 / \|b\|_2$ , measured in the Euclidean norm, is below the prescribed tolerance, in this case  $10^{-6}$ .

Next, we study the residual reduction factor  $\rho = (r_k/r_0)^{1/k}$ , where  $k$  is the number of iterations required for the prescribed tolerance. We solve the problems above for different refinement levels  $\ell$ , where the number of nodes per dimension is  $2^\ell + 1$ . We observe in Table 4.4.1 that for even  $\ell$  the residual reduction factor  $\rho$  corresponding to geometric multigrid cycles for  $\sigma = 2I_2$  is significantly worse than that for  $\sigma = \sqrt{5}I_2$ . This is due to the lack of interpolation when  $2^{\frac{\ell}{2}} \in \mathbb{N}$ , as the step is a multiple of  $\Delta x$ , so for any mesh node  $x_{\Delta x} \in \Omega_{\Delta x}$   $x_{\Delta x} \pm \sqrt{\Delta x} \sigma_i \in \Omega_{\Delta x}$ . The lack of interpolation, and the equal stencil lengths in both directions, gives  $\mu_{\text{loc}} = 1$  in (4.3.2). Moreover, as

---

<sup>1</sup>Incomplete LU factorization with the same sparsity pattern as the original system matrix.



(a) Residual after the  $k$ -th iteration for  $\sigma = 2I_2$ . (b) Residual after the  $k$ -th iteration for  $\sigma = \sqrt{5}I_2$ .

Figure 4.4.1: Residual  $\|b - Ax^k\|_2$  in the Euclidean norm at the end of the  $k$ -th iteration of different geometric and algebraic multigrid cycles when solving (4.2.1) on equispaced Cartesian grid of  $[0, 1]^2$  with 257 nodes per dimension and with homogeneous Dirichlet boundary conditions. Geometric  $V(\nu_1, \nu_2)$  and  $W(\nu_1, \nu_2)$  cycles are considered, where  $\nu_1$  and  $\nu_2$  denote the number of pre- and post-smoothing steps. Their performance is compared to the iterative method BICGSTAB with and without preconditioner, and to two algebraic algorithms, AMG and AGMG from [69] and [63], respectively (see also Sections 4.5 and 4.6). Notice the almost overlapping of lines for geometric  $V(\nu_1, \nu_2)$  and  $W(\nu_1, \nu_2)$  cycles for equal  $\nu_1$  and  $\nu_2$  (see also Table 4.4.1, which shows almost identical rates for  $\ell = 8$ ).

shown in Figure 4.2.1, the eigenvectors corresponding to small eigenvalues are highly oscillatory and hence not resolved sufficiently on the coarse mesh.

Regarding the BICGSTAB iterative solver, we observe the benefit of using ILU(0) as preconditioner, however, the significant increase in the convergence rate (smaller residual reduction factor  $\rho$ ) as the mesh is refined (and hence the condition number of the matrix increases) suggests that convergence is not asymptotically mesh size independent. To further illustrate this, Table 4.4.2 contains the residual reduction factors for AGMG and BICGSTAB with and without preconditioner when solving (4.2.1) in the one-dimensional domain  $[0, 1]$  with homogeneous Dirichlet boundary conditions and discretized using the LISL scheme. We consider the cases  $\sigma = 2$  and  $\sigma = \sqrt{5}$ . As discussed previously, for  $\sigma = 2$  and  $\ell$  even, no interpolation is required. In this case, ILU(0) exactly factorises the system matrix  $A$  and hence  $\rho = 0$ . However,

when interpolation is needed,  $\rho$  approaches 1 for BICGSTAB as the mesh is refined but not for AGMG.

(a) $\rho$ for $\sigma = 2I_2$							
$\ell$	$m$	GMG V(1,1)	GMG W(1,1)	AMG	AGMG	BICGSTAB	BICGSTAB with ILU(0)
6	16	0.4193	0.4204	0.0415	0.1015	0.2502	0.0151
7	22	0.2612	0.2666	0.0633	0.1502	0.5158	0.0767
8	32	0.7561	0.7564	0.0981	0.1551	0.5763	0.1234
9	45	0.5076	0.4905	0.1216	0.1858	0.7109	0.2392
10	64	0.8823	0.8841	0.1219	0.2001	0.7621	0.3382

(b) $\rho$ for $\sigma = \sqrt{5}I_2$							
$\ell$	$m$	GMG V(1,1)	GMG W(1,1)	AMG	AGMG	BICGSTAB	BICGSTAB with ILU(0)
6	17	0.2999	0.2857	0.0403	0.1162	0.3718	0.0262
7	25	0.2565	0.2568	0.0532	0.1367	0.4408	0.0302
8	35	0.4348	0.4300	0.0740	0.1656	0.5938	0.1302
9	50	0.4480	0.4314	0.1124	0.2030	0.6751	0.1746
10	71	0.5547	0.4799	0.1272	0.1992	0.7478	0.3374

Table 4.4.1: The residual reduction factor  $\rho$  for different mesh sizes and different multigrid algorithms, for the two-dimensional Laplace equation; the length of the stencil  $m$  as per (4.2.2).

Returning to the two-dimensional case, the grid hierarchies in the geometric (GMG) and algebraic (AMG) multigrid have 5 levels, including the finest one. In the geometric case, the number of unknowns is 4 times smaller from one level to the next. In the AGMG case, the hierarchy is at most 4 levels deep. Table 4.4.3 reports the complexities for the three categories of algorithms considered. The results confirm the assertion in [82] that AMG coarsening, generally, need not lead to a significant reduction of the unknowns. In the present setting, contrasting the case  $\sigma = \sqrt{5}I_2$  with  $\sigma = 2I_2$  shows that the growth in complexity is due to the interpolation, which creates a denser connectivity graph on the coarser levels.

(a) $\rho$ for $\sigma = 2$				(b) $\rho$ for $\sigma = \sqrt{5}$			
$\ell$	AGMG	BICGSTAB	BICGSTAB with ILU(0)	$\ell$	AGMG	BICGSTAB	BICGSTAB with ILU(0)
10	0.2341	0.6569	0	10	0.3323	0.7545	0.2826
15	0.4684	0.9391	0.7009	15	0.5646	0.9279	0.7867
20	0.5291	0.9948	0	20	0.6345	0.9925	0.9392
21	0.6524	0.9995	0.9498	21	0.4781	0.9987	0.9504

Table 4.4.2: Comparison of the residual reduction factor  $\rho$  for different system sizes and different solvers for the one dimensional Laplace equation. The system size is  $2^\ell + 1$ .

(a) Complexities for $\sigma = 2I_2$							(b) Complexities for $\sigma = \sqrt{5}I_2$						
$\ell$	GMG		AMG		AGMG		$\ell$	GMG		AMG		AGMG	
	$c_G$	$c_A$	$c_G$	$c_A$	$c_G$	$c_A$		$c_G$	$c_A$	$c_G$	$c_A$	$c_G$	$c_A$
6	1.31	3.66	1.75	1.74	1.24	1.18	6	1.31	2.59	1.67	3.78	1.18	1.10
7	1.32	2.69	1.76	6.92	1.26	1.26	7	1.32	2.69	1.74	6.48	1.24	1.22
8	1.33	3.90	1.72	2.05	1.33	1.28	8	1.33	2.65	1.71	7.47	1.20	1.22
9	1.33	2.75	1.71	11.79	1.25	1.31	9	1.33	2.76	1.69	9.34	1.22	1.30
10	1.33	3.97	1.70	2.16	1.25	1.23	10	1.33	2.67	1.64	7.39	1.32	1.45

Table 4.4.3: Comparison of the grid and algebraic complexities as per Definitions 4.1.1 and 4.1.2 for different mesh sizes and different multigrid algorithms, for the two-dimensional case.

## 4.5 Properties of the LISL matrix

In this section, we discuss the theoretical foundation of *Aggregation-based Multigrid* (AGMG) for our specific application of wide stencil discretisations.

The key result is Lemma 3.1 in [63]. The non-negativity of the row sums of a LISL discretization matrix is obtained almost by construction. To see this, let  $A \in \mathbb{R}^{N \times N}$  be the discretization matrix, where  $N := |\Omega_{\Delta x}|$  is the number of mesh points, then the sum for the  $i$ -th row is

$$\sum_{j=1}^N (A)_{ij} = 1 + \Delta t \left( \frac{M}{\Delta x} - c_i^{\alpha, n} \right) - \frac{\Delta t}{\Delta x} M \geq 0,$$

where we have used the fact that for any  $z \in \bar{\Omega}$ ,  $\sum_{j=1}^N w_j(z) = 1$  and the CFL-type condition  $1 - \Delta t c_i^{\alpha, n} \geq 0$ , which is satisfied for sufficiently small  $\Delta t$  independent of  $\Delta x$ .

The following analysis of the non-negativity of the column sum makes use of the regularity of the coefficients  $b$  and  $\sigma$ . In particular, we assume that the coefficients are such that for all  $p \in [[1, P]]$ , and for any mesh points  $x_i, x_l$  and corresponding controls  $\alpha_i, \alpha_l \in \mathcal{A}$  and  $s \in [0, T]$  we have that

$$\|\sigma_p^{\alpha_i}(s, x_i) - \sigma_p^{\alpha_l}(s, x_l)\|_{\infty} \leq L_{\sigma} \|x_i - x_l\|_{\infty}^{\eta}, \quad (4.5.1)$$

$$\|b^{\alpha_i}(s, x_i) - b^{\alpha_l}(s, x_l)\|_{\infty} \leq L_b \|x_i - x_l\|_{\infty}^{\beta}, \quad (4.5.2)$$

where  $\beta \in (0, 1]$ ,  $\eta \in (\frac{1}{2}, 1]$ .

**Remark 4.5.1.** As stated in the introduction, we are working under the standard assumption of Lipschitz continuity of the coefficients in  $x$  and continuity in  $\alpha$ . However, what we require in (4.5.1) is stronger, namely, if the control is inserted in the coefficients as a function of the state  $x$ , the resulting functions are Hölder continuous in  $x$ . This situation arises in every step of the policy iteration algorithm: a control



vector  $(\alpha_i)$  is determined by the optimisation step, and then a linear system with this control vector is solved for  $(x)_i$ . Generally, the optimal control is not a (Hölder) continuous function of the space variables, but there are many important examples where it is at least piecewise Hölder. It can be seen from the proof below that Proposition 4.5.1 still holds in this situation.

**Remark 4.5.2.** Lemma 3.1 in [63] also assumes that the system matrix is irreducible. LISL discretization matrices need not be irreducible, e.g.  $L_{SL}^{2K,2,1}$  for any  $K \in \mathbb{N}$  as in (4.2.3), however, this technical requirement could be overcome by adding an irreducible M-matrix, multiplied by a sufficiently small factor, to the LISL discretization matrix.

We also assume the use of multi-linear interpolation requiring  $2^d$  points to approximate function values in  $\mathbb{R}^d$  and the use of Cartesian grids.

**Proposition 4.5.1.** *Let  $A$  be the LISL discretization matrix of (1.1.1) for a given time step, on an equispaced Cartesian grid  $\Omega_{\Delta x}$  of  $\Omega \subset \mathbb{R}^d$  with  $\Delta x > 0$ , and a given vector of control values  $(\alpha_i)_{i=1,\dots,N}$ ,  $\alpha_i \in \mathcal{A}$ , associated with the mesh points  $x_i$ ,  $1 \leq i \leq N := |\Omega_{\Delta x}|$ . Assume that (4.5.1) holds.*

*Then the column sum of the matrix is non-negative provided*

$$\Delta t \leq \frac{\Delta x}{\sup_{\alpha \in \mathcal{A}} |c^{\alpha,+}| + (\mathcal{M} - 1)(P + 1)}, \quad (4.5.3)$$

where  $\mathcal{M}$  depends on the dimension of the domain  $d$  and the Lipschitz constants  $L_\sigma$  and  $L_b$ , but not on the mesh parameter  $\Delta x$ . Indeed,  $\mathcal{M} = 3^d$  for sufficiently small  $\Delta x$ .

*Proof.* We carry out the proof for Scheme 2 of [26], but an analogous analysis holds for Schemes 1 and 3 in the introduction. We also note that we can restrict the analysis to steps where no truncation is required, as in the case of truncation the weights only

contribute (positively) to the diagonal of the matrix and the right-hand side of the equation (see Remark 2.2.3).

For simplicity of notation, we omit the dependence of the coefficient functions  $b$  and  $\sigma_p$  on the time variable  $t$  and the control. For any  $i \neq j$  the matrix entry  $(A)_{ij} \neq 0$  if and only if for any  $1 \leq m \leq P+1$  we require  $\phi(x_j)$  to approximate – by means of linear interpolation –  $\phi(x_i + y_m^\pm(x_i))$ , where  $y_m^\pm(x_i)$  is either  $y_p^\pm(x_i) = \pm\sqrt{\Delta x}\sigma_p(x_i)$  for  $1 \leq p \leq P$  or  $y_{P+1}^\pm(x_i) = \Delta x b(x_i)$ . For any two nodes  $i$  and  $l$  to contribute to the sum of column  $j$ , it is necessary that

$$\|x_l + y_m^\pm(x_l) - (x_i + y_m^\pm(x_i))\|_\infty < 2\Delta x.$$

As  $x_l$  and  $x_i$  lie on the grid, there exists a positive constant  $M$  such that  $\|x_l - x_i\|_\infty = M\Delta x$ . Then  $(M+1)^d$  constitutes an upper bound on the number of terms the step  $y_m^\pm(\cdot)$  contributes to the sum of column  $j$ .

We consider the different possible values for  $y_m^\pm(x_l)$  separately. First, assume  $y_m^\pm(x_l) = \Delta x b(x_l)$  and let  $\Delta x \leq 1/(L_b\sqrt{d})$ , then

$$2\Delta x > \|x_l + \Delta x b(x_l) - (x_i + \Delta x b(x_i))\|_\infty \geq M\Delta x - \Delta x^{1+\beta} L_b M^\beta, \quad (4.5.4)$$

where we have used the triangle inequality and the Hölder regularity of  $b$ . Rearranging,

$$M < \frac{2}{1 - L_b M^{\beta-1} \Delta x^\beta},$$

such that  $M \leq 2$  for sufficiently small  $\Delta x$ . Proceeding similarly for  $y_m^\pm(x) = \pm\sqrt{\Delta x}\sigma_m(x)$ , we obtain again  $M \leq 2$  as  $\Delta x \rightarrow 0$ .

Denote  $\mathcal{M}$  to be the maximum of all of the  $(M+1)^d$ s above, i.e. for different mesh

points, then the column sum gives

$$\begin{aligned} \sum_{i=1}^N (A)_{ij} &= 1 + \Delta t \left( \frac{P+1}{\Delta x} - c_j^{\alpha,n} \right) - \frac{\Delta t}{2\Delta x} \sum_{i=1}^N \sum_{p=1}^{P+1} w_j(x_i + y_p(x_i)) \\ &\geq 1 - \frac{\Delta t}{\Delta x} (c_j^{\alpha,n} + (\mathcal{M} - 1)(P + 1)). \end{aligned} \quad (4.5.5)$$

Therefore, non-negativity of the sum is guaranteed by condition (4.5.3).

□

**Remark 4.5.3.** For the LISL scheme to be first order accurate, it is required that  $\Delta t \sim \mathcal{O}(\Delta x)$ . Therefore, the bound (4.5.3) does not impose problematic restrictions on the size of the time steps. We recall that  $\Delta t = \mathcal{O}(\Delta x)$  and  $\Delta t = \mathcal{O}(\Delta x^{3/2})$  are the CFL conditions for the explicit schemes without and with truncation, respectively, see Corollary 2.2.5. Therefore, fully implicit time stepping with policy iteration and AGMG preconditioning is the computationally most efficient overall algorithm among the ones considered.

## 4.6 Performance of the algebraic approaches

We compare the performance of the classical AMG implementation in the HSL library [44], and AGMG from [61] for the benchmark optimal control problems in Section 2.2.4. Both of these methods are used as preconditioners for a Krylov subspace method that is assumed to have converged when the relative residual is smaller than  $10^{-6}$ . In particular, we use MATLAB’s implementation of GMRES [71] for AMG and GCR [28] for AGMG. We use GMRES for AMG as GCR is not among the solvers implemented by default in MATLAB. However, as noted in [71] GMRES “is theoretically equivalent to the Generalized Conjugate Residual (GCR)”, moreover “it also requires only half the storage required by the GCR method and needs fewer arithmetic operations than GCR”.

The AMG preconditioner consists of one iteration of the standard V-cycle with two Gauss-Seidel pre- and post-smoothing steps, whereas AGMG uses one Gauss-Seidel pre- and post-smoothing step and the enhanced multigrid cycles mentioned in the introduction, see [62, 63]. For completeness, we also include as benchmark MATLAB's sparse direct solver using UMFPACK [25]. The problems considered have smooth closed form solutions linear in  $t$ . As mentioned in the previous section, we employ policy iteration to solve the resulting non-linear discrete problem. The tests were run on a Linux machine under MATLAB 2015a, on a quad-core AMD 4.2GHz with 7.5GB of RAM and 15GB of swap.

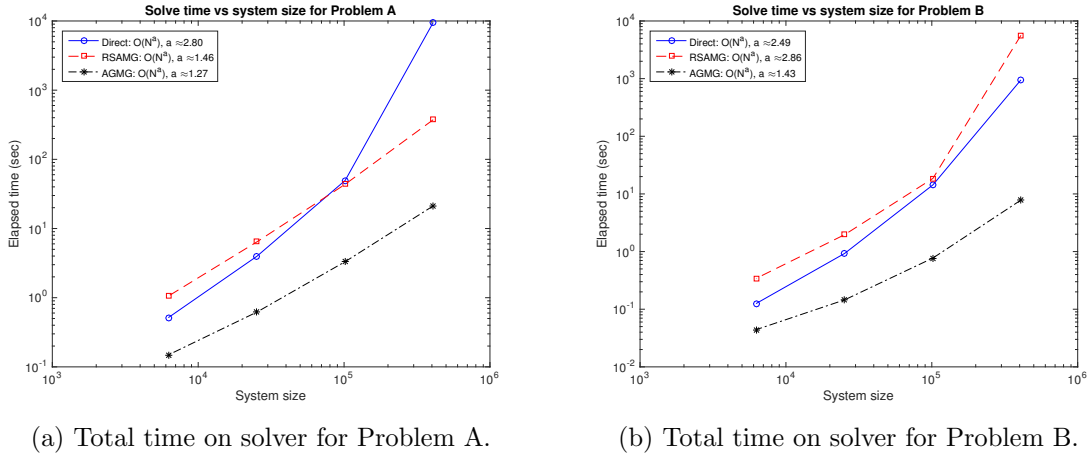
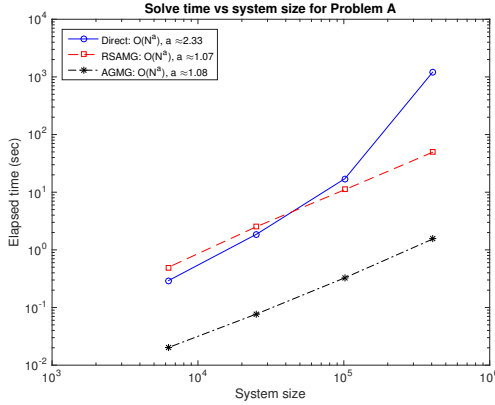
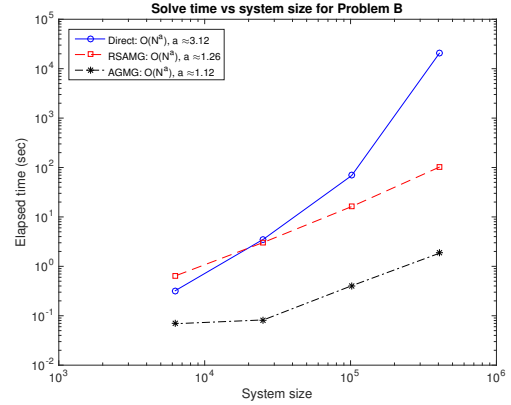


Figure 4.6.1: Total number of seconds for solving the linear systems versus the size of the systems for each of the linear system solvers considered. We use equispaced Cartesian grids in space with 81, 161, 321 and 641 nodes per dimension and one time step.

In Figure 4.6.1 we present the elapsed time solving linear systems for a single time step ( $\Delta t = T$ ). Both MG methods provide a solution with the same accuracy as the sparse direct solver but with improved scalability. AGMG outperforms AMG and the sparse direct solver in both problems. Figure 4.6.2 shows the average time spent solving linear systems per time step when  $\Delta t = \Delta x$ . Reducing  $\Delta t$  makes the system matrix more diagonally dominant and as a consequence easier to precondition. This effect is noticeable for Problem B using AMG as preconditioner, see Table 4.6.1.



(a) Average time on solver for Problem A.



(b) Average time on solver for Problem B.

Figure 4.6.2: Average number of seconds per time step for solving the linear systems versus the size of the systems. We use equispaced Cartesian grids in space with 81, 161, 321 and 641 nodes per dimension and  $\Delta t = \Delta x$ .

	$N_x$	Direct		AMG		AGMG	
		$\Delta t = T$	$\Delta t = \Delta x$	$\Delta t = T$	$\Delta t = \Delta x$	$\Delta t = T$	$\Delta t = \Delta x$
Problem A	321	49.60	17.93	43.62	10.07	3.31	0.41
	641	9.50e+03	4.33e+03	373.77	48.82	21.22	1.84
Problem B	321	14.35	68.70	18.59	16.19	0.77	0.40
	641	950.62	2.09e+04	5.64e+03	102.57	7.82	1.85

Table 4.6.1: Average seconds per time step solving linear systems.

Table 4.6.3 and Table 4.6.4 report memory consumption and quantities related to the Krylov subspace method and to the coarsening. As commented in the previous sections, AMG results in grid and algebraic complexities higher than AGMG's. The coarsening for both of the methods is stopped when the coarse level system is cheap to solve exactly compared to the starting system, specifically, we stop whenever the number of unknowns at the coarse level is comparable to the cubic root of the initial number of unknowns. The effect of simplifying the intergrid transfer operators can be observed on the coarse to fine stencil ratio (C/F stencil). For AGMG, the stencil on the coarsest level is similar to the initial one, whereas for AMG it is significantly denser. The fact that aggregation-based coarsening strategies yield coarse matrices with similar sparsity as the original one was noted in [48]. Moreover, AGMG yields

shallower hierarchies due to higher coarsening factors. The effect of reducing  $\Delta t$  is also appreciated in this ratio. We observe that the direct method's memory consumption increases dramatically while for the MG methods, we note the relation between the memory requirement and the algebraic complexity of the method.

The number of Krylov iterations highlight previous comments on the fact that aggregation-based multigrid methods are not efficient if used as stand-alone solvers: in all test cases, AGMG used more iterations than AMG per policy iteration. However, AGMG used as a preconditioner to a Krylov subspace method provides accurate solutions faster and cheaper than the other two solvers considered.

	$N_x$	Direct		AMG		AGMG	
		$\Delta t = T$	$\Delta t = \Delta x$	$\Delta t = T$	$\Delta t = \Delta x$	$\Delta t = T$	$\Delta t = \Delta x$
Problem A	161	44.17%	69.93%	64.21%	75.42%	12.40%	7.63%
	321	53.34%	82.36%	69.24%	77.24%	15.09%	10.82%
	641	78.01%	85.91%	34.20%	67.47%	5.53%	10.37%
Problem B	161	8.46%	80.71%	49.16%	79.35%	7.27%	8.92%
	321	26.09%	94.36%	77.48%	76.28%	12.88%	8.83%
	641	95.25%	97.65%	98.64%	87.68%	2.48%	17.06%

Table 4.6.2: Percentage of computational time spent in linear solvers for the Examples in Section 4.6.

(a) Peak memory consumption measured in GB for  $\Delta t = T$

	$N_x$	Direct		AMG		AGMG	
		VIRT	RES	VIRT	RES	VIRT	RES
Problem A	321	11.40	5.29	10.63	5.10	9.83	4.14
	641	25.99	7.13	14.40	7.05	12.66	6.76
Problem B	321	11.71	6.50	9.84	5.14	9.84	5.15
	641	23.83	7.10	18.51	7.13	9.90	5.12

(b) Peak memory consumption measured in GB for  $\Delta t = \Delta x$

	$N_x$	Direct		AMG		AGMG	
		VIRT	RES	VIRT	RES	VIRT	RES
Problem A	321	8.39	3.20	6.41	2.38	6.48	2.39
	641	21.80	7.22	10.61	6.68	10.52	6.64
Problem B	321	13.76	7.09	9.83	3.67	9.83	3.71
	641	26.11	7.26	12.72	7.23	9.90	5.17

Table 4.6.3: Peak memory consumption statistics in gigabytes (GB) of the MATLAB process sampled using the shell command `top`. VIRT is the total amount of virtual memory used by MATLAB, whereas RES is the non-swapped physical memory (limited to 7.5).

(a) Krylov iterations and coarsening related quantities for  $\Delta t = T$ 

	Solver	$N_x$	Avg Krylov It	# levels	C/F stencil	$c_G$	$c_A$
Problem A	AMG	321	4.00	9.0	26.25	2.61	7.06*
		641	4.29	11.0	22.33	2.42	4.63*
	AGMG	321	12.67	5.83	1.30	1.81	1.97
		641	17.14	6.14	1.36	1.61	1.77
Problem B	AMG	321	5.00	7.0	86.26	2.08	9.60*
		641	6.67	9.5	221.43	2.23	11.61*
	AGMG	321	12.00	5.00	0.50	1.60	1.92
		641	15.00	5.50	0.39	1.53	1.57

(b) Krylov iterations and coarsening related quantities for  $\Delta t = \Delta x$ 

	Solver	$N_x$	Avg Krylov It	# levels	C/F stencil	$c_G$	$c_A$
Problem A	AMG	321	3.00	9.98	16.30	2.76	5.90*
		641	3.00	12.0	16.60	2.75	5.12*
	AGMG	321	6.00	2.00	0.23	1.00	1.00
		641	6.00	2.00	0.26	1.00	1.00
Problem B	AMG	321	2.98	8.61	48.65	2.47	8.61*
		641	2.99	11.27	107.77	2.70	11.12*
	AGMG	321	4.99	2.96	0.31	1.07	1.02
		641	5.01	2.97	0.33	1.15	1.10

Table 4.6.4: Quantities related to the Krylov subspace iteration and multigrid coarsening. *Avg Krylov It* contains the average number of Krylov iterations over all time steps and all policy iterations; *# levels* contains the average depth in the grid hierarchy; *C/F stencil* contains the ratio between the stencil at the coarsest level and that on the finest level (lower is better). On the finest level, the stencil is close to 11 for Problem A and close to 8 for Problem B. The last two columns report the grid and algebraic complexity as per Definitions 4.1.1 and 4.1.2. As the full matrix hierarchy was not available from [44] for AMG, but only the coarsest and finest matrices, the starred algebraic complexities are estimates based on the assumption of a geometrically decreasing complexity between the coarsest and finest level, which is likely to be a significant underestimate.



# Chapter 5

## Conclusion

This thesis focused on the study of wide-stencil discretizations of second order non-linear parabolic differential operators. First, we studied the truncation of the stencil for problems on bounded domains, as the method may overstep the boundaries for nodes in a surrounding layer. Our main result detailed the construction of such truncation and proves that the resulting scheme remains consistent, monotone and conditionally stable. Numerical examples confirmed that the truncation improves the accuracy of the approach compared to constant and linear extrapolation of the boundary conditions. These examples also confirmed the worsening of the CFL condition for explicit time-stepping schemes. For the Black-Scholes problem we showed that, under suitable domain transformations, the semi-Lagrangian stencil does not step outside the transformed domain.

Second, we have obtained error bounds applicable to numerical schemes approximating the Cauchy-Dirichlet problem for HJB equations on bounded domains. The truncated schemes described in this thesis being one of such schemes.

Third, driven by the stringent CFL condition of explicit time-stepping schemes for the method with truncated stencils, we considered implicit schemes and the application of multigrid preconditioners to efficiently solve the resulting discrete non-

linear system of equations. Using theoretical and empirical arguments, we showed the need to employ multigrid methods based on algebraic ideas. We demonstrated that aggregation-based methods are well suited for the discretization schemes considered and justified their use by proving that under mild conditions on the mesh refinement parameters, the LISL discretization matrices are M-matrices with non-negative row and column sums. These algorithms are shown to compare favourably against other multigrid methods and sparse direct solvers.

Although we only considered linear interpolation, much of the analysis, in particular the matrix properties in Section 4.5, will also hold if other limited interpolations (see, e.g., [26], [81]), are used, as only the properties in (2.2.14) are critical.

A possible direction of future research is to generalise the ideas in Section 2.3 and the analysis in Chapter 3. Regarding the domain transformation for the Black-Scholes equation, the usual well-posedness assumptions limit the growth of the drift and diffusion coefficients. Therefore, if the transformation compensates for this growth, it seems possible to generalise the ideas to other models. It would also be interesting to benchmark the domain transformation against the domain localisation for problems like lookback options given that, as pointed out in [5], it is not clear how the solution of this problem behaves at infinity. This affects the quality of the artificial boundary condition.

The domain stretching used for the error analysis in Chapter 3, works well for star shaped domains centred at the origin, but it would be desirable to consider such transformations in a wider sense. We also found that most of the literature on semi-Lagrangian schemes or on regularity for non-linear second order parabolic equations focuses on unbounded domains, it seems that the routine pointed by Ishii and Lions<sup>1</sup> has ceased to exist and that the treating the unbounded case is actually the goal. In that sense another direction of research would be to better understand the regularity

---

<sup>1</sup>“It is a routine work to adapt these results to the case of unbounded domains” [46, p. 29].

of viscosity solutions in bounded domains with boundary conditions and improve the set of assumptions in Chapter 3: for instance, to study if there are ways to relate the regularity of the solution to the shape of the domains as in [39] for the linear case.

In Chapter 4 we used spectral analysis of the discretization matrices to justify the use of algebraic multigrid over geometric multigrid. Further analysis of the subspaces resulting from the aggregation procedure may help to modify the algorithm so that it is better suited for LISL discretization matrices. For instance, we may be able to use the fact that the coefficients of the matrix are essentially interpolation weights to construct better aggregation algorithms for LISL discretization matrices.



# Bibliography

- [1] R. Abgrall. Numerical discretization of boundary conditions for first order Hamilton–Jacobi equations. *SIAM Journal on Numerical Analysis*, 41(6):2233–2261, 2003.
- [2] Y. Achdou and M. Falcone. A semi-Lagrangian scheme for mean curvature motion with nonlinear Neumann conditions. *Interfaces and Free Boundaries*, 14(4):455–485, 2012.
- [3] M. Akian, P. Séquier, and A. Sulem. A finite horizon multidimensional portfolio selection problem with singular transactions. In *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, volume 3, pages 2193–2198. IEEE, 1995.
- [4] M. Avellaneda, A. Levy, and A. Parás. Pricing and hedging derivative securities in markets with uncertain volatilities. *Applied Mathematical Finance*, 2(2):73–88, 1995.
- [5] G. Barles. Convergence of numerical schemes for degenerate parabolic equations arising in finance theory. In L. C. G. Rogers and D. Talay, editors, *Numerical methods in finance*, volume 13 of *Publications of the Newton Institute*, chapter 1, pages 1–21. Cambridge University Press, 1997.
- [6] G. Barles, C. Daher, and M. Romano. Convergence of numerical schemes for parabolic equations arising in finance theory. *Mathematical Models and Methods in Applied Sciences*, 5(1):125–143, 1995.
- [7] G. Barles and E. R. Jakobsen. On the convergence rate of approximation schemes for Hamilton–Jacobi–Bellman equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 36(1):33–54, 2002.
- [8] G. Barles and E. R. Jakobsen. Error bounds for monotone approximation schemes for parabolic Hamilton–Jacobi–Bellman equations. *Mathematics of Computation*, 76(260):1861–1893, 2007.
- [9] G. Barles and E. Rouy. A strong comparison result for the Bellman equation arising in stochastic exit time control problems and its applications. *Communications in Partial Differential Equations*, 23(11-12):552–562, 1998.

- [10] G. Barles and P. E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Analysis*, 4(3):271–283, 1991.
- [11] F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973.
- [12] M. Bloß and R. H. W. Hoppe. Numerical computation of the value function of optimally controlled stochastic switching processes by multi-grid techniques. *Numerical Functional Analysis and Optimization*, 10(3-4):275–304, 1989.
- [13] O. Bokanowski, M. Falcone, R. Ferretti, D. Kalise, L. Grüne, and H. Zidani. Value iteration convergence of  $\varepsilon$ -monotone schemes for stationary Hamilton–Jacobi equations. *Discrete and Continuous Dynamical Systems - Series A*, 35(9):4041–4070, 2015.
- [14] O. Bokanowski, S. Maroso, and H. Zidani. Some convergence results for Howard’s algorithm. *SIAM Journal on Numerical Analysis*, 47(4):3001–3026, 2009.
- [15] O. Bokanowski, A. Picarelli, and H. Zidani. Dynamic programming and error estimates for stochastic control problems with maximum cost. *Applied Mathematics & Optimization*, 71(1):125–163, 2015.
- [16] L. Bonaventura and R. Ferretti. Semi-Lagrangian methods for parabolic problems in divergence form. *SIAM Journal on Scientific Computing*, 36(5):A2458–A2477, 2014.
- [17] J. F. Bonnans, É. Ottenwaelter, and H. Zidani. A fast algorithm for the two dimensional HJB equation of stochastic control. *ESAIM: Mathematical Modelling and Numerical Analysis*, 38(4):723–735, 2004.
- [18] L. A. Caffarelli and P. E. Souganidis. A rate of convergence for monotone finite difference approximations to fully nonlinear uniformly elliptic PDEs. *Communications on Pure and Applied Mathematics*, 61(1):1–17, 2008.
- [19] F. Camilli and M. Falcone. An approximation scheme for the optimal control of diffusion processes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 29(1):97–122, 1995.
- [20] S. Chaumont. Uniqueness to elliptic and parabolic Hamilton–Jacobi–Bellman equations with non-smooth boundary. *Comptes Rendus Mathématique*, 339(8):555–560, 2004.
- [21] J. C. Cox, J. E. Ingersoll, and S. A. Ross. A theory of the term structure of interest rates. *Econometrica*, 53(2):385–407, 1985.
- [22] M. G. Crandall and H. Ishii. The maximum principle for semicontinuous functions. *Differential Integral Equations*, 3(6):1001–1014, 1990.

- [23] M. G. Crandall, H. Ishii, and P.-L. Lions. User's guide to viscosity solutions of second order partial differential equations. *Bulletin of the American Mathematical Society*, 27(1):1–67, 1992.
- [24] M. G. Crandall and P.-L. Lions. Convergent difference schemes for nonlinear parabolic equations and mean curvature motion. *Numerische Mathematik*, 75(1):17–41, 1996.
- [25] T. A. Davis. Algorithm 832: UMFPACK V4.3—an Unsymmetric-pattern Multifrontal Method. *ACM Transactions on Mathematical Software*, 30(2):196–199, June 2004.
- [26] K. Debrabant and E. R. Jakobsen. Semi-Lagrangian schemes for linear and fully non-linear diffusion equations. *Mathematics of Computation*, 82(283):1433–1462, 2013.
- [27] K. Debrabant and E. R. Jakobsen. Semi-Lagrangian schemes for linear and fully non-linear Hamilton–Jacobi–Bellman equations. In F. Ancona et al., editors, *Hyperbolic Problems: Theory, Numerics, Applications*, volume 8 of *AIMS Series on Applied Mathematics*, pages 483–490, Springfield, MO, 2014. American Institute of Mathematical Sciences (AIMS).
- [28] S. C. Eisenstat, H. C. Elman, and M. H. Schultz. Variational iterative methods for nonsymmetric systems of linear equations. *SIAM Journal on Numerical Analysis*, 20(2):345–357, 1983.
- [29] M. Falcone. Some remarks on the synthesis of feedback controls via numerical methods. In J. L. Menaldi, E. Rofman, and A. Sulem, editors, *Optimal Control and Partial Differential Equations*, pages 456–465. IOS Press, 2001.
- [30] M. Falcone and R. Ferretti. *Semi-Lagrangian Approximation Schemes for Linear and Hamilton–Jacobi Equations*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013.
- [31] W. Feller. Two singular diffusion problems. *The Annals of Mathematics*, 54(1):173–182, 1951.
- [32] X. Feng, R. Glowinski, and M. Neilan. Recent developments in numerical methods for fully nonlinear second order partial differential equations. *SIAM Review*, 55(2):205–267, 2013.
- [33] R. Ferretti. On the relationship between Semi-Lagrangian and Lagrange-Galerkin schemes. *Numerische Mathematik*, 124(1):31–56, 2013.
- [34] W. H. Fleming and H. M. Soner. *Controlled Markov processes and viscosity solutions*, volume 25. Springer Science & Business Media, 2006.

- [35] P. A. Forsyth and K. R. Vetzal. Numerical methods for nonlinear PDEs in finance. In J.-C. Duan, W. K. Härdle, and J. E. Gentle, editors, *Handbook of Computational Finance*, Springer Handbooks of Computational Statistics, pages 503–528. Springer Berlin Heidelberg, 2012.
- [36] M. I. Freidlin. *Functional Integration and Partial Differential Equations*, volume 109 of *Annals of Mathematics Studies*. Princeton University Press, 1985.
- [37] A. Friedman. *Stochastic Differential Equations and Applications*. Number v. 2 in Probability and mathematical statistics. Academic Press, 1975.
- [38] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*, volume 224 of *Classics in Mathematics*. Springer, 2001.
- [39] P. Grisvard. *Singularities in boundary value problems*, volume 22 of *Research Notes in Applied Mathematics*. Springer, 1992.
- [40] J. Guyon and P. Henry-Labordère. *Nonlinear option pricing*. CRC Press, 2013.
- [41] D. Han and J. W. L. Wan. Multigrid methods for second order Hamilton–Jacobi–Bellman and Hamilton–Jacobi–Bellman–Isaacs equations. *SIAM Journal on Scientific Computing*, 35(5):S323–S344, 2013.
- [42] J. Heinonen. *Lectures on analysis on metric spaces*. Springer Science & Business Media, 2012.
- [43] R. H. Hoppe. Multi-grid methods for Hamilton–Jacobi–Bellman equations. *Numerische Mathematik*, 49(2-3):239–254, 1986.
- [44] HSL. A collection of Fortran codes for large scale scientific computation, <http://www.hsl.rl.ac.uk/>, 2015.
- [45] H. Ishii and S. Koike. Viscosity solutions of a system of nonlinear second-order elliptic PDEs arising in switching games. *Funkcialaj Ekvacioj*, 34(1):143–155, 1991.
- [46] H. Ishii and P.-L. Lions. Viscosity solutions of fully nonlinear second-order elliptic partial differential equations. *Journal of Differential Equations*, 83(1):26–78, 1990.
- [47] E. R. Jakobsen and K. H. Karlsen. Continuous dependence estimates for viscosity solutions of fully nonlinear degenerate parabolic equations. *Journal of Differential Equations*, 183(2):497–525, 2002.
- [48] H. Kim, J. Xu, and L. Zikatanov. A multigrid method based on graph matching for convection–diffusion equations. *Numerical Linear Algebra with Applications*, 10(1-2):181–195, 2003.



- [49] S. Koike. Uniqueness of viscosity solutions for monotone systems of fully nonlinear PDEs under Dirichlet condition. *Nonlinear Analysis: Theory, Methods & Applications*, 22(4):519–532, 1994.
- [50] N. V. Krylov. On the rate of convergence of finite-difference approximations for Bellman’s equations. *St. Petersburg Mathematical Journal*, 9(3):639–650, 1998.
- [51] N. V. Krylov. On the rate of convergence of finite-difference approximations for Bellman’s equations with variable coefficients. *Probability Theory and Related Fields*, 117(1):1–16, 2000.
- [52] N. V. Krylov. On the rate of convergence of finite-difference approximations for elliptic Isaacs equations in smooth domains. *Communications in Partial Differential Equations*, 40(8):1393–1407, 2015.
- [53] H. J. Kushner and P. Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time*, volume 24. Springer Science & Business Media, 2001.
- [54] P.-L. Lions. Optimal control of diffusion processes and Hamilton–Jacobi–Bellman equations part 2 : viscosity solutions and uniqueness. *Communications in Partial Differential Equations*, 8(11):1229–1276, 1983.
- [55] T. J. Lyons. Uncertain volatility and the risk-free synthesis of derivatives. *Applied Mathematical Finance*, 2(2):117–133, 1995.
- [56] K. Ma and P. A. Forsyth. An unconditionally monotone numerical scheme for the two factor uncertain volatility model. *IMA Journal of Numerical Analysis*, 2016. To appear.
- [57] E. J. McShane. Extension of range of functions. *Bulletin of the American Mathematical Society*, 40(12):837–842, 1934.
- [58] J.-L. Menaldi. Some estimates for finite difference approximations. *SIAM Journal on Control and Optimization*, 27(3):579–607, 1989.
- [59] T. S. Motzkin and W. Wasow. On the approximation of linear elliptic differential equations by difference equations with positive coefficients. *Journal of Mathematics and Physics*, 31(1):253–259, 1952.
- [60] A. Napov and Y. Notay. An algebraic multigrid method with guaranteed convergence rate. *SIAM Journal on Scientific Computing*, 34(2):A1079–A1109, 2012.
- [61] Y. Notay. Homepage for AGMG, <http://homepages.ulb.ac.be/~ynotay/AGMG/>.
- [62] Y. Notay. An aggregation-based algebraic multigrid method. *Electronic Transactions on Numerical Analysis*, 37(6):123–146, 2010.
- [63] Y. Notay. Aggregation-based algebraic multigrid for convection-diffusion equations. *SIAM Journal on Scientific Computing*, 34(4):A2288–A2316, 2012.

- [64] Y. Notay and P. S. Vassilevski. Recursive Krylov-based multigrid cycles. *Numerical Linear Algebra with Applications*, 15(5):473–487, 2008.
- [65] A. M. Oberman. Convergent difference schemes for degenerate elliptic and parabolic equations: Hamilton–Jacobi equations and free boundary problems. *SIAM Journal on Numerical Analysis*, 44(2):879–895, 2006.
- [66] O. A. Oleinik and E. V. Radkevich. *Second order equations with nonnegative characteristic form*. American Mathematical Society Providence, 1973.
- [67] H. Pham. On some recent aspects of stochastic control and their applications. *Probability Surveys*, 2:506–549, 2005.
- [68] C. Reisinger and J. Rotaetxe Arto. Boundary treatment and multigrid preconditioning for semi-Lagrangian schemes applied to Hamilton–Jacobi–Bellman equations. *Journal of Scientific Computing*, pages 1–33, 2017.
- [69] J. W. Ruge and K. Stüben. Algebraic multigrid. *Multigrid methods*, 3:73–130, 1987.
- [70] Y. Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, second edition, 2003.
- [71] Y. Saad and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986.
- [72] I. Smears and E. Süli. Discontinuous Galerkin finite element approximation of Hamilton–Jacobi–Bellman equations with Cordès coefficients. *SIAM Journal on Numerical Analysis*, 52(2):993–1016, 2014.
- [73] I. Smears and E. Süli. Discontinuous Galerkin finite element methods for time-dependent Hamilton–Jacobi–Bellman equations with Cordès coefficients. *Numerische Mathematik*, 133(1):141–176, 2016.
- [74] R. M. Stulz. Options on the minimum or the maximum of two risky assets. *Journal of Financial Economics*, 10(2):161–185, 1982.
- [75] K. Stüben. A review of algebraic multigrid. *Journal of Computational and Applied Mathematics*, 128(1–2):281–309, 2001. Numerical Analysis 2000. Vol. VII: Partial Differential Equations.
- [76] D. Tavella and C. Randall. *Pricing financial instruments: The finite difference method*, volume 13. John Wiley & Sons, 2000.
- [77] U. Trottenberg, C. Oosterlee, and A. Schüller. *Multigrid*. Academic Press, 2001.
- [78] H. A. van der Vorst. Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 13(2):631–644, 1992.

- [79] P. S. Vassilevski. *Multilevel block factorization preconditioners: Matrix-based analysis and algorithms for solving finite element equations*. Springer Science & Business Media, 2008.
- [80] J. Wang and P. A. Forsyth. Maximal use of central differencing for Hamilton–Jacobi–Bellman PDEs in finance. *SIAM Journal on Numerical Analysis*, 46(3):1580–1601, 2008.
- [81] X. Warin. Some non-monotone schemes for time dependent Hamilton–Jacobi–Bellman equations in stochastic control. *Journal of Scientific Computing*, 66(3):1122–1147, 2016.
- [82] A. J. Wathen. Preconditioning. *Acta Numerica*, 24:329–376, 2015.
- [83] H. Windcliff, P. A. Forsyth, and K. R. Vetzal. Analysis of the stability of the linear boundary condition for the Black-Scholes equation. *Journal of Computational Finance*, 8:65–92, 2004.
- [84] J. Young and X. Y. Zhou. *Stochastic Controls: Hamiltonian Systems and HJB Equations*, volume 43 of *Applications of Mathematics*. Springer, 1999.
- [85] Y.-L. Zhu and J. E. Li. Multi-factor financial derivatives on finite domains. *Communications in Mathematical Sciences*, 1(2):343–359, 2003.