



## **Aproximação Numérica de Soluções de Problemas de Obstáculo**

**Rui Manuel Lança dos Santos**

Dissertação para obtenção do Grau de Mestre em

**Matemática Aplicada e Computação**

Orientador: Prof. Juha Hans Videman

### **Júri**

Presidente: Prof. Pedro Miguel Rita da Trindade e Lima  
Orientador: Prof. Juha Hans Videman  
Vogal: Prof. Pedro Ricardo Simão Antunes

**junho 2024**

This work was created using  $\text{\LaTeX}$  typesetting language  
in the Overleaf environment ([www.overleaf.com](http://www.overleaf.com)).

# **Declaração**

Declaro que o presente documento é um trabalho original da minha autoria e que cumpre todos os requisitos do Código de Conduta e Boas Práticas da Universidade de Lisboa.



# **Agradecimentos**

Primeiramente quero agradecer à minha família, em particular à minha mãe e ao meu pai, mas especialmente à minha mãe, Margarida, pois sem ela nada disto seria possível. Um cumprimento a todos os meus amigos, colegas e professores que de alguma forma estiveram presentes e tiveram algum tipo de influência na minha jornada. Obrigado.



# **Abstract**

The main objective of this dissertation was to develop numerical methods for approximating solutions to elliptic and parabolic obstacle problems, using finite difference and finite element methods. These problems were formulated as variational inequalities, complementarity problems and other suitable mathematical approaches that include a free boundary as one of the unknowns.

The dissertation addressed the elliptic obstacle problem applied to the elastic membrane, investigating its mathematical characteristics, introducing Lagrange multipliers and implementing numerical methods for its solution. In financial maths, the Black-Scholes model for the valuation of American options was considered, treated as a parabolic obstacle problem.

Finite difference and finite element methods were implemented to numerically solve various obstacle problems. The dissertation evaluated the accuracy and convergence of the proposed methods in solving the problems, showing their robustness and applicability in practical scenarios.

# **Keywords**

Obstacle Problem; Finite Element; Finite Differences; Elastic Membrane; Black-Scholes Model.



# Resumo

O objetivo principal desta dissertação foi desenvolver métodos numéricos para a aproximação de soluções para problemas de obstáculo elípticos e parabólicos, utilizando os métodos das diferenças finitas e dos elementos finitos. Estes problemas foram formulados como inequações variacionais, problemas de complementaridade e outras abordagens matemáticas adequadas, que incluem uma fronteira livre como uma das incógnitas.

A dissertação abordou o problema de obstáculo elíptico aplicado à membrana elástica, analisando as suas características matemáticas, introduzindo multiplicadores de Lagrange e a implementação de métodos numéricos para obter a sua solução. Na matemática financeira, considerou-se o modelo de Black-Scholes para a valorização de opções americanas, tratado como um problema de obstáculo parabólico.

Foram implementados métodos das diferenças finitas e dos elementos finitos para resolver numericamente diversos problemas de obstáculo. A dissertação avaliou a precisão e a convergência dos métodos propostos na solução dos problemas, mostrando a sua robustez e aplicabilidade em cenários práticos.

# Palavras Chave

Problema do Obstáculo; Elementos Finitas; Diferenças Finitas; Membrana Elástica; Modelo de Black-Scholes.



# Conteúdo

<b>1</b>	<b>Introdução</b>	<b>1</b>
<b>2</b>	<b>Definição e análise matemática do problema do obstáculo</b>	<b>5</b>
2.1	Problema do Obstáculo Clássico: Membrana Elástica . . . . .	5
2.1.1	Energia Potencial da Membrana . . . . .	5
2.1.2	Conjunto de soluções admissíveis e problema de minimização . . . . .	6
2.1.3	Problema Variacional . . . . .	7
2.1.4	Existência de solução . . . . .	9
2.1.5	Problema de Valor na Fronteira - Formulação Forte (ou clássica) . . . . .	11
2.1.6	Regularidade de solução . . . . .	14
2.1.7	Problema do Obstáculo: Formulação com Multiplicadores de Lagrange . . . . .	15
2.2	Problemas de Obstáculo com Aplicações à Matemática Financeira . . . . .	16
2.2.1	Noções e conceitos básicos na matemática financeira . . . . .	17
2.2.2	Problema de valor final: Maturidade . . . . .	19
2.2.2.1	Opção de compra na maturidade . . . . .	19
2.2.2.2	Opção de venda na maturidade . . . . .	19
2.2.3	Modelo de Black-Scholes . . . . .	20
2.2.3.1	Modelo de Black-Scholes para Opções Europeias . . . . .	20
2.2.3.2	Modelo de Black-Scholes para Opções Americanas . . . . .	21
2.2.4	Formulação fraca da inequação de Black-Scholes para American Options . . . . .	24
2.2.5	Existência, unicidade e regularidade . . . . .	24
<b>3</b>	<b>Métodos numéricos de discretização</b>	<b>25</b>
3.1	Método das Diferenças Finitas . . . . .	25
3.1.1	Diferenças finitas a 1D . . . . .	25
3.1.1.1	Diferenças FInitas em 2D . . . . .	27
3.2	Método dos Elementos Finitos . . . . .	29
3.2.1	Discretização geométrica do domínio: Mesh . . . . .	29
3.2.2	Elementos Finitos - Tripleto . . . . .	30

3.2.3	Formulação Variacional Discreta - Aproximação de Galerkin . . . . .	31
3.2.4	Elementos de Lagrange . . . . .	35
3.2.5	Elementos a 1D . . . . .	35
3.2.6	Elementos a 2D: Triângulo . . . . .	36
3.2.6.1	Elementos de Lagrange - 2D: Triângulo . . . . .	37
3.2.6.2	Lagrange de 1º Grau - Lineares . . . . .	38
3.2.7	Integração Numérica . . . . .	38
3.2.7.1	Integração de Gauss-Legendre a 2D: Domínio Triangular . . . . .	38
3.2.7.2	Transformação do triângulo de referência para o quadrado de referência . . . . .	39
3.2.7.3	Algumas considerações sobre $\mathcal{P}_1$ . . . . .	41
<b>4</b>	<b>Resolução Numérica de Problemas do Obstáculo</b> . . . . .	<b>45</b>
4.1	Problema de Complementariedade Linear . . . . .	45
4.1.1	Algoritmo iterativo: Semi-smooth Newton Method (SSNM) . . . . .	46
4.2	Problema do Obstáculo Elíptico: Membrana Elástica . . . . .	47
4.2.1	Discretização com Diferenças finitas . . . . .	48
4.2.2	Discretização com Elementos Finitos . . . . .	50
4.2.2.1	Formulação mista . . . . .	51
4.2.2.2	Formulação com estabilização . . . . .	56
4.3	Problema Parabólico: Problema de valor final - Black Scholes para American Options . . . . .	60
4.3.1	Aproximação Finita do Domínio . . . . .	60
4.3.2	Método das Diferenças Finitas para Black Scholes . . . . .	61
4.3.3	Esquema Explícito (Puro) . . . . .	64
4.3.4	Esquema Implícito (Puro) . . . . .	64
4.3.5	Esquema- $\theta$ . . . . .	65
4.3.5.1	Esquema- $\theta$ para Opções Europeias . . . . .	65
4.3.5.2	Esquema- $\theta$ para American Black-Scholes . . . . .	66
4.3.6	Método dos Elementos Finitos para Black-Scholes . . . . .	66
4.3.6.1	Formulação fraca . . . . .	66
4.3.6.2	Malhagem do domínio: Mesh . . . . .	67
4.3.6.3	Espaço de dimensão finita . . . . .	67
4.3.6.4	Sistema Matricial . . . . .	68
4.3.7	Diferenças finitas no tempo: Esquema- $\theta$ . . . . .	70
4.3.7.1	Esquema- $\theta$ para European Black-Scholes . . . . .	72
4.3.7.2	Esquema- $\theta$ para American Black-Scholes . . . . .	72

4.3.8 Consistência, Estabilidade e Convergência de FDM e FEM com o esquema- $\theta$ no tempo . . . . .	72
<b>5 Resolução Numérica e resultados</b>	<b>73</b>
5.1 Problema Elíptico: Problema Estacionário - Membrana Elástica . . . . .	73
5.1.1 Problema prático: Domínio Circular . . . . .	73
5.1.2 Solução Analítica . . . . .	73
5.1.3 Implementação dos Algoritmos . . . . .	75
5.1.3.1 Discretização com diferenças finitas e implementação do SSNM para o LCP	76
5.1.3.2 Formulação com multiplicadores de Lagrange: elementos finitos mistos .	76
5.1.3.3 Formulação com multiplicadores de lagrange: elementos finitos com método de estabilização . . . . .	77
5.1.4 Análise e comparação dos resultados . . . . .	78
5.2 Modelo de Black-Scholes . . . . .	79
5.2.1 Opções Europeias . . . . .	79
5.2.2 Opções Americanas . . . . .	82
<b>6 Conclusão e trabalho futuro</b>	<b>85</b>
<b>A Algoritmos</b>	<b>88</b>

**x**

# Listas de Figuras

2.1 Curva $S_f$ , de separação das duas regiões para Opções de compra . . . . .	23
2.2 Curva $S_f$ , de separação das duas regiões para Opções de venda . . . . .	23
5.1 Função obstáculo $g(x, y)$ . . . . .	74
5.2 Solução $u(x, y)$ . . . . .	74
5.3 Funções $u(x, y)$ e $g(x, y)$ sobrepostas. . . . .	75
5.4 Diferentes domínios do problema. . . . .	75
5.5 Grelha da discretização do domínio circular $\Omega$ , para diferenças finitas . . . . .	75
5.6 Malha triangular do domínio circular $\Omega$ , usado para elementos finitos . . . . .	75
5.7 Valores de $\lambda_h$ obtidos para cada elemento com o Primal-dual active set para o método misto com $h = 0.2$ . . . . .	77
5.8 Valores de $\lambda$ obtidos para cada elemento com o Primal-dual active set para o método estabilizado com $h = 0.2$ . . . . .	78
5.9 Superfície $V(S, t)$ para opções de venda Europeias, obtida com: $T = 1, r = 0.06, \sigma = 0.3, K = 10$ . . . . .	80
5.10 $V(S, t)$ para opções de venda Europeias, obtida com $T = 1, r = 0.06, \sigma = 0.3, K = 10$ , avaliada em três momentos diferentes: $t = T, t = T/2$ e $t = 0$ . . . . .	80
5.11 Convergência do erro na escala logarítmica . . . . .	81
5.12 Superfície $V(S, t)$ para opções de venda Americanas, obtida com: $T = 1, r = 0.06, \sigma = 0.3, K = 10$ . . . . .	82
5.13 $V(S, t)$ para opções de venda Americanas, obtida com $T = 1, r = 0.06, \sigma = 0.3, K = 10$ , avaliada em três momentos diferentes: $t = T, t = T/2$ e $t = 0$ . . . . .	82
5.14 Curva de separação $S_f(t)$ da região de paragem (stop) e da região de continuação (hold), para uma opção americana de venda obtida com os parâmetros: $r = 0.06, \sigma = 0.3, T = 1$ e $K = 10, S_{max} = 15$ . . . . .	83



# **Lista de Tabelas**

5.1	Discretização com diferenças finitas, e resolução do LCP com SSNM, com tolerância $TOL = 10^{-12}$ . . . . .	76
5.2	Resultados numéricos para o método misto, com $TOL = 10^{-32}$ . . . . .	77
5.3	Resultados numéricos para o método estabilizado, com $TOL = 10^{-32}$ . . . . .	78
5.4	Erros para opções de venda Europeias, com FDM no espaço e esquema implícito no tempo, em $V(S_0, 0)$ , com: $T = 1, r = 0.06, \sigma = 0.3, K = 10, S_0 = 8, S_{max} = 200$ . . . . .	80
5.5	Erros para opções de venda Europeias, com FEM no espaço e esquema implícito no tempo, em $V(S_0, 0)$ , com: $T = 1, r = 0.06, \sigma = 0.3, K = 10, S_0 = 8, S_{max} = 200$ . . . . .	80
5.6	FDM com $h_t = 0.1, h_S = 0.015625$ . . . . .	81
5.7	FEM com $h_t = 0.1, h_S = 0.015625$ . . . . .	81
5.8	Valor $V(S_0, 0)$ para diferentes $S_0$ de uma opção de venda Americana, com FDM no espaço e esquema implícito no tempo, e SSNM para cada iteração, com: $T = 1, r = 0.06, \sigma = 0.3, K = 10, S_{max} = 200, TOL = 10^{-12}, h_S = 0.5, h_t = 0.1$ (Algoritmo A.6). . . . .	83
5.9	Valor $V(S_0, 0)$ para diferentes $S_0$ de uma opção de venda Americana, com FEM no espaço e esquema implícito no tempo, e SSNM para cada iteração, com: $T = 1, r = 0.06, \sigma = 0.3, K = 10, S_{max} = 200, TOL = 10^{-12}, h_S = 0.5, h_t = 0.1$ (Algoritmo A.7). . . . .	83
5.10	Erros relativos entre os dois métodos implementados para diferentes valores de $S_0$ , $e_{rela} =  V_{FEM}(S_0, 0) - V_{FDM}(S_0, 0) $ , para valores $V(S_0, 0)$ das tabelas 5.8 e 5.9 . . . . .	84



# **Lista de Algoritmos**

A.1	Iterative Solution Algorithm: Semi-smooth Newton Method (SSNM) for Linear Complementarity Problem (LCP) . . . . .	88
A.2	Iterative Solution Algorithm: Primal-dual active set for Mixed Method . . . . .	89
A.3	Iterative Solution Algorithm: Primal-dual active set for Stabilized FEM . . . . .	90
A.4	Diferenças finitas no espaço e esquema- $\theta$ no tempo para Opções Europeias . . . . .	90
A.5	Elementos finitos no espaço e esquema- $\theta$ no tempo para Opções Europeias . . . . .	91
A.6	Diferenças finitas no espaço, esquema- $\theta$ no tempo, com SSNM para o problema de complementaridade, para Opções Americanas . . . . .	91
A.7	Elementos finitos no espaço, esquema- $\theta$ no tempo, com SSNM para o problema de complementaridade, para Opções Americanas . . . . .	92



# Capítulo 1

## Introdução

As **equações diferenciais parciais** (EDPs), do inglês *Partial Differential Equations* (PDEs), são ferramentas essenciais para modelar matematicamente uma variedade de fenômenos físicos e naturais. Entre elas, destacam-se os problemas elípticos e os problemas parabólicos, que representam classes distintas de comportamentos.

As equações parciais elípticas normalmente descrevem situações de equilíbrio (estacionárias), portanto, sem depender de como a solução evolui no tempo, como problemas de estática e problemas de valor de contorno. Estas equações são caracterizadas com um operador diferencial elíptico de segunda ordem do tipo:

$$\mathcal{L}_{\text{ele}}(u) := Lu$$

Por outro lado os problemas parabólicos, modelando evolução temporal, como a difusão (de calor, por exemplo), e outros variadíssimos tipos de problemas. Eles são caracterizados por um operador diferencial de segunda ordem que inclui uma derivada no tempo:

$$\mathcal{L}_{\text{par}}(u) := \frac{\partial u}{\partial t} + Lu$$

onde para ambos os casos,  $L$  denota um operador diferencial parcial de segunda ordem no espaço.

Em ambos os casos assumiremos a presença de um conjunto aberto e limitado  $\Omega \subseteq \mathbb{R}^n$ , com fronteira  $\partial\Omega$  regular (com regularidade significa que ser de  $C^2$  é suficiente, mas podemos considerar  $C^\infty$ ).

No caso parabólico, temos uma variável tempo  $t$ , consideramos  $T > 0$ , definindo assim todo o domínio  $U_T = \Omega \times (0, T)$ .

Também para ambos os casos, a letra  $L$  denota para cada tempo  $t$  um operador diferencial parcial de segunda ordem, tendo ou a forma de divergência (em inglês *divergence form*):

$$\begin{aligned} Lu &= -\nabla_x \cdot (A(x, t)\nabla_x u) + b^T(x, t)\nabla_x u(x, t) + c(x, t)u(x, t) = \\ &= -\sum_{i,j=1}^N \frac{\partial}{\partial x_i} \left\{ a_{ij}(x, t) \frac{\partial u}{\partial x_j} \right\} + \sum_{i=1}^N b_i(x, t) \frac{\partial u}{\partial x_i} + c(x, t)u \end{aligned} \tag{1.1}$$

ou então a forma de não divergência (em inglês *nondivergence form*):

$$\begin{aligned} Lu &= (1)^T [A(x, t) \odot \nabla(\nabla u)](1) + b^T(x, t) \nabla_x u(x, t) + c(x, t)u(x, t) = \\ &= - \sum_{i,j=1}^N a_{ij}(x, t) \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^N b_i(x, t) \frac{\partial u}{\partial x_i} + c(x, t)u \end{aligned} \quad (1.2)$$

o vetor (1) é um vetor coluna de dimensão N, só de valores '1', e  $\odot$  é o produto matricial de Hadamard, entrada a entrada.  $A(x, t)$  é uma matriz de dimensão  $N \times N$ , e  $b(x, t)$  um vetor de dimensão  $N$ :

$$A(x, t) = \begin{bmatrix} a_{11}(x, t) & \dots & a_{1N}(x, t) \\ \vdots & \dots & \vdots \\ a_{N1}(x, t) & \dots & a_{NN}(x, t) \end{bmatrix}, \quad b(x, t) = \begin{bmatrix} b_1(x, t) \\ \vdots \\ b_N(x, t) \end{bmatrix}$$

Doravante assumiremos sempre as seguintes condições:

- A matriz  $A$ , será simétrica, ou seja  $A = A^T$ ;
- Vamos assumir sempre que  $a_{ij}, b_i, c$  são limitadas, ou por outras palavras:  $a_{ij}, b_i, c \in L^\infty(U_T)$ , sabendo que  $f \in L^\infty(\Omega) \implies \exists k \in \mathbb{R}^+ : |f| < k \ \forall (x, t) \in U_T$ ;
- O operador  $L$ , está na forma de divergência, sendo  $\mathcal{L}_{ele}$  uniformemente elíptico, e  $\mathcal{L}_{par}$  uniformemente parabólico para todo  $\Omega_T$ . Para tal a matriz  $A$  têm de ser semidefinida positiva, ou seja:

$$\exists \theta > 0 : \xi^T A(x, t) \xi = \sum_{i,j=1}^N a_{ij} \xi_i \xi_j > \theta \|\xi\|^2$$

para todo  $(x, t) \in U_T$ , e todo  $\xi \in \mathbb{R}^N \setminus \{0\}$ .

Para mais detalhes sobre os operadores diferenciais consultar [1].

Mesmo com a distinção clara entre problemas elípticos e parabólicos, surgem desafios comuns na análise e na resolução dessas EDPs. Um desses desafios são os **problemas de fronteira livre** (FBPs, sigla do inglês significa *Free Boundary Problems*) são uma classe de EDPs que envolvem fronteiras desconhecidas. Ao contrário das EDPs clássicas, onde as fronteiras são especificadas *a priori*, os FBPs exigem a determinação tanto da solução quanto da própria fronteira como parte do processo de solução, tornando este problema significativamente mais desafiador para analisar e resolver.

Um dos mais importantes FBPs, é o **Problema do Obstáculo**. Nestes problemas a fronteira desconhecida não é totalmente livre a ponto de assumir qualquer forma; pois, dado um domínio  $U_T$  do problema, onde a solução  $u \in U_T$  estará definida, teremos a presença de um *obstáculo*, que muitas vezes é escrito como uma função  $g \in U_T$ , e com ele é imposto restrições/barreiras na nossa solução

final (por exemplo na forma de desigualdades, como  $u \geq g$ , refletindo a limitação dentro do domínio, impedindo a solução de ultrapassar determinados valores).

Contudo, muitas vezes, o obstáculo e a solução não são independentes; elas interagem por meio de uma *condição complementar*. Essa condição garante que a *dinâmica* da solução seja direcionada para o obstáculo, apenas em regiões onde a solução atinge o limite do obstáculo ( $u = g$ ), criando uma região de contacto, que designaremos por  $U_C \subset U_T$ . Assim a posição no domínio, onde estas situações limite ocorrem, é desconhecida a priori, o que cria uma "fronteira livre",  $\partial U_C$ , desconhecida, tal como mencionado. Assim, o domínio  $U_T$  é dividido em:

**Domínio de contacto:**  $U_C = \{x \in U_T : u = g\}$

**Domínio de livre contacto:**  $U_f = \{x \in U_T : u > g\}$ , onde temos  $\mathcal{L}u = 0$

**Fronteira livre:**  $\Gamma = \partial U_C \cap \partial U_f$ .

A física e a matemática estão intrinsecamente ligadas, muitas vezes são encontradas interseções fascinantes que iluminam problemas complexos, cruzando estes campos. Um desses problemas é o "*Problema do Obstáculo Clássico: Membrana Elástica*", este problema é frequentemente utilizado como um exemplo motivador neste tipo de matéria, tendo sido intensamente estudado ao longo das últimas décadas. O estudo destes problemas requer técnicas analíticas sofisticadas, como métodos variacionais. E sendo este um dos mais simples, em termos de desigualdades variacionais, a sua escolha torna-se natural. E será, portanto usado como exemplo motivador para o estudo do Problema do Obstáculo associado a operadores elípticos, sendo este um problema estático.

Mais uma vez a matemática e, em específico, o nosso problema do obstáculo, encontra mais uma intersecção com o mundo físico real e, sem nenhum espanto, é no campo das finanças que essa intersecção acontece. O problema, em concreto, é derivado do **modelo de Black-Scholes**, para **Opções Americanas**, tratando-se de **problemas de evolução**, problemas cuja solução decorrerá não apenas no espaço estático, mas sim ao longo do tempo, adicionando uma nova variável ao nosso problema. Este é um problema de finanças com uma aplicação prática de elevada importância, tornando-a uma escolha atraente para este estudo. Como se trata de um problema uni-dimensional no espaço e com uma variável no tempo, fornece uma estrutura simples, tornando o modelo básico e elegante. Tornando a sua escolha extremamente acessível, dentro deste contexto. E será este o problema associado ao **operador parabólico** que motivará o seu estudo.

No entanto, a resolução analítica dessas equações é frequentemente inviável devido à complexidade das condições iniciais e de contorno associadas, ou a outros motivos diversos. Este é particularmente o caso no contexto do problema do obstáculo, onde a equação diferencial é combinada com restrições

adicionais que tornam a solução analítica ainda mais desafiadora. Os métodos numéricos emergem como ferramentas indispensáveis. Eles fornecem meios eficientes e flexíveis para obter aproximações precisas das soluções de EDPs em situações onde métodos analíticos falham. Técnicas como o Método dos Elementos Finitos (FEM), o Método das Diferenças Finitas (FDM) são amplamente utilizadas para discretizar e resolver EDPs numericamente. Essas abordagens permitem a consideração de geometria complexa, condições de contorno variadas e não-linearidades que são típicas no problema do obstáculo.

Assim, a utilização de métodos numéricos na resolução de EDPs associadas ao problema do obstáculo não é apenas uma questão de conveniência, mas uma necessidade para avançar no entendimento e na aplicação de modelos.

No capítulo 2, define-se e analisa-se matematicamente o problema do obstáculo, no caso elíptico aplicado à membrana elástica, e no caso parabólico com aplicações na matemática financeira para opções europeias e americanas; discutindo as suas formulações variacionais e a regularidade das soluções. No capítulo 3 abordam-se os métodos de discretização numérica, nos quais se inclui o método das diferenças finitas e o método dos elementos finitos, os quais serão usados na resolução numérica destes problemas. Posteriormente, no capítulo 4 aplicamos os métodos de discretização aos nossos problemas do obstáculo, derivamos métodos e algoritmos iterativos para a obtenção da sua solução numérica. Já no capítulo 5 implementam-se os métodos estudados no capítulo anterior a problemas concretos, quer para o caso parabólico quer para o caso elíptico, e finalmente tiram-se algumas conclusões sobre a sua eficácia.

# Capítulo 2

## Definição e análise matemática do problema do obstáculo

### 2.1 Problema do Obstáculo Clássico: Membrana Elástica

O problema do obstáculo clássico é um problema que provém da teoria da elasticidade (linear) clássica. Descreve como uma membrana elástica, sujeita a uma força vertical  $f$ , se situa acima de um obstáculo  $g$ . Portanto pretende-se determinar a posição de equilíbrio (deslocamento vertical), da membrana, denotada por  $u$ .

Na teoria clássica da elasticidade, esta membrana é uma placa fina que não oferece nenhuma resistência à flexão, atuando apenas em tensão (interna à membrana). Assume-se que é uma membrana homogénea ocupando um domínio  $\Omega \subset \mathbb{R}^n$ , da reta  $Ox$  em 1D, ou do plano  $Oxy$  em 2D. Supõe-se que esteja igualmente esticada em todas as direções por uma tensão uniforme, e carregada por uma força  $f : \Omega \rightarrow \mathbb{R}$ , normalmente, distribuída uniformemente (que, por exemplo, podemos pensar como sendo a força gravítica).

A posição da membrana na fronteira  $\partial\Omega$  está fixa, mas não é necessariamente constante, portanto impõe-se:  $u = h$  sobre  $\partial\Omega$ .

Considere agora o problema de encontrar a posição de equilíbrio da membrana, ou seja  $u$ , restringida de modo a ser obrigada a ficar por cima de um objeto/corpo ao qual chamaremos obstáculo e representaremos por uma função  $g : \Omega \rightarrow \mathbb{R}$ , definida em  $\Omega$ , que verifica  $g \leq u$  em  $\Omega$  e verifica  $g \leq h = u|_{\partial\Omega}$  em  $\partial\Omega$  (cf. [2], [3]).

#### 2.1.1 Energia Potencial da Membrana

Vamos assumir que a energia potencial de deformação da membrana é proporcional ao aumento de área da sua superfície. A área da membrana é dada por:

$$A(u) = \int_{\Omega} \sqrt{1 + \|\nabla u\|_{l^2}^2} dx$$

onde para  $x \in \mathbb{R}^n$  a norma  $\|x\|_{l^p} = (\sum_{i=1}^n |x_i|^p)^{1/p}$ . Para pequenas deformações da membrana, temos  $\sqrt{1 + \|\nabla u\|_{l^2}^2} \approx 1 + \frac{1}{2}\|\nabla u\|_{l^2}^2$ . Assim:

$$A(u) = \int_{\Omega} \sqrt{1 + \|\nabla u\|_{l^2}^2} dx \approx \int_{\Omega} 1 + \frac{1}{2}\|\nabla u\|_{l^2}^2 dx$$

Portanto, a variação da área da membrana é  $\int_{\Omega} \|\nabla u\|_{l^2}^2 dx$ , pelo que a energia potencial de deformação da membrana tem a seguinte expressão:

$$D(u) = \frac{\lambda}{2} \int_{\Omega} |\nabla u|^2 dx$$

sendo  $\lambda > 0$  é uma constante que depende das propriedades de elasticidade da membrana. Mas por simplicidade, e sem perda de generalidade, definimos  $\lambda = 1$ . O trabalho realizado pelas forças externas durante o deslocamento real é

$$F(u) = \int_{\Omega} f u dx$$

então a energia potencial final  $E = D - F$  irá ser:

$$E(u) = \frac{1}{2} \int_{\Omega} |\nabla u|^2 dx - \int_{\Omega} f u dx. \quad (2.1)$$

### 2.1.2 Conjunto de soluções admissíveis e problema de minimização

Conforme [4] e [5], introduzimos agora o conjunto de soluções admissíveis ao problema:

$$K = \{v \in V : v(x) \geq g(x), x \in \Omega\} \quad (2.2)$$

onde  $V$  é um espaço vetorial de funções com energia finita, sendo um espaço de Hilbert munido do produto interno  $(\cdot, \cdot)_V$ , e seja associada a norma  $\|\cdot\|_V$ , induzida por esse produto interno. De modo a simplificar o nosso problema, vamos considerar que a fronteira, da membrana elástica, é mantida como  $u = 0$  em  $\partial\Omega$ . E portanto teremos

$$V = \{v \in H^1(\Omega) : v = 0, x \in \partial\Omega\} = H_0^1(\Omega) \quad (2.3)$$

Assumindo que o subconjunto  $K \subset V$  é fechado, convexo e não vazio, conforme [6], seja  $a : V \times V \rightarrow \mathbb{R}$  um operador bilinear simétrico, elítico e contínuo dado por:

$$a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v dx \quad (2.4)$$

e  $L : V \rightarrow \mathbb{R}$  um operador (funcional) linear dado por:

$$L(v) = \int_{\Omega} f v dx \quad (2.5)$$

recrevemos o funcional  $E : V \rightarrow \mathbb{R}$  como:

$$E(u) = \frac{1}{2} \int_{\Omega} |\nabla u|^2 dx - \int_{\Omega} f u dx = \frac{1}{2} a(u, u) - L(u) \quad (2.6)$$

num subconjunto convexo, fechado e não vazio  $K \subset V$ .

Na posição de equilíbrio aplica-se o *princípio da energia potencial mínima*, e é declarado na forma:

$$\exists u \in K : E(u) \leq E(v), \forall v \in K \quad (2.7)$$

que nos permite finalmente formular o **Problema de minimização**:

$$\begin{cases} \text{Encontrar } u \in K \text{ tal que:} \\ E(u) := \min_{v \in K} E(v). \end{cases} \quad (2.8)$$

Observe que  $K$  não é um subespaço, mas um subconjunto convexo do espaço de Hilbert  $H^1(\Omega)$ . Isto torna o problema não linear, uma vez que o mínimo não é procurado num espaço linear. Consequentemente, o problema de minimização é caracterizado por uma desigualdade variacional em vez de uma igualdade, como vamos ver.

### 2.1.3 Problema Variacional

Como veremos, uma outra maneira de formular o problema do obstáculo é com uma inequação variacional. O **problema variacional** define-se como:

$$\begin{cases} \text{Encontrar } u \in K \text{ tal que:} \\ a(u, v - u) \geq L(v - u), \forall v \in K \end{cases} \quad (2.9)$$

Seja  $u \in K$  solução para o problema de minimização (2.8), e seja  $v \in K$  é um elemento arbitrário em  $K$ , como  $K$  é um conjunto convexo então a sua combinação convexa pertence também a  $K$ , ou seja:

$$vt + u(1 - t) = u + t(v - u) \in K, \forall t \in [0, 1].$$

Definimos o funcional de energia  $e(t) = E(u + t(v - u))$ , com  $t \in [0, 1]$ , tendo um mínimo em  $t = 0$  dado por  $e(0) = E(u)$ . Portanto  $e(0) \leq e(t)$  para qualquer  $t \in [0, 1]$ , e podemos ver que é válido que:

$$e'(0) = \left. \frac{\partial e(t)}{\partial t} \right|_{t=0} = \left. \frac{\partial E(u + t(v - u))}{\partial t} \right|_{t=0} \geq 0$$

portanto, calculando a derivada:

$$\begin{aligned}\frac{\partial E(vt + u(1-t))}{\partial t} &= \frac{\partial}{\partial t} \left\{ \frac{1}{2}a(u + t(v-u), u + t(v-u)) - (f, u + t(v-u)) \right\} \\ &= \frac{\partial}{\partial t} \left\{ \frac{1}{2}a(u, u) + ta(u, v-u) + t^2 \frac{1}{2}a(v-u, v-u) - (f, u) - t(f, v-u) \right\} \\ &= a(u, v-u) + ta(v-u, v-u) - (f, v-u)\end{aligned}$$

substituindo na fórmula por  $t = 0$  (onde ocorre o mínimo), obtemos:

$$\begin{aligned}\frac{\partial E(u + t(v-u))}{\partial t} \Big|_{t=0} \geq 0 &\Leftrightarrow a(u, v-u) - (f, v-u) \geq 0 \Leftrightarrow \\ &\Leftrightarrow a(u, v-u) \geq (f, v-u), \quad \forall v \in K\end{aligned}$$

o que mostra que a solução do problema de minimização é uma solução para o problema variacional.

Para mostrar a equivalência entre os dois problemas, falta apenas mostrar que uma solução do Problema Variacional é também solução do Problema de Minimização.

Considere-se agora que  $u \in K$  é solução do problema variacional, e que se tem para  $\forall t \in [0, 1]$  e para  $\forall v \in K$ :

$$\begin{aligned}E(u + (v-u)t) &= \frac{1}{2}a(u + t(v-u), u + t(v-u)) - (f, u + t(v-u)) \\ &= \frac{1}{2}a(u, u) + ta(u, v-u) + t^2 \frac{1}{2}a(v-u, v-u) - (f, u) - t(f, v-u)\end{aligned}$$

Escolhendo  $t = 1$ , tem-se:

$$\begin{aligned}E(v) &= \frac{1}{2}a(u, u) + a(u, v-u) + \frac{1}{2}a(v-u, v-u) - (f, u) - (f, v-u) = \\ &= \frac{1}{2}a(u, u) - (f, u) + \frac{1}{2}a(v-u, v-u) - (f, v-u) + a(u, v-u) = \\ &= E(u) + \frac{1}{2}a(v-u, v-u) - (f, v-u) + a(u, v-u)\end{aligned}$$

Tendo em conta que a desigualdade do Problema variacional, é dada por  $a(u, v-u) \geq L(v-u)$ , ficamos com:

$$E(v) \geq E(u) + \frac{1}{2}a(v-u, v-u) - (f, v-u) + (f, v-u) \Leftrightarrow E(v) \geq E(u) + \frac{1}{2}a(v-u, v-u)$$

Sendo  $a(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$  um operador bilinear elíptico, e sendo este também coercivo, existe uma constante  $C > 0$  tal que  $a(v, v) \geq C\|v\|_V^2$  para todo  $v \in V$ , portanto:

$$E(v) \geq E(u) + \frac{C}{2}\|v-u\|^2, \text{ ou seja } E(u) \leq E(v), \quad \forall v \in K$$

E assim mostramos que a solução do problema variacional é a solução do problema de minimização, e

vice-versa.

#### 2.1.4 Existência de solução

Para provar a existência de solução deste problema, precisamos de alguns teoremas fundamentais.

**Teorema 1** (Teorema da Projecção). *Seja  $K$  um subconjunto convexo, fechado e não-vazio de um espaço de Hilbert  $V$  (ou seja  $K \subset V$ ). Para cada  $w \in V$ , existe um único  $u \in K$  tal que:*

$$\|w - u\| = \min_{v \in K} \|w - v\|. \quad (2.10)$$

Além disso o elemento  $u \in K$  satisfaz:

$$(u - w, u - v) \geq 0, \quad \forall v \in K \quad (2.11)$$

e a equação (2.10) define um operador de projecção  $P_K : V \rightarrow K$  onde  $P_K(w) := u$ .

**Demonstração:** Ver [6], [7] e [8].

**Teorema 2** (Representação de Riesz). *Seja  $V$  um espaço de Hilbert e seja  $a : V \times V \rightarrow \mathbb{R}$  uma forma bilinear contínua. Então existe um único operador linear limitado  $A : V \rightarrow \mathbb{R}$  tal que*

$$a(u, v) = (Au, v), \quad \forall u, v \in V. \quad (2.12)$$

Além disso

$$\|A\| = \|a\| \quad (2.13)$$

onde

$$\|a\| = \sup_{u, v \neq 0} \frac{|a(u, v)|}{\|u\| \|v\|}. \quad (2.14)$$

**Demonstração:** Ver [6], [7] e [8].

**Teorema 3** (Ponto fixo de Banach). *Seja  $X$  um conjunto não vazio e fechado num espaço de Banach  $B$ , e seja  $T : X \rightarrow X$  um operador contrativo em  $X$ , isto é, existe uma constante  $L$ ,  $0 \leq L < 1$ , tal que:*

$$\|T(u) - T(v)\|_B \leq L \|u - v\|_B, \quad \forall u, v \in X. \quad (2.15)$$

Então  $T$  tem precisamente um único ponto fixo  $x$  em  $X$ , isto é

$$\exists^1 x \in X : T(x) = x. \quad (2.16)$$

**Demonstração:** Ver [6], [7] e [8].

De modo a provar o pretendido, a existência de solução, começemos por recordar o problema variacional dado em (2.9), encontrar  $u \in K$  tal que:

$$a(u, v - u) \geq (f, v - u), \quad \forall v \in K. \quad (2.17)$$

Recorrendo ao Teorema da representação de Riesz (Teorema 2), existe um operador linear e limitado  $A$  tal que:

$$(Au, v - u) \geq (f, v - u), \quad \forall v \in K. \quad (2.18)$$

Por outro lado, para todo  $\rho \geq 0$  é verdade que:

$$\begin{aligned} (Au, v - u) &\geq (f, v - u) \\ (Au - f, v - u) &\geq 0 \\ -\rho(Au - f, v - u) &\leq 0 \\ \rho(Au - f, u - v) &\leq 0 \\ (u - u, v - u) + \rho(Au - f, u - v) &\leq 0 \\ (u - (u - \rho[Au - f]), u - v) &\leq 0 \end{aligned}$$

Pelo Teorema da Projeção, notamos que a desigualdade anterior é equivalente a encontrar  $u = P_K(u - \rho[Au - f])$ . Defina-se o operador projeção  $T_\rho : V \rightarrow V$  como

$$T_\rho(u) := P_K(u - \rho[Au - f])$$

Pelas propriedades do operador projeção, é um operador linear e limitado, tem-se:

$$\begin{aligned} \|P_K(u)\|^2 &= (P_K(u), P_K(u)) = (P_K(P_K(u)), u) = (P_K^2(u), u) = (P_K(u), u) \leq \|P_K(u)\| \cdot \|u\| \Leftrightarrow \\ &\Leftrightarrow \|P_K(u)\| \leq \|u\| \end{aligned}$$

Vamos mostrar agora que  $T_\rho$  é contractivo,

$$\begin{aligned} \|T_\rho(u) - T_\rho(v)\|^2 &= \|P_K(u - \rho[Au - f]) - P_K(v - \rho[Av - f])\|^2 = \\ &= \|P_K((u - \rho[Au - f]) - (v - \rho[Av - f]))\|^2 \leq \\ &\leq \|(u - \rho[Au - f]) - (v - \rho[Av - f])\|^2 = \\ &= \left( (u - \rho[Au - f]) - (v - \rho[Av - f]), (u - \rho[Au - f]) - (v - \rho[Av - f]) \right) = \\ &= \left( u - v - \rho A(u - v), u - v - \rho A(u - v) \right) = \end{aligned}$$

$$\begin{aligned}
&= \left( u - v, u - v \right) + \left( \rho A(u - v), \rho A(u - v) \right) + 2 \left( -\rho A(u - v), u - v \right) = \\
&= \|u - v\|^2 + \rho^2 \|A(u - v)\|^2 - 2\rho \left( A(u - v), u - v \right) = \\
&= \|u - v\|^2 + \rho^2 \|A(u - v)\|^2 - 2\rho \cdot a(u - v, u - v) \leq \\
&\leq \|u - v\|^2 + \rho^2 \|A\|^2 \|u - v\|^2 - 2\rho \cdot C \|u - v\|^2 = \\
&\leq (1 + \rho^2 \|A\|^2 - 2\rho \cdot C) \|u - v\|^2
\end{aligned}$$

de onde observamos que  $T_\rho$  será uma contração se

$$1 + \rho^2 \|A\|^2 - 2\rho C < 1 \Leftrightarrow \rho(\rho \|A\|^2 - 2C) < 0$$

escolhendo um  $\rho \in \left]0, \frac{2C}{\|A\|^2}\right[$  para o operador  $T_\rho$ , pelo teorema , admitirá um único ponto fixo e assim existe uma única solução do Problema Variacional.

### 2.1.5 Problema de Valor na Fronteira - Formulação Forte (ou clássica)

Agora, a partir do problema variacional (2.9), vamos formular o **Problema de Valor na Fronteira**. Seja  $u \in K$  a solução do nosso problema, e seja  $v$  uma função arbitrária de  $K$ , satisfazendo a condição de fronteira  $v = h$ . Então claramente, teremos  $w = v - u = 0$  sobre  $\partial\Omega$ , e portanto  $w = v - u \in H_0^1(\Omega)$ .

O problema variacional pode ser visto como encontrar  $u \in K$ :

$$a(u, v - u) \geq L(v - u), \quad \forall v \in K$$

ou, na forma integral:

$$\int_{\Omega} \nabla u \cdot \nabla(v - u) dx \geq \int_{\Omega} f(v - u) dx, \quad \forall v \in K$$

Vamos dividir agora o nosso domínio  $\Omega$ , em **domínio de contacto**  $\Omega_C$  e em **domínio livre de contacto**  $\Omega_f$ , que definimos como:

$$\Omega_C = \{x \in \Omega : u(x) = g(x)\} \tag{2.19}$$

$$\Omega_f = \{x \in \Omega : u(x) > g(x)\}. \tag{2.20}$$

Além disso seja  $\Gamma = \partial\Omega_C \cap \partial\Omega_f$  a fronteira que separa as duas regiões distintas  $\Omega_C$  e  $\Omega_f$ , que será chamada de **fronteira livre**. Note que estes conjuntos não têm necessariamente de ser conexos.

Agora fazendo a divisão do domínio na forma integral:

$$\int_{\Omega_C} \nabla u \cdot \nabla(v - u) dx + \int_{\Omega_f} \nabla u \cdot \nabla(v - u) dx \geq \int_{\Omega} f(v - u) dx, \quad \forall v \in K$$

De modo a obter a formulação forte, supomos que  $u \in C^2(\Omega)$ , para ter suficiente regularidade. Integrando por partes a expressão anterior:

$$-\int_{\Omega_C} \Delta u(v - u) dx + \int_{\Gamma} (\nabla u \cdot \vec{n}_C)(v - u) dx - \int_{\Omega_f} \Delta u(v - u) dx + \int_{\Gamma} (\nabla u \cdot \vec{n}_f)(v - u) dx \geq \int_{\Omega} f(v - u) dx, \quad \forall v \in K \quad (2.21)$$

onde  $\vec{n}_C$  representa o vector normal exterior a  $\Omega_C$  e  $\vec{n}_f$  representa o vector normal exterior a  $\Omega_f$ , e é claro que temos  $\vec{n}_f = -\vec{n}_C$ .

É espectável que a solução não seja de classe  $C^2$  em todo o domínio  $\Omega$ , pelo que não será possível integrar por partes em todo o domínio  $\Omega$ .

- Considere-se agora um novo  $w$  definido em  $\Omega$ , sendo  $w \geq 0$  em  $\Omega$ , suave de suporte compacto no domínio livre de contacto ou seja  $C_0^\infty(\Omega_f)$ , e extendido  $w = 0$  em  $\overline{\Omega_C}$ . Como  $w$  tem um suporte compacto na região onde  $u > g$ , existe uma constante  $c$  tal que

$$u(x) \geq c > g(x), \quad x \in \text{supp}\{w\}.$$

Consequentemente é possível escolher  $\delta > 0$ , de modo a que  $u \pm \delta w \geq g$  e assim  $u \pm \delta w \in K$ .

Usando  $v = u \pm \delta w$  como função teste, e substituindo na desigualdade integral (2.21). Ficamos com:

$$\pm \delta \int_{\Omega_f} (\Delta u - f)w dx \geq 0, \quad \forall w \in C_0^\infty(\Omega_f). \quad (2.22)$$

Visto que  $\delta > 0$ , segue-se que

$$\int_{\Omega_f} (\Delta u - f)w dx = 0, \quad \forall w \in C_0^\infty(\Omega_f) \quad (2.23)$$

de onde sai

$$-\Delta u - f = 0, \quad x \in \Omega_f. \quad (2.24)$$

- Além disso nós sabemos que na região de contacto temos:

$$u = g, \quad \text{em } \Omega_C \quad (2.25)$$

o que nos permite escrever as duas equações anteriores, em apenas uma condição, chamada de

**condição de complementaridade:**

$$(u - g)(-\Delta u - f) = 0, \quad x \in \Omega. \quad (2.26)$$

Além disso, como  $u \in V$  temos  $u = h$  para  $x \in \partial\Omega$ .

- Finalmente, na desigualdade integral, escolhendo uma função teste  $v = u \pm w$ , onde  $\delta > 0$  e  $w \in C^\infty(\Omega_f)$  com  $w = 0$  em  $\Omega_C$  e  $w \geq 0$  em  $\Omega_f$ , e substituindo, ficamos com:

$$\pm\delta \left[ \int_{\Gamma} (\nabla u \cdot \vec{n}_C) w \, dx + \int_{\Gamma} (\nabla u \cdot \vec{n}_f) w \, dx + \int_{\Omega_f} (-\Delta u - f) w \, dx \right] \geq 0, \quad \forall w \in C^\infty(\Omega_f)$$

como  $-\Delta u - f = 0$  em  $\Omega_f$ , simplificando:

$$\pm\delta \int_{\Gamma} (\nabla u \cdot \vec{n}_C + \nabla u \cdot \vec{n}_f) w \, dx \geq 0, \quad \forall w \in C^\infty(\Omega_f)$$

sendo  $\vec{n}_C = -\vec{n}_f$ , dado a definição de derivada direcional (por limite):

$$D_{\vec{n}} u(x) = \nabla u \cdot \vec{n} = \frac{\partial u}{\partial \vec{n}} = \lim_{\epsilon \rightarrow 0} \frac{u(x + \epsilon \vec{n}) - u(x)}{\epsilon}$$

assim a expressão anterior:

$$\begin{aligned} \nabla u \cdot \vec{n}_C &= \lim_{\epsilon \rightarrow 0} \frac{u(x + \epsilon \vec{n}_C) - u(x)}{\epsilon} \Leftrightarrow \\ &\Leftrightarrow \nabla u \cdot (-\vec{n}_f) = \lim_{\epsilon \rightarrow 0} \frac{u(x + \epsilon (-\vec{n}_f)) - u(x)}{\epsilon} \Leftrightarrow \\ &\Leftrightarrow -\nabla u \cdot \vec{n}_f = \lim_{\epsilon \rightarrow 0} \frac{u(x - \epsilon \vec{n}_f) - u(x)}{\epsilon} \Leftrightarrow \\ &\Leftrightarrow \nabla u \cdot \vec{n}_f = -\lim_{\epsilon \rightarrow 0} \frac{u(x - \epsilon \vec{n}_f) - u(x)}{\epsilon} \end{aligned}$$

agora, substituindo pela definição de limite na seguinte expressão:

$$\begin{aligned} \nabla u \cdot \vec{n}_C + \nabla u \cdot \vec{n}_f &= -\lim_{\epsilon \rightarrow 0} \frac{u(x - \epsilon \vec{n}_f) - u(x)}{\epsilon} + \lim_{\epsilon \rightarrow 0} \frac{u(x + \epsilon \vec{n}_f) - u(x)}{\epsilon} = \\ &= \lim_{\epsilon \rightarrow 0} \frac{u(x + \epsilon \vec{n}_f) - u(x - \epsilon \vec{n}_f)}{\epsilon}. \end{aligned}$$

Define-se o salto da derivada direcional em  $x \in \Gamma$  em ordem à normal como:

$$\left[ \left[ \frac{\partial u}{\partial \vec{n}} \right] \right](x) = \lim_{\epsilon \rightarrow 0} \frac{u(x + \epsilon \vec{n}) - u(x - \epsilon \vec{n})}{\epsilon} \quad (2.27)$$

e portanto obtém-se

$$\pm \delta \int_{\Gamma} \left[ \left[ \frac{\partial u}{\partial \vec{n}} \right] \right] w d\sigma \geq 0, \quad \forall w \in \mathcal{C}^{\infty}(\Omega_f) \quad (2.28)$$

como neste caso  $w \geq 0$  em  $\Omega$ , e  $w = 0$  em  $\Omega_C$ , e  $\Gamma \subset \Omega$  então  $w \geq 0$  em  $x \in \Gamma$ , portanto:

$$\pm \delta \left[ \left[ \frac{\partial u}{\partial \vec{n}} \right] \right] \geq 0, \quad x \in \Gamma$$

então, como  $\delta > 0$ , conclui-se que

$$\left[ \left[ \frac{\partial u}{\partial \vec{n}} \right] \right] = 0, \quad x \in \Gamma.$$

Como  $u = g$  em  $\Omega_C$ , e  $u > g$  em  $\Omega_f$  (assume-se que a fronteira livre é suficientemente suave), a condição  $u = g$  em  $x \in \Gamma$  é automaticamente satisfeita.

Estão assim reunidas as condições para definir o **Problema de valor na fronteira**: encontrar  $u \in H^2(\Omega) \cap C(\bar{\Omega})$ , com  $f \in L^2(\Omega)$  e  $g \in H^1(\Omega) \cap C(\bar{\Omega})$ , tal que juntando toda esta informação nos permite rescrever o problema numa *formulação forte* (ou clássica):

$$\begin{cases} -\Delta u - f \geq 0 & , \text{em } \Omega \\ u \geq g & , \text{em } \Omega \\ (u - g)(-\Delta u - f) = 0 & , \text{em } \Omega \\ u = 0 & , \text{sobre } \partial\Omega \\ u = g & , \text{sobre } \Gamma \\ \left[ \left[ \frac{\partial u}{\partial \vec{n}} \right] \right] = 0 & , \text{sobre } \Gamma. \end{cases} \quad (2.29)$$

A terceira equação em (2.29) é a condição complementar que garante que uma das duas primeiras inequações tem de ser uma igualdade. O que nos permite rescrever este problema através de um **problema de complementaridade**:

$$\begin{cases} \min \{-\Delta u - f, u - g\} = 0 & , x \in \Omega \\ u = 0 & , x \in \partial\Omega. \end{cases} \quad (2.30)$$

## 2.1.6 Regularidade de solução

No sistema (2.29) as suas duas últimas equações reforçam a continuidade da solução  $u$  e da sua derivada normal ao longo da fronteira livre  $\Gamma$ . Como  $-\Delta u - f = 0$  em  $x \in \Omega_f$ , e  $u = g$  em  $x \in \Omega_C$ , é possível ver que as segundas derivadas de  $u$  podem dar um salto ao longo da fronteira livre, pelo que não podemos garantir que as segundas derivadas de  $u$  serão contínuas. A solução, em geral, não será mais regular que  $C^{1,1}(\Omega)$ , i.e., as suas primeiras derivadas são Lipschitz contínuas (ou Lipschitzianas). E a solução pode pertencer a  $H^2(\Omega)$  mas não  $H^3(\Omega)$ , sendo que a regularidade estará sempre dependente da regularidade do obstáculo e das condições de contorno.

### 2.1.7 Problema do Obstáculo: Formulação com Multiplicadores de Lagrange

Considerando de novo o conjunto de solução:

$$K = \{v \in H_0^1(\Omega) : v \geq g, \text{ q.t.p. em } \Omega\}. \quad (2.31)$$

Seguindo [9] e [10], introduzimos um multiplicador de Lagrange não-negativo, sendo ele uma função  $\lambda : \Omega \rightarrow \mathbb{R}$ , que nos permite reescrever o problema do obstáculo como:

$$\begin{cases} \Delta u - \lambda = f & , \text{em } \Omega \\ u - g \geq 0 & , \text{em } \Omega \\ \lambda \geq 0 & , \text{em } \Omega \\ (u - g)\lambda = 0 & , \text{em } \Omega \\ u = 0 & , \text{sobre } \partial\Omega. \end{cases} \quad (2.32)$$

O multiplicador de Lagrange está no espaço dual de  $V = H_0^1(\Omega)$ , ou seja:

$$\lambda \in Q = H^{-1}(\Omega) = V'$$

com norma:

$$\|\xi\|_{H^{-1}(\Omega)} = \sup_{v \in H_0^1(\Omega)} \frac{\langle v, \xi \rangle_{H_0^1(\Omega) \times H^{-1}(\Omega)}}{\|v\|_{H_0^1(\Omega)}}$$

onde  $\langle \cdot, \cdot \rangle_{H_0^1(\Omega) \times H^{-1}(\Omega)} : V \times Q \rightarrow \mathbb{R}$  é o produto interno do dual, ou só produto dual.

De forma a obter a formulação fraca (formulação variacional), começemos por considerar a igualdade  $\Delta u - \lambda = f$ , multipliquemos por uma função teste e integremos por partes, obtendo:

$$(\nabla u, \nabla v) - \langle v, \lambda \rangle = (f, v) , \forall v \in V$$

considerando a inequação  $u - g \geq 0$ , multipliquemos por  $\mu \in \Lambda$  e integremos, obtendo:

$$\langle u - g, \mu \rangle \geq 0 , \forall \mu \in \Lambda$$

onde

$$\Lambda = \{\mu \in Q : \langle v, \mu \rangle \geq 0 , \forall v \in V, v \geq 0 , \text{ q.t.p. em } \Omega\} \quad (2.33)$$

e no caso de  $(u - g)\lambda = 0$ , integremos apenas em  $\Omega$ , obtendo:

$$\langle u - g, \lambda \rangle = 0.$$

Colocando todas elas num sistema obtemos:

$$\begin{cases} (\nabla u, \nabla v) - \langle v, \lambda \rangle = (f, v) & , \forall v \in V \\ \langle u - g, \mu \rangle \geq 0 & , \forall \mu \in \Lambda \\ \langle u - g, \lambda \rangle = 0. \end{cases} \quad (2.34)$$

Assim, subtraindo a 3<sup>a</sup> equação na 2<sup>a</sup> desigualdade do sistema (2.34), a formulação variacional passa a ser: encontrar o par  $(u, \lambda) \in V \times \Lambda$  tal que:

$$\begin{cases} (\nabla u, \nabla v) - \langle v, \lambda \rangle = (f, v) & , \forall v \in V \\ \langle u - g, \mu - \lambda \rangle \geq 0 & , \forall \mu \in \Lambda. \end{cases} \quad (2.35)$$

A existência de uma solução única  $(u, \lambda) \in V \times \Lambda$  para o problema misto (2.35) e a equivalência entre a formulação variacional normal e a formulação variacional com multiplicadores de Lagrange estão provadas por Haslinger, Hlaváček, e Nečas (c.f [11]).

De forma a escrever a formulação variacional numa forma compacta, subtraímos a 2<sup>a</sup> desigualdade na 1<sup>a</sup> igualdade do sistema (2.35), obtendo:

$$(\nabla u, \nabla v) - \langle v, \lambda \rangle - \langle u, \mu - \lambda \rangle \leq (f, v) - \langle g, \mu - \lambda \rangle , \forall (v, \mu) \in V \times \Lambda$$

sendo  $U = V \times Q$  e definimos a forma bilinear  $\mathcal{B} : U \times U \rightarrow \mathbb{R}$ , e a forma linear,  $\mathcal{L} : U \rightarrow \mathbb{R}$ , como:

$$\mathcal{B}(w, \xi; v, \mu) = (\nabla w, \nabla v) - \langle v, \xi \rangle - \langle w, \mu \rangle \quad (2.36)$$

$$\mathcal{L}(v, \mu) = (f, v) - \langle g, \mu \rangle. \quad (2.37)$$

E o problema variacional, pode ser agora reescrito como: encontrar o par  $(u, \lambda) \in V \times \Lambda$  tal que:

$$\mathcal{B}(u, \lambda; v, \mu - \lambda) \leq \mathcal{L}(v, \mu - \lambda) , \forall (v, \mu) \in V \times \Lambda. \quad (2.38)$$

## 2.2 Problemas de Obstáculo com Aplicações à Matemática Financeira

Sabemos que este operador que estamos a tratar no problema do obstáculo (operador parabólico), trata de **problemas de evolução**, problemas cuja solução decorrerá não apenas no espaço estático, mas sim ao longo do tempo, adicionando uma nova variável ao nosso problema. O problema, em concreto, é derivado do **modelo de Black-Scholes**, para **American Options**, cuja formulação forte (strong form)

pode ser escrita como:

$$\begin{cases} \frac{\partial u}{\partial t} + Lu \geq f, & \text{em } \Omega \times (0, T] \\ u \geq g, & \text{em } \Omega \times (0, T] \\ (\frac{\partial u}{\partial t} + Lu - f)(g - u) = 0, & \text{em } \Omega \times (0, T] \\ u(x, 0) = u_0(x), & \text{em } \Omega \times \{t = 0\} \\ u = h, & \text{sobre } \partial\Omega \times (0, T] \end{cases} \quad (2.39)$$

onde  $g$  será a nossa função barreira.

Antes de formular diretamente o nosso problema, precisamos de abordar alguns conceitos preliminares básicos que são necessários para a abordagem do mesmo, pois estamos agora no campo da matemática financeira.

### 2.2.1 Noções e conceitos básicos na matemática financeira

Nesta secção, vamos abordar alguns conceitos preliminares, necessários para o entendimento do problema do obstáculo associado a este contexto (cf. cap. 3 [12], [13] e [14]).

**Ativos** são *bens* e *direitos* que podem ser convertidos em dinheiro. Os bens são as coisas que se possui, e os direitos são dívidas a receber ou benefícios futuros.

Os diferentes ativos podem ser divididos em diversos tipos de grupos e subgrupos, conforme a sua necessidade, e área de atuação. Uma dessas subdivisões que se considera interessante a enunciar no contexto deste trabalho, é a diferenciação entre: **ativos tangíveis**, **ativos intangíveis** e **ativos financeiros**.

- *Ativos tangíveis* são ativos que têm substância física. Por exemplo: imóveis, terrenos, equipamentos, mercadorias, objetos, etc...
- *Ativos intangíveis* são ativos sem substância física. Por exemplo: marcas registradas, patentes, direitos autorais, licenças, software, etc...
- *Ativos financeiros* representam direitos que se têm sobre outras entidades, são instrumentos intangíveis (cujo valor deriva de um acordo contratual existente), sendo representados apenas por papéis / documentos. Por exemplo: dinheiro no bolso ou no banco (o dinheiro é sem dúvida o tipo mais comum de ativo financeiro), empréstimo (os credores permitem, aos mutuários, o uso de fundos por um período de tempo, em troca de um valor adicional pago pelo empréstimo), seguro (o segurado transfere o risco de uma possível perda financeira para a seguradora para mitigar esse risco em troca de uma compensação monetária), acções-stocks (cada ação representa uma pequena parte do capital social

de uma companhia de capital aberto - capital social de uma empresa é o valor total investido pelos sócios, ou acionistas, para financiar ), etc...

Um **derivado** ou *derivativo* são contratos financeiros, cujo valor *deriva* (daí o nome) do valor de outros ativos subjacentes. Um contrato que é definido, ou celebrado, a partir de hoje, em  $t = 0$ , nele é também definida uma data de vencimento/liquidação do contrato, em  $t = T > 0$ , que é chamada de **maturidade**.

As **opções** são um tipo específico de derivativos. Estas opções são vendidas por uma das partes, o **vendedor**, para a outra parte, o **comprador/titular**, por um certo preço, chamado de **prémio**, que é imposto pelo vendedor. Depois de pagar o prémio, o *titular* adquire o direito, mas não a obrigação, de comprar (**opção de compra**, em inglês *call option*) ou vender (**opção de venda**, em inglês *put option*) o ativo subjacente a um preço determinado que representaremos por  $K$ , e é chamado de **preço de exercício**. A principal diferença entre as opções e os outros derivados financeiros é o facto, de após o titular pagar o preço pela opção e o contrato ser estabelecido e estar em vigor, o titular fica com o direito, e não a obrigação de exercer a opção; enquanto o vendedor tem a obrigação de cumprir a sua parte.

Define-se por  $S = S(t) = S_t$  o preço de mercado de um certa ativo subjacente, no tempo; e por  $S(t_0) = S_0$  o preço do ativo subjacente, no tempo  $t_0$ , ou seja no inicio do contrato.

Além disso, supõe-se que  $S$  é um processo estocástico que segue um movimento geométrico Browniano, não iremos abordar a fundo essa questão, mas é uma parte importante para a formulação do nosso problema final.

O problema e o nosso objetivo é a valorização (ou precificação) destas opções, ou seja, é obter o *preço justo da opção*, representado por  $V(S_0, t_0)$ , sendo este o valor que o titular deveria de pagar ao vendedor no inicio do contrato, para a opção entrar em vigor. Este valor teoricamente deveria ser igual ao prémio, que é imposto pelo vendedor, mas não é necessariamente igual, pois o vendedor terá os seus próprios métodos e modelos de precificação.

A função  $V(S, t)$  retorna o valor da opção para qualquer preço do ativo  $S \geq 0$  em qualquer momento  $t \in [0, T]$ . Esta função irá depender de vários fatores, como do preço de exercício  $K$ , do tempo de maturidade  $T$ , da **taxa de juro** representada por  $r$ , da **volatilidade** representada por  $\sigma$  (mede o tamanho das flutuações do preço do ativo subjacente  $S(t)$ , de forma que é um indicador de risco).

O valor da opção na maturidade é chamado de **retorno (payoff)**:  $V(S_T, T)$ , que representará o que

iremos receber no fim do contrato.

Existem vários tipos de opções, mas dentro de todas elas existe um grupo que para além de serem as mais negociadas, são também as mais simples e as mais standard, o que as faz serem conhecidas por *vanilla options* ou *plain vanilla options*, e são duas: as **Opções Europeias** e as **Opções Americanas**.

## 2.2.2 Problema de valor final: Maturidade

Para ambas as opções, calculemos o seu valor na maturidade, o que é comum nos dois casos. Estando na maturidade, o ativo atingiu um determinado valor  $S(T) = S_T$ . Calculemos o **retorno**:  $V(S_T, T)$ , para os dois casos distintos separadamente:

### 2.2.2.1 Opção de compra na maturidade

Neste caso o titular terá 2 hipóteses:

1. Exerce a opção de compra, e pagando o preço de exercício  $K$ , recebendo os ativos pagando apenas esse preço;
2. Não exerce a opção de compra, podendo, se quiser comprar os ativos ao preço de mercado  $S_T$

o titular para minimizar a quantia paga pelo ativo, terá de analizar:

- Se  $S_T \leq K$ , a opção é inútil e o melhor é deixar a opção expirar. Tendo  $V(S_T, T) = 0$ ;
- Se  $S_T > K$ , o titular deverá exercer a opção, e pagar o preço de exercício  $K$ , tendo recebido o ativo ao preço de mercado  $S_T$ , fazendo no total um lucro de  $S_T - K$ . Tendo  $V(S_T, T) = S_T - K$ .

Então, o retorno da opção de compra é

$$\begin{aligned} V_C(S_T, T) &= \begin{cases} 0, & S_T \leq K \\ S_T - K, & S_T > K \end{cases} \\ &= \max \{S_T - K; 0\} \\ &= (S_T - K)^+. \end{aligned} \tag{2.40}$$

### 2.2.2.2 Opção de venda na maturidade

Neste caso o titular terá 2 hipóteses:

1. Exerce a opção de venda, e recebe o preço de exercício  $K$ , vendendo os ativos a esse preço;

2. Não exerce a opção de venda, podendo, se quiser vender os ativos ao preço de mercado  $S_T$ .

o titular para maximizar a quantia recebida pelo ativo, terá de analizar:

- Se  $S_T \geq K$ , a opção é inútil e o melhor é deixar a opção expirar. Tendo  $V(S_T, T) = 0$ ;
- Se  $S_T < K$ , o titular deverá exercer a opção, e recebendo o preço de exercício  $K$ , tendo vendido o ativo ao preço de mercado  $S_T$  por  $K$ , fazendo no total um lucro de  $K - S_T$ , ou seja,  $V(S_T, T) = K - S_T$ .

Então, o retorno da opção de venda é

$$\begin{aligned} V_P(S_T, T) &= \begin{cases} 0, & S_T \geq K \\ K - S_T, & S_T < K \end{cases} \\ &= \max \{K - S_T; 0\} \\ &= (K - S_T)^+. \end{aligned} \tag{2.41}$$

### 2.2.3 Modelo de Black-Scholes

O modelo de Black-Scholes é amplamente utilizado na valoração deste tipo de opções (ver [15]). Este modelo tratará estas questões através de problemas diferenciais parabólicos.

Consideremos o seguinte operador diferencial linear de segunda ordem, associado ao modelo de Black-Scholes dado por:

$$\mathcal{L}_{BS}(V) := \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS \frac{\partial V}{\partial S} - rV \tag{2.42}$$

onde  $r$ , e  $\sigma$  serão constantes sabidas do nosso problema.

Por facilidade de notação, chamaremos a função  $g(S, t)$  o retorno, dado por:

$$\begin{aligned} g_c(S, t) &= \max \{S - K; 0\}, \quad (\text{opção de compra}) \\ g_p(S, t) &= \max \{K - S; 0\}, \quad (\text{opção de venda}) \end{aligned} \tag{2.43}$$

#### 2.2.3.1 Modelo de Black-Scholes para Opções Europeias

As Opções Europeias só podem ser exercidas na maturidade ( $t = T$ ). Numa opção de compra, se  $S_T = 0$  (na maturidade), teremos ganho 0, o que tornará a opção inútil, e irá permanecer inútil ao longo de todo o tempo se  $S(t) = 0$ , e desta forma  $V_C(S = 0, t) = 0$  para  $\forall t \in [0, T]$ . Com o aumento do preço do ativo torna-se mais provável o exercício da opção e menos importante a magnitude real do preço de exercício. Pode-se então escrever com  $S \rightarrow \infty$  o valor da opção  $V_C(S, t) \rightarrow S$  com  $S \rightarrow \infty$ . Mas podemos escrever  $V_C(S, t) \approx S - Ke^{-r(T-t)}$  para ter em consideração o preço de exercício. E portanto

as condições de fronteira para uma opção Europeia de compra são dadas por:

$$\begin{cases} V_C(S, t) = 0 & , \forall (S, t) \in \{0\} \times [0, T] \\ V_C(S, t) = S - Ke^{-r(T-t)} & , \forall (S, t) \in \{\infty\} \times [0, T] \end{cases} \quad (2.44)$$

as condições de contorno correspondentes para uma opção Europeia de venda são dadas por:

$$\begin{cases} V_P(S, t) = Ke^{-r(T-t)} & , \forall (S, t) \in \{0\} \times [0, T] \\ V_P(S, t) = 0 & , \forall (S, t) \in \{\infty\} \times [0, T]. \end{cases} \quad (2.45)$$

Condensamos as condições de fronteira numa função  $h$ , definida por partes, dada por

$$\begin{aligned} h_c^E(S, t) &= \begin{cases} 0, & \forall (S, t) \in \{0\} \times (0, T) \\ S - Ke^{-r(T-t)}, & \forall (S, t) \in \{\infty\} \times (0, T) \end{cases}, \quad (\text{Opção Europeia de compra}) \\ h_p^E(S, t) &= \begin{cases} Ke^{-r(T-t)}, & \forall (S, t) \in \{0\} \times (0, T) \\ 0, & \forall (S, t) \in \{\infty\} \times (0, T) \end{cases}, \quad (\text{Opção Europeia de venda}) \end{aligned} \quad (2.46)$$

e com algum abuso de notação impomos que na fronteira do domínio  $V = h$ ,  $\forall (S, t) \in \partial\{(0, \infty)\} \times [0, T]$ .

Propõe-se a formulação forte do problema:

$$\begin{cases} \frac{\partial V}{\partial t} + \mathcal{L}_{BS}V = 0 & , \text{em } (0, \infty) \times (0, T) \\ V = g & , \text{sobre } (0, \infty) \times \{t = T\} \\ V = h & , \text{sobre } \partial\{(0, \infty)\} \times [0, T] \end{cases} \quad (2.47)$$

cuja solução analítica existe, e é dada pelo sistema (2.48)

$$\begin{cases} V_C(S, t) = SN(d_1) - Ke^{-r(T-t)}N(d_2) \\ V_P(S, t) = -SN(-d_1) + Ke^{-r(T-t)}N(-d_2) \\ d_1 = \frac{\log(S/K) + (r + \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}} \\ d_2 = \frac{\log(S/K) + (r - \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}} \\ N(d) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^d e^{-1/2 \times x^2} dx \end{cases} \quad (2.48)$$

onde  $N(d)$  é a função de distribuição de uma normal standard.

### 2.2.3.2 Modelo de Black-Scholes para Opções Americanas

Nas Opções Americanas, o titular agora tem o direito de exercer a opção em qualquer momento do tempo de vida do contrato ( $0 \leq t \leq T$ ), ao contrário das Opções Europeias, que apenas podem ser exercidas na maturidade ( $t = T$ ). A possibilidade de **exercer antecipadamente** (*early exercise*) nas Opções Americanas dá maiores direitos ao titular do que as Opções Europeias, sendo estas mais restritivas, de tal modo, que as Opções Americanas têm potencialmente um valor mais alto que as Opções Europeias:  $V^{Am} \geq V^{Eur}$ . Note-se que uma Opção Europeia pode ter valores inferiores ao retorno, isto

é  $V^{Eur}(S, t) \leq g(S, t)$ .

Um dos conceitos onde assenta a teoria da matemática financeira é o de não poder haver hipóteses de **arbitragem**. A arbitragem neste contexto, refere-se à exploração de discrepâncias de preço entre ativos para obter lucro sem risco. Por exemplo, se o preço do mesmo derivado é diferente em dois mercados, podemos comprar o derivado no mercado que apresenta um preço mais baixo e vendê-lo no outro, saindo seguramente a ganhar.

De modo a não haver hipóteses de arbitragem o seu valor  $V(S, t)$  deverá ser maior ou igual a  $g(S, t)$ . Assim, a primeira condição geral para opções americanas é:

$$V(S, t) \geq g(S, t) \Leftrightarrow (g(S, t) - V(S, t)) \leq 0 , \quad \forall (S, t) \in (0, \infty) \times [0, T] \quad (2.49)$$

A avaliação de uma Opção Americana é, portanto, mais complicada do que a da Opção Europeia com os mesmos parâmetros. Porque é necessário não só calcular o valor da opção, mas também analisar para cada valor de  $S$  se a opção deve ou não ser exercida.

Denotaremos como  $S_f(t)$  o valor crítico do ativo, onde o valor da opção coincide com o valor do retorno:  $V(S_f, t) = g(S_f, t)$ . O conjunto de pontos  $S_f(t)$  para  $t \in [0, T]$  forma uma **curva de contacto**

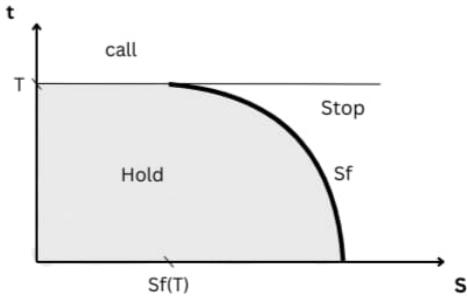
$$\Gamma_{S_f} := \{(S, t) \in \mathbb{R}^2 : t \in [0, T], S = S_f(t)\} \quad (2.50)$$

No contexto financeiro, a localização de  $\Gamma_{S_f}$  é muito importante, pois esta curva, define a separação de duas regiões distintas:

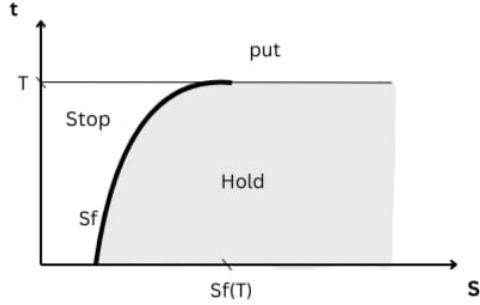
- uma região para a qual o valor da opção é igual ao retorno, chamada **região de paragem** (*stopping region*), na qual exercer antecipadamente a opção é vantajoso;
- e outra região para a qual o valor da opção é maior que o retorno, chamada **região de continuação** (*continuation region ou holding region*).

Para opções de compra americanas, para  $S > S_f(t)$ , temos a igualdade  $V_C^{Am}(S, t) = g(S, t)$  (região de paragem), para  $S < S_f(t)$  temos desigualdade estrita  $V_C^{Am}(S, t) > g(S, t)$  (região de continuação), o que está representado na figura 2.1.

Para opções de venda americanas, para  $S < S_f(t)$ , temos a igualdade  $V_P^{Am}(S, t) = g(S, t)$  (região de paragem), para  $S > S_f(t)$  temos desigualdade estrita  $V_P^{Am}(S, t) > g(S, t)$  (região de continuação), o que está representado na figura 2.2.



**Figura 2.1:** Curva  $S_f$ , de separação das duas regiões para Opções de compra



**Figura 2.2:** Curva  $S_f$ , de separação das duas regiões para Opções de venda

No entanto, como a localização desta curva é desconhecida e *livre*, o que explica porque estes problemas são chamados de **problemas de fronteira livre** (*free boundary problems*).

Precisamos de impor uma condição extra para definir a sua localização, para Opções Americanas de compra e de venda, respetivamente:

$$\begin{cases} V_C(S_f(t), t) = S_f(t) - K \\ \frac{\partial V_C}{\partial S}(S_f(t), t) = 1 \end{cases} \quad (2.51)$$

$$\begin{cases} V_P(S_f(t), t) = K - S_f(t) \\ \frac{\partial V_P}{\partial S}(S_f(t), t) = -1. \end{cases} \quad (2.52)$$

Para impor condições de fronteira ao nosso problema, temos de levar em conta que  $V^{Am} \geq g$ , e podemos mesmo impor  $h^{Ame}(S, t) = \max\{h^{Eur}, g\}$ . Portanto, as condições de fronteira serão muito semelhantes às do problema para opções europeias. Apenas diferem no valor do desconto da taxa de juro ao preço de exercício ( $K$ ), para uma opção de venda quando  $S = 0$ , e para uma opção de compra quando  $S = \infty$ . Assim, as condições de fronteira estão definidas numa função  $h$ , dadas por:

$$\begin{aligned} h_c^{Am}(S, t) &= \begin{cases} 0, & \forall (S, t) \in \{0\} \times (0, T) \\ S - K, & \forall (S, t) \in \{\infty\} \times (0, T) \end{cases}, & (\text{Opção Americana de compra}) \\ h_p^{Am}(S, t) &= \begin{cases} K, & \forall (S, t) \in \{0\} \times (0, T) \\ 0, & \forall (S, t) \in \{\infty\} \times (0, T) \end{cases}, & (\text{Opção Americana de venda}) \end{aligned} \quad (2.53)$$

O problema de valoração das opções americanas pode ser formulado como um problema de obstáculo parabólico e visto como um problema de fronteira livre, onde existe uma fronteira desconhecida que de-

pende do tempo para o exercício da opção:

$$\begin{cases} \frac{\partial V}{\partial t} + \mathcal{L}_{BS}V \leq 0 & , \text{em } (0, \infty) \times (0, T) \\ (g - V) \leq 0 & , \text{em } (0, \infty) \times (0, T) \\ \left( \frac{\partial V}{\partial t} + \mathcal{L}_{BS}V \right) (g - V) = 0 & , \text{em } (0, \infty) \times (0, T) \\ V = g & , \text{sobre } (0, \infty) \times \{t = T\} \\ V = h & , \text{sobre } \partial\{(0, \infty)\} \times (0, T) \end{cases} \quad (2.54)$$

formando-se assim o nosso problema do obstáculo parabólico, onde a existência e unicidade de solução desta formulação forte estão provados em [13].

## 2.2.4 Formulação fraca da inequação de Black-Scholes para American Options

O problema de obstáculo parabólico pode ser formulado como uma inequação variacional. Considere-se então a desigualdade  $u_t + \mathcal{L}_{BS}u \leq 0$  como sendo:

$$u_t + a(x, t)u_{xx} + b(S, t)u_S + c(x, t)u \leq 0, \quad (x, t) \in (0, \infty) \times (0, T) \quad (2.55)$$

onde  $a = \frac{1}{2}\sigma^2x^2$ ,  $b = rS$  e  $c = -r$ . Multiplicando por uma função  $v(x) \in H^1(\Omega) = V$  e integrando em  $\Omega = (0, \infty)$ , obtemos:

$$\langle u_t, v \rangle_{V' \times V} + B[u, v] \leq 0, \quad \forall v \in V \quad (2.56)$$

onde o operador bilinear:

$$B[u, v] = -\frac{1}{2}\sigma^2 \int_{\Omega} x^2 \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} dx + (r - \sigma^2) \int_{\Omega} x \frac{\partial u}{\partial x} v dx - r \int_{\Omega} uv dx + \frac{1}{2}\sigma^2 [x_f^2 u_x(x_f)v(x_f) - x_i^2 u_x(x_i)v(x_i)]$$

e o produto interno dual:

$$\langle u_t, v \rangle_{V' \times V} = \int_{\Omega} u_t v dx.$$

Então a nossa formulação fraca passa a ser encontrar uma solução  $u \in L^2(0, T; V)$  e  $u_t \in L^2(0, T; V')$ , tal que a inequação variacional (2.56) seja válida.

## 2.2.5 Existência, unicidade e regularidade

É possível provar que a formulação fraca do problema (2.54), ou seja (2.56), possui uma única solução fraca, cf. [16]. E tal como no problema do obstáculo elíptico, a regularidade não será superior a  $C^{1,1}$  cf. [17], e estará limitada pela regularidade do obstáculo.

# Capítulo 3

## Métodos numéricos de discretização

Neste capítulo abordaremos alguns métodos numéricos que serão usados para discretizar os nossos problemas. Os resultados aqui apresentados encontram-se aprofundados, e.g., em [18] ou [19].

### 3.1 Método das Diferenças Finitas

O **Método das Diferenças Finitas** (MDF), do inglês para *Finite Difference Method* (FDM), é baseia-se na aproximação dos operadores diferenciais, podendo ser obtido diretamente da expansão de Taylor, usando o método dos coeficientes indeterminados.

#### 3.1.1 Diferenças finitas a 1D

Considere-se um intervalo  $I = [a, b] \subset \mathbb{R}$ , e dividido este em  $N + 1$  subintervalos:  $I = I_0 \cup I_1 \cup \dots \cup I_N$ , ficando com um total de  $N + 2$  pontos ordenados:  $\{x_0; x_1; \dots; x_N; x_{N+1}\}$ . Queremos aproximar uma qualquer derivada, de uma função  $u$ , num qualquer nó  $i$ , ou seja  $D^\alpha u(x_i) \approx D^\alpha U_i$ , à custa dos nós vizinhos a  $i$ . Vamos denotar a função  $u$  avaliada num ponto  $x_i$ , como sendo  $u(x_i) = u_i$ , e a sua aproximação numérica  $U_i \approx u_i$ .

Definimos o passo, ou *step*, (tamanho dos subintervalos) como:

$$h_{i,j} = |x_i - x_j|, \text{ e } h_i = |x_{i+1} - x_i| \quad (3.1)$$

de modo a que:

$$\begin{aligned} h_{j-2,j} &= |x_{j-2} - x_j| = h_{j-2} + h_{j-1} \\ h_{j,j+2} &= |x_j - x_{j+2}| = h_j + h_{j+1}. \end{aligned} \quad (3.2)$$

**Diferenças Centradas (Central Difference):** Na diferenciação centrada, a derivada de uma função  $u(x)$  em torno de um ponto  $x_j$ , é calculada à custa de pontos em ambos os lados do ponto atual, quer pontos olhando para frente quer pontos olhando para trás, relativamente à direção do aumento do eixo  $x$ . Isto é usando os pontos tendo em conta alguma simetria:

$$\{u_{j-n}; \dots; u_{j-1}; u_j; u_{j+1}; \dots; u_{j+n}\}, \quad n \in \mathbb{N}.$$

Aproximemos a primeira derivada,  $u'(x)$ , usando os pontos  $u_{j-1}$ ,  $u_j$  e  $u_{j+1}$ , pelo método dos coeficientes indeterminados:

tes indeterminados, fica finalmente com:

$$\begin{aligned} u'(x_j) &= \frac{-h_j}{h_{j-1}^2 + h_j h_{j-1}} u_{j-1} + \frac{h_j - h_{j-1}}{h_j h_{j-1}} u_j + \frac{h_{j-1}}{h_j^2 + h_{j-1} h_j} u_{j+1} + \dots \\ &\dots + \frac{-h_j}{h_{j-1}^2 + h_j h_{j-1}} \frac{1}{3!} (-h_{j-1})^3 u^{(3)}(\xi_{j-1}) + \frac{h_{j-1}}{h_j^2 + h_{j-1} h_j} \frac{1}{3!} h_j^3 u^{(3)}(\xi_{j+1}). \end{aligned} \quad (3.3)$$

De modo semelhante, aproximemos a segunda derivada,  $u''(x)$ , usando os pontos  $u_{j-1}$ ,  $u_j$  e  $u_{j+1}$ :

$$\begin{aligned} u''(x_j) &= \frac{2}{h_{j-1}^2 + (h_{j-1} h_j)} u_{j-1} + \frac{-2}{h_{j-1} h_j} u_j + \frac{2}{h_j^2 + (h_{j-1} h_j)} u_{j+1} + \dots \\ &\dots + \frac{2}{h_{j-1}^2 + (h_{j-1} h_j)} \frac{1}{3!} (-h_{j-1})^3 u^{(3)}(\xi_{j-1}) + \frac{2}{h_j^2 + (h_{j-1} h_j)} \frac{1}{3!} h_j^3 u^{(3)}(\xi_{j+1}). \end{aligned} \quad (3.4)$$

É claro que considerando um espaçamento uniforme, ou seja  $h = h_j = h_{j-1}$  para todo  $j$ , então as formulas serão mais simplificadas:

$$\begin{aligned} u'(x_j) &= D_x u_j + O(h^2) = \frac{u_{j+1} - u_{j-1}}{2h} + O(h^2) \\ u''(x_j) &= D_{xx} u_j + O(h^2) = \frac{u_{j-1} - 2u_j + u_{j+1}}{h^2} + O(h^2). \end{aligned} \quad (3.5)$$

**Diferenças Progressivas (Forward Difference):** A diferenciação para a frente aproxima a derivada de uma função  $u(x)$  em torno de um ponto  $x_j$ , olhando para frente na direção do aumento do eixo  $x$ . Isso significa que se usa o valor da função no ponto atual e pontos à frente para calcular a derivada. Ou seja:

$$\{u_j; u_{j+1}; \dots; u_{j+n}\}, \quad n \in \mathbb{N}.$$

A aproximação da primeira derivada,  $u'(x)$ , usando os pontos  $u_j$  e  $u_{j+1}$ , é dada por:

$$u'(x_j) = \frac{u_{j+1} - u_j}{h_j} + \frac{1}{2} h_j u^{(2)}(\xi_{j+1}). \quad (3.6)$$

Para aproximar a segunda derivada,  $u''(x)$ , já é necessário usarmos 3 nós, usemos então os nós  $u_j$ ,  $u_{j+1}$  e  $u_{j+2}$ , ficamos com:

$$\begin{aligned} u''(x) &= \frac{2}{h_j h_{j,j+2}} u_j + \frac{-2}{h_j^2 + h_j h_{j,j+2}} u_{j+1} + \frac{-2}{h_{j,j+2}^2 + h_j h_{j,j+2}} u_{j+2} + \\ &+ \frac{-2h_j^3}{-h_j^2 + h_j h_{j,j+2}} \frac{1}{3!} u^{(3)}(\xi_{j+1}) + \frac{-2h_{j,j+2}^3}{-h_{j,j+2}^2 + h_j h_{j,j+2}} \frac{1}{3!} u^{(3)}(\xi_{j+2}) \end{aligned} \quad (3.7)$$

Considerando um espaçamento uniforme, simplificamos as fórmulas:

$$\begin{aligned} u'(x_j) &= D_x^+ u_j + O(h) = \frac{u_{j+1} - u_j}{h} + O(h) \\ u''(x_j) &= D_{xx}^+ u_j + O(h) = \frac{u_j - 2u_{j+1} + u_{j+2}}{h^2} + O(h) \end{aligned} \quad (3.8)$$

verificando que  $D_{xx}^+ u_j = D_x^+ D_x^+ u_j$ .

**Diferenças Regressivas (Backward Difference):** A diferenciação para trás aproxima a derivada de uma função  $u(x)$  em torno de um ponto  $x_j$  olhando para trás, relativamente à direção de aumento do eixo  $x$ . Isso significa que se usa o valor da função no ponto atual e pontos a trás para calcular a derivada. Ou seja:

$$\{u_{j-n}; \dots; u_{j-1}; u_j\}, \quad n \in \mathbb{N}$$

A aproximação da primeira derivada,  $u'(x)$ , usando os pontos  $u_{j-1}$  e  $u_j$ , é dada por:

$$u'(x_j) = \frac{u_j - u_{j-1}}{h_{j-1}} - \frac{1}{2} h_{j-1} u^{(2)}(\xi_{j-1}) \quad (3.9)$$

para aproximar a segunda derivada,  $u''(x)$ , tal como na diferença progressiva, é necessário usarmos 3 nós, usemos então os nós  $u_{j-2}$ ,  $u_{j-1}$  e  $u_j$ : ficamos com

$$\begin{aligned} u''(x) &= \frac{-2}{-h_{j-2,j}^2 + h_{j,j-2}h_{j-1,j}} u_{j-2} + \frac{-2}{-h_{j-1,j}^2 + h_{j,j-2}h_{j-1,j}} u_{j-1} + \frac{2}{h_{j-2,j}h_{j-1,j}} u_j + \\ &- \frac{-2h_{j-2,j}^3}{-h_{j-2,j}^2 + h_{j,j-2}h_{j-1,j}} \frac{1}{3!} u^{(3)}(\xi_{j-2}) - \frac{-2h_{j-1,j}^3}{-h_{j-1,j}^2 + h_{j,j-2}h_{j-1,j}} \frac{1}{3!} u^{(3)}(\xi_{j-1}) \end{aligned} \quad (3.10)$$

Considerando um espaçamento uniforme, simplificamos as formulas:

$$\begin{aligned} u'(x_j) &= D_x^- u_j + O(h) = \frac{u_j - u_{j-1}}{h} + O(h) \\ u''(x_j) &= D_{xx}^- u_j + O(h) = \frac{u_j - 2u_{j-1} + u_{j-2}}{h^2} + O(h) \end{aligned} \quad (3.11)$$

verificando que  $D_{xx}^- u_j = D_x^- D_x^- u_j$ .

### 3.1.1.1 Diferenças FInitas em 2D

A diferença crucial do caso a 2D, comparativamente com o caso a 1D, é que agora temos de ter em consideração a discretização de 2 direções distintas. Os métodos de diferenças finitas são fáceis de implementar em domínios simples em forma de retângulo ou caixa. Formas mais complicadas do domínio exigem técnicas e esforços de implementação substancialmente mais avançados. Num

domínio retangular ou em forma de caixa, os pontos da malha são introduzidos separadamente nas várias direções do espaço:

$$x_0 < x_1 < \dots < x_{N_x} < x_{N_x+1}$$

$$y_0 < y_1 < \dots < y_{N_y} < y_{N_y+1}$$

com space steps:

$$\begin{aligned} &\{h_0^x; h_1^x; \dots; h_{N_x-1}^x; h_{N_x}^x\}, \text{ onde } h_j^x = x_{j+1} - x_j \text{ para } j = 0, \dots, N_x \\ &\{h_0^y; h_1^y; \dots; h_{N_y-1}^y; h_{N_y}^y\}, \text{ onde } h_i^y = y_{i+1} - y_i \text{ para } i = 0, \dots, N_y \end{aligned}$$

e ainda, definimos como no caso unidimensional, o passo (*space step*) entre diferentes pontos:

$$h_{m,n}^x = |x_m - x_n|, \quad h_{m,n}^y = |y_m - y_n|$$

Seja  $u(x_j, y_i) = u_{j,i}$ . A aproximação das derivadas  $\frac{\partial u}{\partial x}$ ,  $\frac{\partial^2 u}{\partial x^2}$ ,  $\frac{\partial u}{\partial y}$  e  $\frac{\partial^2 u}{\partial y^2}$ , é feita exatamente como no caso unidimensional, quer para as diferenças progressivas, regressivas ou centradas. Como podemos ver, por exemplo no caso das diferenças centradas, temos:

$$\begin{aligned} \frac{\partial u(x_j, y_i)}{\partial x} &\approx \frac{-h_j}{h_{j-1}^2 + h_j h_{j-1}} u_{j-1,i} + \frac{h_j - h_{j-1}}{h_j h_{j-1}} u_{j,i} + \frac{h_{j-1}}{h_j^2 + h_{j-1} h_j} u_{j+1,i} \\ \frac{\partial u(x_j, y_i)}{\partial y} &\approx \frac{-h_j}{h_{j-1}^2 + h_j h_{j-1}} u_{j-1,i} + \frac{h_j - h_{j-1}}{h_j h_{j-1}} u_{j,i} + \frac{h_{j-1}}{h_j^2 + h_{j-1} h_j} u_{j+1,i} \end{aligned} \quad (3.12)$$

$$\begin{aligned} \frac{\partial^2 u(x_j, y_i)}{\partial x^2} &\approx \underbrace{\frac{2}{(h_{j-1}^x)^2 + (h_{j-1}^x h_j^x)}}_{=a_{j-1}^x} u_{j-1,i} + \underbrace{\frac{-2}{h_{j-1}^x h_j^x}}_{=b_j^x} u_{j,i} + \underbrace{\frac{2}{(h_j^x)^2 + (h_{j-1}^x h_j^x)}}_{=c_{j+1}^x} u_{j+1,i} = \\ &= a_{j-1}^x u_{j-1,i} + b_j^x u_{j,i} + c_{j+1}^x u_{j+1,i} \end{aligned} \quad (3.13)$$

$$\begin{aligned} \frac{\partial^2 u(x_j, y_i)}{\partial y^2} &\approx \underbrace{\frac{2}{(h_{i-1}^y)^2 + (h_{i-1}^y h_i^y)}}_{=a_{i-1}^y} u_{j,i-1} + \underbrace{\frac{-2}{h_{i-1}^y h_i^y}}_{=b_i^y} u_{j,i} + \underbrace{\frac{2}{(h_i^y)^2 + (h_{i-1}^y h_i^y)}}_{=c_{i+1}^y} u_{j,i+1} \\ &= a_{i-1}^y u_{j,i-1} + b_i^y u_{j,i} + c_{i+1}^y u_{j,i+1} \end{aligned} \quad (3.14)$$

e portanto o Laplaciano:

$$\begin{aligned} \Delta u(x_j, y_i) &= \frac{\partial^2 u(x_j, y_i)}{\partial x^2} + \frac{\partial^2 u(x_j, y_i)}{\partial y^2} = \\ &= [a_{j-1}^x] u_{j-1,i} + [a_i^y] u_{j,i-1} + (b_j^x + b_i^y) u_{j,i} + [c_{j+1}^x] u_{j+1,i} + [c_{i+1}^y] u_{j,i+1} \end{aligned} \quad (3.15)$$

se a malha for igualmente espaçada no mesmo eixo, ou seja  $h_j^x = h_x$  e  $h_i^y = h_y$  para todo  $j$  e todo o  $i$ :

$$\Delta u(x_j, y_i) = \frac{u_{j-1,i} - 2u_{j,i} + u_{j+1,i}}{h_x^2} + \frac{u_{j,i-1} - 2u_{j,i} + u_{j,i+1}}{h_y^2} + O(h_x^2) + O(h_y^2). \quad (3.16)$$

## 3.2 Método dos Elementos Finitos

O **Método dos Elementos Finitos** (MEF), do inglês *Finite Element Method* (FEM), é mais uma alternativa para a resolução de EDPs, e tende a ser mais adequados para lidar com geometrias complexas.

A ideia essencial, desenvolvida a partir da década de 50 e remontando a Galerkin (1915), consiste em reformular o problema usando uma formulação variacional equivalente, utilizando funções teste suficientemente regulares. Ao fazer isso, a maior regularidade das funções teste compensa uma menor regularidade da solução, através de uma transferência que se baseia no uso da fórmula de Green.

Assumindo que a solução  $u$  pertence a um dado espaço de Hilbert, denotado por  $V$  (que é um espaço de dimensão infinita), o objetivo deste método passará por encontrar uma solução  $u_h$ , contida num subespaço de dimensão finita (finite dimensional subspace), denotado por  $V_h \subset V$ , e será uma aproximação do espaço de solução original.

Note, passamos a resolver o nosso problema num espaço de dimensão finita, mas a questão é qual devem ser os espaços  $V_h$  que garantem que a solução aproximada converge para a solução exata. A resposta é clara usando o FEM.

Como veremos, o espaço de dimensão finita  $V_h$  consiste em funções compostas por polinómios.

### 3.2.1 Discretização geométrica do domínio: Mesh

Considerando um domínio  $\Omega \subset \mathbb{R}^n$ , o qual queremos discretizar. Primeiro temos de o aproximar por um domínio poligonal  $\tilde{\Omega}$  através de uma subdivisão do domínio  $\Omega$  em uma partição ou mesh  $\mathcal{T} = \{T_1, \dots, T_{N_{el}}\}$ , i.e., numa colecção de elementos geometricamente simples. Estes elementos dependendo da dimensão  $n$  do domínio, em 1D serão intervalos ou segmentos, em 2D serão polígonos (como triângulos ou quadriláteros), em 3D serão poliedros (como tetraedros).

Este processo de discretização geométrica do domínio é chamado de *malhagem* (ou por *triangularização* (triangulation) que o uso de triângulos é o mais vulgar, mas sendo um nome abusivo, no caso de não se tratarem de triângulos).

Os elementos  $T_i$ ,  $i = 1, \dots, N_{el}$ , satisfazem as seguintes suposições:

1. Cada subdomínio  $T_i$  é chamado de *elemento*, e está contido em  $\Omega$ , ou seja:  $T_i \subset \Omega$ ;

2. Cada elemento  $T_i$  é um conjunto compacto, conexo, com fronteira lipschitziana, e interior não vazio;
3. A união finita de todos os  $N_{el}$  elementos é chamada de *malha* (mesh), que cobre por completo o domínio  $\bar{\Omega}$ , ou seja,

$$\bar{\Omega}_h = \bigcup_{i=1}^{N_{el}} T_i = \bigcup_{T \in \mathcal{T}_h} T, \quad \Omega_h = \text{int} \left\{ \bigcup_{T \in \mathcal{T}_h} T \right\}.$$

Iremos referir esta triangulação como  $\mathcal{T}_h$  ou  $\Omega_h$ , revelando que se trata de uma discretização do domínio  $\Omega$ ;

4. Vamos denotar por  $h_T$ , o diâmetro de um elemento arbitrário  $T$ , e definimos  $h$  sendo o máximo valor  $h_T, T \in \mathcal{T}_h$  de toda a malha, ou seja:

$$\forall T \in \mathcal{T}_h : h_T = \text{diam} \{T\} \text{ e seja então } h = \max_{T \in \mathcal{T}_h} \{h_T\}$$

ou de outra maneira

$$h = \max_{T \in \mathcal{T}_h} \text{diam} \{T\}$$

onde o sub-índice  $h$  refere-se ao nível de refinamento da malha, de modo que à medida que  $h \rightarrow 0$  os elementos ficam cada vez menores e obtemos uma malha fina em  $\Omega$ .

5. Dois elementos distintos  $T_i, T_j \in \mathcal{T}, i \neq j$  não se sobrepõem, verificando sempre:

$$\text{int} \{T_i\} \cap \text{int} \{T_j\} = \emptyset, \quad \forall i \neq j.$$

### 3.2.2 Elementos Finitos - Tripleto

Para além de uma definição e caracterização geométrica dos elementos finitos, vamos também associar-lhe um espaço de funções e um conjunto de pontos (nós) que nos interessam para efeitos de interpolação. Muitas vezes, para além dos vértices dos elementos finitos, para fazer a interpolação precisamos de mais nós desse elemento finito. No caso de elementos mais complicados, em que se pretende interpolar também derivadas, não chega considerar apenas os nós, já que num mesmo nó condicionamos não apenas o valor da função, mas também o valor das suas derivadas. Aparece assim a noção de variáveis nodais.

**Definição 1** (Elemento finito (por Ciarlet)). *Um elemento finito é o tripleto  $(T, \mathcal{P}, \mathcal{N})$ , em que:*

1.  $T$  é o elemento geométrico, um subconjunto compacto e conexo de  $\mathbb{R}^n$ , com fronteira regular (tipicamente Lipschitz-contínuo);

2.  $\mathcal{P}$  é um espaço vetorial de dimensão finita,  $\dim \{\mathcal{P}\} = N$  onde se encontram funções definidas em  $T$  (o espaço das funções de forma (shape functions)):

$$\mathcal{P} = \text{span} \{\psi_1, \dots, \psi_M\}$$

tal que para  $p \in \mathcal{P}$  temos  $p : T \rightarrow \mathbb{R}$ ;

3. o conjunto de variáveis nodais:  $\mathcal{N} = \{\nu_1, \dots, \nu_N\}$  é uma base nodal para  $\mathcal{P}^* = \mathcal{L}(\mathcal{P}, \mathbb{R})$ , que é o espaço dual de  $\mathcal{P}$ ,

$$\nu_j : \mathcal{P} \rightarrow \mathbb{R}$$

$$p \mapsto p_j = \nu_j(p)$$

e portanto o espaço dual  $\mathcal{P}^*$ , é constituído por aplicações lineares  $\nu$  que transformam funções  $p \in \mathcal{P}$  em números reais, designando-se normalmente por

$$\langle \nu, p \rangle = \nu(p)$$

**Definição 2** (Triangulação ou malhagem). Designamos por triangulação uma família de elementos finitos

$$\mathcal{T}_h = \bigcup_{E \subseteq \tilde{\Omega}} (E, \mathcal{P}_E, \mathcal{N}_E)$$

em que o parâmetro  $h$  é definido por  $h = \max h_E$ .

### 3.2.3 Formulação Variacional Discreta - Aproximação de Galerkin

Embora os problemas que estamos a tratar neste trabalho não sejam igualdades mas sim desigualdades, para motivar o uso do FEM, iremos considerar o problema variacional composto pela seguinte seguinte igualdade:

$$\begin{cases} \text{Encontrar } u \in V \text{ tal que:} \\ a(u, v) = L(v), \quad \forall v \in V. \end{cases} \quad (3.17)$$

Para existir uma e uma só solução para o problema variacional em  $V$ , basta garantir algumas propriedades dos operadores que a compõem, o que está enunciado no seguinte teorema:

**Teorema 4** (Lax-Milgram). Seja  $V$  um espaço de Hilbert, com norma  $\|\cdot\|_V$ , induzida pelo produto interno. Seja  $a : V \times V \rightarrow \mathbb{R}$  uma forma bilinear (bilinear form) contínua e coerciva em  $V$ , e seja  $L : V \rightarrow \mathbb{R}$  um funcional linear e limitado em  $V$ . Resumindo,  $a$  e  $L$ , satisfazem as seguintes propriedades:

des, para todo  $\forall u, v \in V$ :

$$\begin{aligned} |a(u, v)| &\leq C\|u\|_V\|v\|_V, \text{ (continuidade-continuity)} \\ a(u, v) &\geq \alpha\|u\|_V^2, \text{ (coercividade-coercivity)} \\ |L(v)| &\leq C_L\|v\|_V, \text{ (limitado-boundedness)} \end{aligned} \tag{3.18}$$

e algumas constantes  $C, \alpha, C_L > 0$  independentes de  $u$  e  $v$ . Então existe um único  $u \in V$ , tal que:

$$a(u, v) = L(v), \quad \forall v \in V. \tag{3.19}$$

**Demonstração:** A demonstração irá baseia-se no *Teorema de Representação de Riesz* e no *Teorema do Ponto Fixo de Banach*. Para esta prova pode-se consultar qualquer livro clássico de análise funcional e EDP's, como Evans [1], ou ainda [18] e [19].

Assim, associado ao problema variacional: dado  $L \in V^*$ , encontrar  $u \in V$  tal que

$$a(u, v) = L(v), \quad \forall v \in V.$$

Consideramos um problema variacional em dimensão finita, que designamos como *aproximação de Galerkin*, e que consiste em considerar um subespaço de dimensão finita  $V_h \subset V$ , e encontrar  $u_h \in V_h$  tal que

$$a(u_h, v_h) = L(v_h), \quad \forall v_h \in V_h.$$

Ao garantir que a forma  $a$  é coerciva, e como supomos  $V_h \subset V$ , isso implica imediatamente que na discretização também se tenha

$$a(v_h, v_h) \geq \alpha\|v_h\|^2, \quad \forall v_h \in V_h$$

Como estamos a supor uma aproximação por um espaço de dimensão finita  $V_h$ , existirá uma base finita de funções:

$$\{\psi_1, \dots, \psi_{N_h}\}$$

tal que:

$$V_h = \text{span}\{\psi_1, \dots, \psi_{N_h}\}$$

com o número total de funções bases:

$$N_h = \dim(V_h).$$

Então toda a função teste  $v_h \in V_h$  pode ser escrita como uma combinação linear de funções base:

$$v_h = \sum_{j=1}^{N_h} \alpha_j \psi_j(x), \quad \alpha_j \in \mathbb{R} \quad \forall j = \{1, \dots, N_h\}$$

de um estilo semelhante:

$$u_h = \sum_{i=1}^{N_h} \mu_i \psi_i(x), \quad \mu_i \in \mathbb{R} \quad \forall i = \{1, \dots, N_h\}.$$

Verificando-se a relação, para todo  $j = \{1, \dots, N_h\}$ , escolhendo apenas uma base para função teste  $v_h = \alpha_j \psi_j(x)$ :

$$\begin{aligned} a[u_h, v_h] = L(v_h) &\Leftrightarrow a[u_h, \alpha_j \psi_j(x)] = L(\alpha_j \psi_j(x)) \Leftrightarrow \alpha_j a[u_h, \psi_j(x)] = \alpha_j L(\psi_j(x)) \Leftrightarrow \\ &\Leftrightarrow a[u_h, \psi_j] = L(\psi_j) \end{aligned}$$

substituindo  $u_h = \sum_{i=1}^{N_h} \mu_i \psi_i(x)$ , obtemos o sistema:

$$\sum_{i=1}^{I_h} \mu_i a[\psi_i, \psi_j] = L(\psi_j), \quad \forall j = 1, \dots, I_h$$

que pode ser escrito como um sistema matricial linear para determinar  $\mathbf{U} = [\mu_1, \dots, \mu_{N_h}]^T \in \mathbb{R}^{N_h}$ :

$$\mathbf{B} \mathbf{U} = \mathbf{L}$$

onde a matriz  $\mathbf{B} \in \mathbb{R}^{N_h \times N_h}$ , tem as suas entradas dadas por:

$$b_{ij} = a[\psi_i, \psi_j]$$

e o vetor  $\mathbf{L} = [l_1, \dots, l_{N_h}]^T \in \mathbb{R}^{N_h}$ , tem as suas entradas dadas por:

$$l_j = L(\psi_j).$$

Podemos agora apresentar um resultado que estabelece uma certa proporcionalidade entre a aproximação de Galerkin e a melhor aproximação possível no espaço  $V_h$ . Este resultado será crucial para estabelecermos estimativas de erro para a aproximação por elementos finitos.

Antes de provar o *Lema de Céa*, provamos um resultado auxiliar.

**Lema 5** (Ortogonalidade de Galerkin). *Sejam  $V$  um espaço de Hilbert e  $V_h \subset V$  um subespaço de dimensão finita de  $V$ . Seja  $a(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$  uma forma bilinear, seja  $u \in V$  e  $u_h \in V_h$ , respectivamente,*

*soluções de:*

$$\begin{aligned} &\text{encontrar } u \in V : a(u, v) = L(v), \forall v \in V \\ &\text{encontrar } u_h \in V_h : a(u_h, v_h) = L(v_h), \forall v_h \in V_h \end{aligned} \quad (3.20)$$

*então,*  $a(u - u_h, v_h) = 0, \forall v_h \in V_h.$

**Demonstração:**

Usando a linearidade de  $a(\cdot, \cdot)$  na sua primeira componente, obtemos para qualquer  $v_h \in V_h$ ,

$$a(u - v_h, v_h) = a(u, v_h) - a(v_h, v_h) = L(v_h) - L(v_h) = 0$$

□

**Teorema 6** (Lema de Céa). *Sejam  $V$  é um espaço de Hilbert com norma  $\|\cdot\|_V$  induzida pelo produto interno e  $V_h$  um subespaço de dimensão finita de  $V$ . Seja  $a(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$  uma forma bilinear coerciva e contínua em  $V$ , e  $L(\cdot) : V \rightarrow \mathbb{R}$  um funcional linear e limitado em  $V$ . Seja ainda  $u \in V$  e  $u_h \in V_h$ , respectivamente, soluções de:*

$$\begin{aligned} &\text{encontrar } u \in V : a(u, v) = L(v), \forall v \in V \\ &\text{encontrar } u_h \in V_h : a(u_h, v_h) = L(v_h), \forall v_h \in V_h \end{aligned} \quad (3.21)$$

*então*

$$\|u - u_h\|_V \leq \frac{C}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V. \quad (3.22)$$

**Demonstração** Ver [19].

Usando o facto de  $a(\cdot, \cdot)$  ser bilinear, temos, para qualquer  $v_h \in V_h$ ,

$$\begin{aligned} a(u - u_h, u - u_h) &= a(u - u_h, u - u_h + v_h - v_h) = \\ &= a(u - u_h, u - v_h + v_h - u_h) = \\ &= a(u - u_h, u - v_h) + a(u - u_h, \underbrace{v_h - u_h}_{w_h \in V_h}) = a(u - u_h, u - v_h) \\ &\quad \underbrace{\quad}_{=0, \text{ pelo lema 5}} \end{aligned}$$

pela coercividade de  $a$ , ou seja  $a(u - u_h, u - u_h) \geq \alpha \|u - u_h\|_V^2$ , então através de algumas manipulações, e usando o resultado do último cálculo, e da sua linearidade:

$$\|u - u_h\|_V^2 \leq \frac{1}{\alpha} a(u - u_h, u - u_h) = \frac{1}{\alpha} a(u - u_h, u - v_h) \leq \frac{C}{\alpha} \|u - u_h\|_V \|u - v_h\|_V, \forall v_h \in V_h$$

dividindo este último resultado por  $\|u - u_h\|_h$  obtemos:

$$\|u - u_h\|_h \leq \frac{C}{\alpha} \|u - u_h\|_V, \quad \forall v_h \in V_h \quad (3.23)$$

e em particular

$$\|u - u_h\|_h \leq \frac{C}{\alpha} \inf_{v_h \in V_h} \|u - u_h\|_V. \quad (3.24)$$

□

Este resultado afirma que o erro é comparável ao erro da melhor aproximação de  $u$  de  $V_h$ .

### 3.2.4 Elementos de Lagrange

Dado um elemento  $T$ , iremos demarcar como  $z_j$ , todos os pontos (nós), que serão usados para interpolação, nestes pontos costumam estar os vértices deste elemento (isto é, as suas extremidades), e também outros pontos estratégicos para a sua interpolação. Exigindo, através das variáveis nodais, para cada nó, a condição de Lagrange:

$$\nu_i(\phi_j) = \phi_j(z_i) = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

a função será portanto  $C^0(T)$ .

### 3.2.5 Elementos a 1D

Considerando um domínio  $\Omega = (a, b)$ , onde a solução estará definida, faça-se uma discretização do intervalo espacial  $\bar{\Omega} = [a, b]$ , num total de  $M_x + 2$  pontos:

$$a = x_0 < x_1 < \dots < x_{M_x} < x_{M_x+1} = b$$

onde dois desses pontos são as pontas (a fronteira):  $x_0 = a$  e  $x_{M_x+1} = b$ ; e os restantes  $M_x$  pontos  $x_m$  com  $m = 1, \dots, M_x$  são do interior.

Ficamos com um total de  $M_x + 1$  intervalos, da forma:

$$I_m = [x_m, x_{m+1}], \quad m = 0, \dots, M_x$$

cada um com comprimento ou passo:

$$h_m = x_{m+1} - x_m, \quad m = 0, \dots, M_x$$

e cada intervalo deste será um elemento finito da nossa descretização.

Dado o elemento de referência  $\hat{I} = [-1, 1]$ , para o qual temos as funções base para elementos de Lagrange lineares, neste elemento, dadas por:

$$\begin{cases} \hat{\phi}_1 = \frac{1}{2}(1 - \hat{x}) \\ \hat{\phi}_2 = \frac{1}{2}(1 + \hat{x}) \end{cases}, \hat{x} \in \hat{I} = [-1, 1] \quad (3.25)$$

No caso linear, uma função base resulta da colagem de funções de forma de elementos vizinhos. A função base associada ao ponto  $z_k$ :

$$\psi_k(x) = \begin{cases} \phi_2^{I_{k-1}} & , x \in I_{k-1} \\ \phi_1^{I_k} & , x \in I_k \quad , k = 0, 1, \dots, M_x + 1 \\ 0 & , c.c. \end{cases}$$

Finalmente as funções bases, para cada intervalo serão dadas por:

$$\begin{aligned} \psi_0(x) &= \begin{cases} \frac{x_1 - x}{x_1 - x_0}, & x \in ]x_0, x_1[ \\ 0, & c.c. \end{cases} \\ \psi_j(x) &= \begin{cases} \frac{x - x_{j-1}}{x_j - x_{j-1}}, & x \in ]x_{j-1}, x_j[ \\ \frac{x_{j+1} - x}{x_{j+1} - x_j}, & x \in ]x_j, x_{j+1}[ \quad , j = 1, \dots, M_x \\ 0, & c.c. \end{cases} \\ \psi_{M_x+1}(x) &= \begin{cases} \frac{x - x_{M_x}}{x_{M_x+1} - x_{M_x}}, & x \in ]x_{M_x}, x_{M_x+1}[ \\ 0, & c.c. \end{cases} \end{aligned} \quad (3.26)$$

### 3.2.6 Elementos a 2D: Triângulo

Seja  $\Omega \subset \mathbb{R}^2$  um conjunto aberto e limitado, faça-se uma partição regular  $\Omega_h = \{T_1, \dots, T_M\}$  em elementos triangulares, onde  $h = \max_{T \in \Omega_h} h_T$  é tamanho máximo da malha, e  $h_T$  é o comprimento máximo de um triângulo arbitrário  $T$ , com 3 vértices que representamos como

$$\{v_1; v_2; v_3\} = \{(v_{1,x}, v_{1,y}); (v_{2,x}, v_{2,y}); (v_{3,x}, v_{3,y})\}$$

$v_i$  e  $i = 1, 2, 3$ , é dado por  $h_T = \text{diam}\{T\}$ , que também pode ser calculado através de:

$$h_T = \max \{\|v_i - v_j\|\}, \quad \forall i, j = 1, 2, 3.$$

Consideremos agora o elemento de referência  $\hat{T}$ , com 3 vértices:

$$\{\hat{v}_1; \hat{v}_2; \hat{v}_3\} = \{(0, 0); (1, 0); (0, 1)\}.$$

Defina-se uma transformação linear,  $X = G_T(\hat{X}) = [A_T]\hat{X} + (b)$ , que mapeia o triângulo de referência,  $\hat{T}$ , para um triângulo genérico  $T$ :

$$\begin{aligned} G_T : \quad \hat{T} &\rightarrow \quad T \\ \hat{X} \mapsto \quad X &= G_T(\hat{X}) = [A_T]\hat{X} + (b) \end{aligned} \tag{3.27}$$

tal que

$$X = G_T(\hat{X}) \Leftrightarrow \begin{pmatrix} x \\ y \end{pmatrix} = G_T(\hat{x}, \hat{y}) = [A_T] \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} + (b_T)$$

onde as variáveis  $\hat{X} = (\hat{x}, \hat{y}) \in \hat{T}$  e  $X = (x, y) \in T$ . A matriz  $[A_T] \in \mathbb{R}^{2 \times 2}$  e o vetor  $(b_T) \in \mathbb{R}^2$  são obtidos impondo:

$$v_1 = G_T(\hat{v}_1), v_2 = G_T(\hat{v}_2), v_3 = G_T(\hat{v}_3) \tag{3.28}$$

obtemos:

$$\begin{aligned} [A_T] &= [(v_2) - (v_1) \mid (v_3) - (v_2)] = \begin{bmatrix} v_{2,x} - v_{1,x} & v_{3,x} - v_{1,x} \\ v_{2,y} - v_{1,y} & v_{3,y} - v_{1,y} \end{bmatrix} \\ (b_T) &= (v_1) = \begin{pmatrix} v_{1,x} \\ v_{1,y} \end{pmatrix} \end{aligned} \tag{3.29}$$

Dado que

$$G(\hat{x}, \hat{y}) = (g_1(\hat{x}, \hat{y}), g_2(\hat{x}, \hat{y})) = \left( \underbrace{(A_T)_{1,1}\hat{x} + (A_T)_{1,2}\hat{y} + b_{T1}}_{:=x(\hat{x}, \hat{y})}, \underbrace{(A_T)_{2,1}\hat{x} + (A_T)_{2,2}\hat{y} + b_{T2}}_{:=y(\hat{x}, \hat{y})} \right)$$

vem que a jacobiana:

$$\frac{\partial(x, y)}{\partial(\hat{x}, \hat{y})} = \nabla_{(\hat{x}, \hat{y})} G(\hat{X}) = DG(\hat{x}, \hat{y}) = \begin{bmatrix} \frac{\partial g_1}{\partial \hat{x}} & \frac{\partial g_1}{\partial \hat{y}} \\ \frac{\partial g_2}{\partial \hat{x}} & \frac{\partial g_2}{\partial \hat{y}} \end{bmatrix} = \begin{bmatrix} (A_T)_{1,1} & (A_T)_{1,2} \\ (A_T)_{2,1} & (A_T)_{2,2} \end{bmatrix} = [A_T]$$

e o determinante da jacobiana:

$$\det \left\{ \frac{\partial(x, y)}{\partial(\hat{x}, \hat{y})} \right\} = \det \{[A_T]\}$$

e a área diferencial:

$$dX = \left| \det \left\{ \frac{\partial(x, y)}{\partial(\hat{x}, \hat{y})} \right\} \right| d\hat{X} = \det \{[A_T]\} d\hat{X}$$

### 3.2.6.1 Elementos de Lagrange - 2D: Triângulo

A 2D, o espaço de funções polinomiais de até grau  $K$ , é representado por  $\mathcal{P}_K(\mathbb{R}^2)$ , cujo a base monómios:

$$\{1; x; y; xy; x^2; y^2; x^2y; xy^2; x^3; y^3; \dots\}.$$

### 3.2.6.2 Lagrange de 1º Grau - Lineares

Definimos as funções lineares em  $\mathbb{R}^2$ , isto é  $p \in \mathcal{P}_1(\mathbb{R}^2)$ , cuja base é dada pelo conjunto de monómios de até 1º grau  $\{1; x; y\}$ , formando:

$$p(x) = a + bx + cy.$$

Considere-se o triângulo de referência  $\hat{T} \subset \mathbb{R}^2$  que é definido pelos seus 3 vértices:

$$\{\hat{v}_1; \hat{v}_2; \hat{v}_3\} = \{(0, 0); (1, 0); (0, 1)\}$$

Para a interpolação iremos considerar, como nós os 3 vértices do triângulo,  $\{z_1; z_2; z_3\} = \{\hat{v}_1; \hat{v}_2; \hat{v}_3\}$ , o que significa que identificamos como variáveis nodais:

$$\nu_j(\phi_i) = \phi_j(z_i) = \delta_{ij}, \quad ij = 1, 2, 3$$

onde podemos definir uma nova base com 3 funções forma:

$$\phi_1(x) = a_1 + b_1x + c_1y; \quad \phi_2(x) = a_2 + b_2x + c_2y; \quad \phi_3(x) = a_3 + b_3x + c_3y$$

que verificam as condições das variáveis nodais para cada base, que neste caso é:

$$\nu_j(\phi_i) = \phi_j(z_i) = \delta_{ij} \Leftrightarrow a_j + b_jz_{i,x} + c_jz_{i,y} = \delta_{ij}, \quad ij = 1, 2, 3$$

combinando isto, obtemos as bases de referência:

$$\begin{cases} \hat{\phi}_1(\hat{x}, \hat{y}) = 1 - \hat{x} - \hat{y} \\ \hat{\phi}_2(\hat{x}, \hat{y}) = \hat{x} \\ \hat{\phi}_3(\hat{x}, \hat{y}) = \hat{y}. \end{cases}, \quad (\hat{x}, \hat{y}) \in \hat{T}$$

### 3.2.7 Integração Numérica

Como as formas bilinear e linear definidas na formulação variacional são normalmente dadas por integrais, pode ser necessário introduzir aproximações no cálculo desses integrais, como a integração de Gauss-Legendre.

#### 3.2.7.1 Integração de Gauss-Legendre a 2D: Domínio Triangular

Consideremos o triângulo de referência  $\hat{T}$ , com 3 vértices:

$$\{\hat{v}_1; \hat{v}_2; \hat{v}_3\} = \{(0, 0); (1, 0); (0, 1)\}$$

notamos que é fácil calcular o integral de monómios sobre este domínio:

$$\begin{aligned} \int_{\hat{T}} x^m y^n dX &= \int_0^1 \int_0^{1-x} x^m y^n dy dx = \int_0^1 x^m \left( \int_0^{1-x} y^n dy \right) dx = \int_0^1 x^m \left[ \frac{y^{n+1}}{n+1} \right]_0^{1-x} dx = \\ &= \int_0^1 x^m \frac{(1-x)^{n+1}}{n+1} dx = \frac{m! n!}{(m+n+2)!} \end{aligned}$$

mas nem sempre são monómios que temos de integrar. E nem sempre é sobre o domínio de referência que temos de integrar.

Considerando a mudança de variável, dada pela transformação linear  $X = G_T(\hat{X}) = [A_T]\hat{X} + (b)$ , descrita em (3.27), finalmente, pelo teorema das mudanças de variável, podemos definir a seguinte mudança de variável no integral:

$$\int_T f(X) dX = \int_{\hat{T}} f(G_T(\hat{X})) \left| \det \left\{ \frac{\partial(x, y)}{\partial(\hat{x}, \hat{y})} \right\} \right| d\hat{X} = |\det[A_T]| \int_{\hat{T}} f(G_T(\hat{X})) d\hat{X}. \quad (3.30)$$

### 3.2.7.2 Transformação do triângulo de referência para o quadrado de referência

No triângulo de referência:

$$\int_{\hat{T}} f(\hat{X}) d\hat{X} = \int_0^1 \int_0^{1-\hat{x}} f(\hat{x}, \hat{y}) d\hat{y} d\hat{x}$$

1) Transformar o integral sobre o triângulo de referência, num integral equivalente, mas na superfície de um quadrado:

$$\{(u, v) \in \mathbb{R}^2 : 0 \leq u, v \leq 1\}$$

através da substituição:

$$\hat{x} = uv ; \hat{y} = u(1-v)$$

considerando a mudança de variável:

$$(\hat{x}, \hat{y}) = P(u, v) = (p_1(u, v), p_2(u, v)) = (uv, u(1-v))$$

cuja jacobiana:

$$\frac{\partial(\hat{x}, \hat{y})}{\partial(u, v)} = \nabla_{(u, v)} P(u, v) = \nabla P(u, v) = DP(u, v) = \begin{bmatrix} \frac{\partial p_1}{\partial u} & \frac{\partial p_1}{\partial v} \\ \frac{\partial p_2}{\partial u} & \frac{\partial p_2}{\partial v} \end{bmatrix} = \begin{bmatrix} v & u \\ (1-v) & -u \end{bmatrix}$$

e o determinante da jacobiana:

$$\det \left\{ \frac{\partial(\hat{x}, \hat{y})}{\partial(u, v)} \right\} = \det \{ \nabla P(u, v) \} = -vu - u(1-v) = -u$$

e a área diferencial:

$$d\hat{y}d\hat{x} = \left| \det \begin{Bmatrix} \partial(\hat{x}, \hat{y}) \\ \partial(u, v) \end{Bmatrix} \right| du dv = |u| du dv$$

portanto:

$$\int_0^1 \int_0^{1-\hat{x}} f(\hat{x}, \hat{y}) d\hat{y} d\hat{x} = \int_0^1 \int_0^1 f(uv, u(1-v)) |u| du dv$$

2) Transformar a integral sobre  $[0; 1]^2$  num integral sobre o quadrado de referência:

$$\{(\xi, \eta) \in \mathbb{R}^2 : -1 \leq \xi, \eta \leq 1\}$$

através da substituição:

$$u = (1 + \xi)/2 ; \quad v = (1 + \eta)/2$$

considerando a mudança de variável:

$$(u, v) = H(\xi, \eta) = (h_1(\xi, \eta); h_2(\xi, \eta)) = ((1 + \xi)/2; (1 + \eta)/2)$$

cuja jacobiana:

$$\frac{\partial(u, v)}{\partial(\xi, \eta)} = DH(\xi, \eta) = \begin{bmatrix} \frac{\partial h_1}{\partial \xi} & \frac{\partial h_1}{\partial \eta} \\ \frac{\partial h_2}{\partial \xi} & \frac{\partial h_2}{\partial \eta} \end{bmatrix} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}$$

e o determinante da jacobiana:

$$\det \left\{ \frac{\partial(u, v)}{\partial(\xi, \eta)} \right\} = \det \{ \nabla H(\xi, \eta) \} = 1/4$$

e a área diferencial:

$$du dv = \left| \det \left\{ \frac{\partial(u, v)}{\partial(\xi, \eta)} \right\} \right| d\xi d\eta = \frac{1}{4} d\xi d\eta$$

finalmente ficamos com a expressão:

$$\begin{aligned} \int_0^1 \int_0^{1-\hat{x}} f(\hat{x}, \hat{y}) d\hat{y} d\hat{x} &= \int_0^1 \int_0^1 f(uv, u(1-v)) |u| du dv = \\ &= \int_{-1}^1 \int_{-1}^1 f \left( \frac{(1 + \xi)(1 + \eta)}{4}, \frac{(1 + \xi)(1 - \eta)}{4} \right) \frac{|(1 + \xi)|}{8} d\xi d\eta. \end{aligned}$$

Substituindo a integração em cada derivada pela formulação de somatório da quadratura de Gauss-Legendre obtemos:

$$I(f) = \sum_{j=1}^{n_y} \sum_{i=1}^{n_x} \frac{|(1 + \xi_i)|}{8} w_j w_i f \left( \frac{(1 + \xi_i)(1 + \eta_j)}{4}, \frac{(1 + \xi_i)(1 - \eta_j)}{4} \right)$$

onde  $\xi_i$  e  $\eta_j$ , para  $i = 1, \dots, n_x$  e  $j = 1, \dots, n_y$ , são as raízes (ou nós) dos polinómios de Gauss-

Legendre  $P_{n_x}$  e  $P_{n_y}$ , e  $w_i$  tal como  $w_j$  são os respetivos pesos.

3) Ter em consideração que podemos não estar no triângulo de referência inicialmente, tentemos simplificar as coisas ainda de uma outra maneira:

$$\begin{aligned} \int_T f(X) dX &= |A_T| \int_{\hat{T}} f(G(\hat{X})) d\hat{X} = |A_T| \int_0^1 \int_0^1 f(G(P(u, v))) |u| du dv = \\ &= |A_T| \int_{-1}^1 \int_{-1}^1 f(G(P(H(\xi, \eta)))) \frac{|(1 + \xi)|}{8} d\xi d\eta = |A_T| \int_{-1}^1 \int_{-1}^1 f(G(z)) \frac{|(1 + \xi)|}{8} d\xi d\eta \end{aligned}$$

sendo

$$G(z) = [A_T](z) + (b_T), \quad z = \left( \frac{(1 + \xi)(1 + \eta)}{4}, \frac{(1 + \xi)(1 - \eta)}{4} \right)$$

finalmente

$$I(f) = |A_T| \sum_{j=1}^{n_y} \sum_{i=1}^{n_x} \frac{|(1 + \xi_i)|}{8} w_j w_i f(G(z_{ij}))$$

com

$$z_{ij} = \left( \frac{(1 + \xi_i)(1 + \eta_j)}{4}, \frac{(1 + \xi_i)(1 - \eta_j)}{4} \right).$$

### 3.2.7.3 Algumas considerações sobre $\mathcal{P}_1$

Como vimos, podemos calcular um integral definido num triângulo arbitrário  $T$ , em termos de um integral definido no triângulo de referência,  $\hat{T}$ , integral este que ainda o podemos aproximar por uma quadratura:

$$\int_T f(x, y) dX = \int_{\hat{T}} f(G(\hat{x}, \hat{y})) \cdot |\det\{[A_T]\}| d\hat{X} \approx \sum_{i=1}^N w_i \cdot f(\hat{t}_i) \cdot |\det\{[A_T]\}|$$

onde o conjunto de nós (ou pontos) da quadratura:

$$(\hat{t}) = \{\hat{t}_1, \dots, \hat{t}_N\} = \{(\hat{t}_{1,x}, \hat{t}_{1,y}), \dots, (\hat{t}_{N,x}, \hat{t}_{N,y})\}$$

e o conjunto de pesos da quadratura:

$$(w) = \{w_1, \dots, w_N\}.$$

Tentemos agora aproximar alguns integrais que nos vão ser úteis ao longo deste trabalho.

- Comecemos por calcular o integral de  $b_i$ , num triângulo genérico,  $T$ , da nossa malha:

$$b_i = \int_T f(x, y) \phi_i(x, y) dx dy = \int_{\hat{T}} f(G(\hat{x}, \hat{y})) \cdot \phi_i(G(\hat{x}, \hat{y})) \cdot |\det\{[A_T]\}| d\hat{A}.$$

Cada base  $\phi_i(G(\hat{x}, \hat{y}))$ , da equação anterior, é um polinómio de grau definido em  $\hat{T}$  que é igual a 1 num

dos vértices de  $\hat{T}$  e zero nos outros dois vértices de  $\hat{T}$ . Portanto  $\phi_i(G(\hat{x}, \hat{y}))$ , coincide com uma das três seguintes funções de base definidas no triângulo de referência  $\hat{T}$ :

$$\Psi_1(\hat{x}, \hat{y}) = 1 - \hat{x} - \hat{y} ; \quad \Psi_2(\hat{x}, \hat{y}) = \hat{x} ; \quad \Psi_3(\hat{x}, \hat{y}) = \hat{y}$$

Cada uma das funções de base  $\Psi_1, \Psi_2$  e  $\Psi_3$  é igual a 1 no vértice correspondente  $\hat{v}_1, \hat{v}_2$  e  $\hat{v}_3$  e igual a zero nos outros dois vértices do triângulo de referência. Assim, o nosso integral:

$$\begin{aligned} b_i &= \int_T f(x, y) \phi_i(x, y) dx dy = \\ &= \int_{\hat{T}} f(G(\hat{x}, \hat{y})) \cdot \phi_i(G(\hat{x}, \hat{y})) \cdot |\det\{[A_T]\}| d\hat{X} = \\ &= \int_{\hat{T}} f(G(\hat{x}, \hat{y})) \cdot \Psi_k(\hat{x}, \hat{y}) \cdot |\det\{[A_T]\}| d\hat{X} = \\ &\approx \sum_{n=1}^N w_n \cdot f(\hat{t}_n) \cdot \Psi_k(\hat{t}_n) \cdot |\det\{[A_T]\}| \end{aligned} \tag{3.31}$$

para algum  $k \in \{1, 2, 3\}$ , correspondendo ao vértice  $v_k$ , onde  $\phi_i$  for igual a 1.

- De igual modo o integral:

$$\begin{aligned} I_{ij} &= \int_T f(x, y) \phi_i(x, y) \phi_j(x, y) dx dy = \\ &= \int_{\hat{T}} f(G(\hat{x}, \hat{y})) \cdot \phi_i(G(\hat{x}, \hat{y})) \cdot \phi_j(G(\hat{x}, \hat{y})) \cdot |\det\{[A_T]\}| d\hat{A} \\ &= \int_{\hat{T}} f(G(\hat{x}, \hat{y})) \cdot \Psi_k(\hat{x}, \hat{y}) \cdot \Psi_l(\hat{x}, \hat{y}) \cdot |\det\{[A_T]\}| d\hat{A} \approx \\ &\approx \sum_{n=1}^N w_i \cdot f(\hat{t}_n) \cdot \Psi_k(\hat{t}_n) \cdot \Psi_l(\hat{t}_n) \cdot |\det\{[A_T]\}| \end{aligned} \tag{3.32}$$

para  $k, l \in \{1, 2, 3\}$ , correspondendo aos vértices  $v_k$  e  $v_l$ , onde  $\phi_i$  e  $\phi_j$  forem iguais a 1 respetivamente.

**Teorema 7.** Seja  $T \in \mathcal{T}$ ,  $g \in C^1(T)$  e  $\hat{g}(\hat{x}, \hat{y}) := g(G(\hat{x}, \hat{y}))$ , onde  $G$  é definido como sendo a transformação linear,  $X = G_T(\hat{X}) = [A_T]\hat{X} + (b)$ , que mapeia o triângulo de referência,  $\hat{T}$ , para um triângulo genérico  $T$ . Então,

$$\hat{\nabla} \hat{g}(\hat{x}, \hat{y}) = [A_T]^T \nabla g(G(\hat{x}, \hat{y}))$$

onde  $\hat{\nabla}$  é o gradiente com respeito as coordenadas  $\hat{x}$  e  $\hat{y}$ .

**Demosntração:** (Ver [19]) Seja  $g := g(x, y)$ , e  $\hat{g}(\hat{x}, \hat{y}) = g(G(\hat{x}, \hat{y}))$ , e denotemos

$$G(\hat{x}, \hat{y}) = \left( \underbrace{(A_T)_{1,1}\hat{x} + (A_T)_{1,2}\hat{y} + b_{T1}}_{:=x(\hat{x}, \hat{y})}, \underbrace{(A_T)_{2,1}\hat{x} + (A_T)_{2,2}\hat{y} + b_{T2}}_{:=y(\hat{x}, \hat{y})} \right)$$

uma das suas derivadas parciais, usando a regra da cadeia:

$$\frac{\partial \hat{g}(\hat{x}, \hat{y})}{\partial \hat{x}} = \frac{\partial g}{\partial x}(G(\hat{x}, \hat{y})) \frac{\partial x}{\partial \hat{x}}(\hat{x}, \hat{y}) + \frac{\partial g}{\partial y}(G(\hat{x}, \hat{y})) \frac{\partial y}{\partial \hat{x}}(\hat{x}, \hat{y}) = \frac{\partial g}{\partial x}(G(\hat{x}, \hat{y}))(A_T)_{1,1} + \frac{\partial g}{\partial y}(G(\hat{x}, \hat{y}))(A_T)_{2,1}$$

e de forma similar:

$$\frac{\partial \hat{g}(\hat{x}, \hat{y})}{\partial \hat{y}} = \frac{\partial g}{\partial x}(G(\hat{x}, \hat{y})) \frac{\partial x}{\partial \hat{y}}(\hat{x}, \hat{y}) + \frac{\partial g}{\partial y}(G(\hat{x}, \hat{y})) \frac{\partial y}{\partial \hat{y}}(\hat{x}, \hat{y}) = \frac{\partial g}{\partial x}(G(\hat{x}, \hat{y}))(A_T)_{1,2} + \frac{\partial g}{\partial y}(G(\hat{x}, \hat{y}))(A_T)_{2,2}$$

desta forma:

$$\hat{\nabla} \hat{g}(\hat{x}, \hat{y}) = \begin{bmatrix} \frac{\partial \hat{g}(\hat{x}, \hat{y})}{\partial \hat{x}} \\ \frac{\partial \hat{g}(\hat{x}, \hat{y})}{\partial \hat{y}} \end{bmatrix} = \begin{bmatrix} (A_T)_{1,1} & (A_T)_{2,1} \\ (A_T)_{1,2} & (A_T)_{2,2} \end{bmatrix} \begin{bmatrix} \frac{\partial g}{\partial x}(G(\hat{x}, \hat{y})) \\ \frac{\partial g}{\partial y}(G(\hat{x}, \hat{y})) \end{bmatrix} = [A_T]^T \nabla g(G(\hat{x}, \hat{y}))$$

□

- Calculemos o seguinte integral:

$$Ig_{ij} = \int_T (\nabla \phi_i \nabla \phi_j) dx dy = \int_{\hat{T}} \nabla \phi_i(G(\hat{x}, \hat{y})) \cdot \nabla \phi_i(G(\hat{x}, \hat{y})) \cdot |\det\{[A_T]\}| d\hat{A}$$

Pelo teorema anterior, sabemos que:

$$\hat{\nabla} \hat{\phi}(\hat{x}, \hat{y}) = [A_T]^T \nabla \phi(G(\hat{x}, \hat{y})) \Leftrightarrow \nabla \phi(G(\hat{x}, \hat{y})) = ([A_T]^T)^{-1} \hat{\nabla} \hat{\phi}(\hat{x}, \hat{y})$$

portanto o integral

$$\begin{aligned} Ig_{ij} &= \int_T (\nabla \phi_i \nabla \phi_j) dx dy = \int_{\hat{T}} \nabla \phi_i(G(\hat{x}, \hat{y})) \cdot \nabla \phi_i(G(\hat{x}, \hat{y})) \cdot |\det\{[A_T]\}| d\hat{A} = \\ &= \int_{\hat{T}} \left[ ([A_T]^T)^{-1} \hat{\nabla} \hat{\phi}_i(\hat{x}, \hat{y}) \right] \cdot \left[ ([A_T]^T)^{-1} \hat{\nabla} \hat{\phi}_j(\hat{x}, \hat{y}) \right] |\det\{[A_T]\}| d\hat{A} = \\ &= \int_{\hat{T}} \left[ ([A_T]^T)^{-1} \hat{\nabla} \hat{\phi}_i \right]^T \left[ ([A_T]^T)^{-1} \hat{\nabla} \hat{\phi}_j \right] |\det\{[A_T]\}| d\hat{A} = \\ &= \int_{\hat{T}} \left[ \hat{\nabla} \hat{\phi}_i^T [A_T]^{-1} \right] \left[ ([A_T]^T)^{-1} \hat{\nabla} \hat{\phi}_j \right] |\det\{[A_T]\}| d\hat{A} = \\ &= \left[ \hat{\nabla} \hat{\phi}_i^T [A_T]^{-1} ([A_T]^T)^{-1} \hat{\nabla} \hat{\phi}_j \right] |\det\{[A_T]\}| \int_{\hat{T}} 1 d\hat{A} = \\ &= \left[ \hat{\nabla} \hat{\phi}_i^T [A_T]^{-1} ([A_T]^T)^{-1} \hat{\nabla} \hat{\phi}_j \right] |\det\{[A_T]\}| \frac{1}{2} \end{aligned}$$

isto, levando em consideração que  $\hat{\phi}$  são polinómios de grau 1, onde  $\hat{\phi}(\hat{x}, \hat{y}) = \Psi(\hat{x}, \hat{y})$ , e os seus gradientes são apenas vetores:

$$\hat{\nabla} \Psi(\hat{x}, \hat{y}) = \begin{cases} \hat{\nabla} \Psi_1 = (-1; -1) \\ \hat{\nabla} \Psi_2 = (1; 0) \\ \hat{\nabla} \Psi_3 = (0; 1) \end{cases}$$

então finalmente, e depois de alguns cálculos:

$$I = \left[ \hat{\nabla} \Psi_k^T [A_T]^{-1} ([A_T]^T)^{-1} \hat{\nabla} \Psi_l \right] | \det\{[A_T]\} | \frac{1}{2}$$

para  $k, l \in \{1, 2, 3\}$ , correspondendo aos vértices  $v_k$  e  $v_l$ , onde  $\phi_i$  e  $\phi_j$  forem iguais a 1 respetivamente.

Assumindo que :

$$[A_T] = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix}$$

e o seu determinante dado por:

$$\det\{[A_T]\} = a_{1,1}a_{2,2} - a_{2,1}a_{1,2}$$

então a matriz

$$[A_T]^{-1} ([A_T]^T)^{-1} = \frac{1}{a_{1,2}a_{2,1} - a_{1,1}a_{2,2}} \begin{bmatrix} a_{1,2}^2 + a_{2,2}^2 & -a_{1,1}a_{1,2} + a_{2,1}a_{2,2} \\ -a_{1,1}a_{1,2} + a_{2,1}a_{2,2} & a_{1,1}^2 + a_{2,1}^2 \end{bmatrix}.$$

# Capítulo 4

## Resolução Numérica de Problemas do Obstáculo

### 4.1 Problema de Complementariedade Linear

Ao longo deste trabalho, ir-nos-emos se deparar com um **Problema de Complementariedade Linear** (LCP, sigla do inglês significa *Linear Complementarity Problem*), na forma matricial, após fazer a discretização dos métodos. Aqui abordaremos essa questão.

Dado  $\mathbf{A} \in \mathbb{R}^{N \times N}$  e  $\mathbf{f}, \mathbf{g} \in \mathbb{R}^N$ , temos o seguinte *problema de complementariedade* algébrico, no qual queremos descobrir  $\mathbf{x} \in \mathbb{R}^N$ , que verifique:

$$\begin{cases} \mathbf{Ax} - \mathbf{f} \leq \mathbf{0} \\ \mathbf{g} - \mathbf{x} \leq \mathbf{0} \\ (\mathbf{Ax} - \mathbf{f})^T(\mathbf{g} - \mathbf{x}) = 0 \end{cases} \quad (4.1)$$

Quando estamos a dizer que  $\mathbf{Ax} - \mathbf{f} \leq \mathbf{0}$  e que  $\mathbf{g} - \mathbf{x} \leq \mathbf{0}$ , estamos a comparar vetores entrada a entrada ou seja, estamos a assumir que todas as entradas de cada vetor são iguais ou menores que zero:

$$\mathbf{V} = \mathbf{Ax} - \mathbf{f} = \begin{bmatrix} v_1 \leq 0 \\ \vdots \\ v_N \leq 0 \end{bmatrix}, \quad \mathbf{U} = \mathbf{g} - \mathbf{x} = \begin{bmatrix} u_1 \leq 0 \\ \vdots \\ u_N \leq 0 \end{bmatrix}$$

Portanto, para o produto interno entre  $\mathbf{U} \leq \mathbf{0}$  e  $\mathbf{V} \leq \mathbf{0}$  ser zero, então significa que: para cada entrada  $i$ , ou  $v_i$  é zero, ou  $u_i$  é zero:

$$(\mathbf{Ax} - \mathbf{f})^T(\mathbf{g} - \mathbf{x}) = 0 \Leftrightarrow \mathbf{U}^T \mathbf{V} = 0 \Leftrightarrow \sum_{i=1}^N v_i u_i = 0$$

$$\implies v_i = 0 \vee u_i = 0, \quad \forall i = \{1, \dots, N\}$$

o que é equivalente a dizer que o máximo entre dois vetores (entrada a entrada) tem de ser um vetor de zeros, e portanto

$$\begin{cases} \mathbf{V} \leq \mathbf{0} \\ \mathbf{U} \leq \mathbf{0} \\ \mathbf{U}^T \mathbf{V} = 0 \end{cases} \Leftrightarrow \max\{\mathbf{V}, \mathbf{U}\} = \begin{bmatrix} \max\{v_1, u_1\} \\ \vdots \\ \max\{v_N, u_N\} \end{bmatrix} = \mathbf{0}$$

portanto, reescrevemos o sistema (4.1) numa condição de máximo:

$$\max\{\mathbf{Ax} - \mathbf{f}, \mathbf{g} - \mathbf{x}\} = \mathbf{0}. \quad (4.2)$$

Isolando  $\mathbf{x}$ , aplicando diretamente a condição de máximo, obtemos  $\mathbf{x} = \max\{[-\mathbf{A}]^{-1}(-\mathbf{f}); \mathbf{g}\}$ , pelo que neste caso estamos a assumir que ocorre preservação da desigualdade linha a linha, ou seja  $\mathbf{Ax} \leq \mathbf{f} \Leftrightarrow \mathbf{x} \leq \mathbf{A}^{-1}\mathbf{f}$ , o que pode não ser verdade. Portanto alguns cuidados devem ser tomados, pois aplicar diretamente a condição de máximo pode resultar na obtenção de valores incorretos, a forma de colmatar essa situação é através de algum método iterativo, como o Semi-smooth Newton method, ou um outro algoritmo iterativo projetado como o Jacobi, Gauss-Seidel, ou SOR, dentro de certas condições.

#### 4.1.1 Algoritmo iterativo: Semi-smooth Newton Method (SSNM)

O método de Newton semi-suave (a sigla SSNM, significa do inglês *Semi-smooth Newton Method*), é uma extensão do método de Newton clássico, projetada para resolver problemas onde as funções envolvidas podem não ser totalmente diferenciáveis, ou problemas de equações não lineares (cf. [20]). Dado um operador  $G : \mathbb{R}^N \rightarrow \mathbb{R}^N$ , o método de Newton é usado para encontrar a raiz  $u^*$ , de um sistema de equações  $G(u^*) = 0$ . Denotemos  $\nabla G : \mathbb{R}^N \rightarrow \mathbb{R}^{N \times N}$  a matriz jacobiana de  $G$ . O algoritmo é dado por:

(i) Inicializar:  $\mathbf{u}^{(0)} \in \mathbb{R}^N$  e  $k = 0$ ;

(ii) Iterar

$$\mathbf{u}^{k+1} = \mathbf{u}^k - [\nabla_u G(\mathbf{u}^k)]^{-1}G(\mathbf{u}^k)$$

(iii) Verificar condição de paragem: se  $\|\mathbf{u}^{(k+1)} - \mathbf{u}^{(k)}\| \leq \text{TOL}$ , termine;

(iv) Caso contrário, incrementa-se  $k = k + 1$  e volta-se ao ponto (ii).

Dado  $\varphi(a, b) = \min(a, b)$ , define-se o gradiente generalizado desta função como sendo

$$\nabla_{(a,b)}\varphi(a, b) = \left[ \frac{\partial \varphi}{\partial a}, \frac{\partial \varphi}{\partial b} \right]^T = \begin{cases} [1, 0]^T & , a < b \\ [0, 1]^T & , a > b \\ \lambda[1, 0]^T + (1 - \lambda)[0, 1]^T & , a = b \end{cases}$$

com  $\lambda \in [0, 1]$ . O caso  $a = b$  corresponde ao *convex hull* do gradiente dos outros 2 casos.

Seja  $u \in \mathbb{R}$ , o operador  $G(u)$  dado por:

$$G(u) = \varphi(a(u), b(u)) = \min\{a(u), b(u)\}$$

e a sua derivada em ordem a  $u$ , pela regra da cadeia

$$\frac{\partial G(u)}{\partial u} = \frac{\partial \varphi}{\partial a} \frac{\partial a}{\partial u} + \frac{\partial \varphi}{\partial b} \frac{\partial b}{\partial u}$$

No caso de  $\mathbf{u} \in \mathbb{R}^N$ , o operador  $G(\mathbf{u}) \in \mathbb{R}^N$ , é dado por:

$$G(\mathbf{u}) = \varphi(a(\mathbf{u}), b(\mathbf{u})) = \begin{bmatrix} \varphi(a_1(\mathbf{u}), b_1(\mathbf{u})) \\ \vdots \\ \varphi(a_N(\mathbf{u}), b_N(\mathbf{u})) \end{bmatrix} = \begin{bmatrix} \min\{a_1(\mathbf{u}), b_1(\mathbf{u})\} \\ \vdots \\ \min\{a_N(\mathbf{u}), b_N(\mathbf{u})\} \end{bmatrix} = \begin{bmatrix} G_1(a, b) \\ \vdots \\ G_N(a, b) \end{bmatrix}$$

sendo  $a = (a_1, \dots, a_N)$  e  $b = (b_1, \dots, b_N)$ . Para calcular o gradiente (jacobiana) generalizada de  $G(\mathbf{u}) \in \mathbb{R}^N$ , ou seja  $\nabla_u G(\mathbf{u}) \in \mathbb{R}^{N \times N}$  usou-se a regra da cadeia (Matrix Version of Chain Rule):

$$\nabla_u G(\mathbf{u}) = \nabla_a G(\mathbf{u}) \nabla_u a(\mathbf{u}) + \nabla_b G(\mathbf{u}) \nabla_u b(\mathbf{u})$$

onde  $\nabla_a G(\mathbf{u}) = D_a$  e  $\nabla_b G(\mathbf{u}) = D_b$  serão matrizes diagonais, para as quais a diagonal irá assumir os valores de 0 ou 1, dependendo de que entrada o mínimo corresponde. Para uma entrada genérica  $(i, i)$  da diagonal:

$$D_a^{(i,i)} = \begin{cases} 1 & , \min\{a_i, b_i\} = a_i \\ 0 & , \min\{a_i, b_i\} = b_i \end{cases}; \quad D_b^{(i,i)} = \begin{cases} 0 & , \min\{a_i, b_i\} = a_i \\ 1 & , \min\{a_i, b_i\} = b_i. \end{cases}$$

No caso do nosso problema nós temos:

$$G(\mathbf{u}) := \varphi\{-\mathbf{A}\mathbf{u} + \mathbf{f}; \mathbf{u} - \mathbf{g}\} = \min\{-\mathbf{A}\mathbf{u} + \mathbf{f}; \mathbf{u} - \mathbf{g}\} = \mathbf{0}$$

e portanto

$$\nabla_u a(\mathbf{u}) = \nabla_u \{-\mathbf{A}\mathbf{u} + \mathbf{f}\} = -\mathbf{A}$$

$$\nabla_u b(\mathbf{u}) = \nabla_u \{\mathbf{u} - \mathbf{g}\} = \mathbf{I}$$

e finalmente  $\nabla_u G(\mathbf{u}) = -D_a \mathbf{A} + D_b$ , não dependendo de  $\mathbf{u}$ . Este algoritmo está descrito com detalhe no apêndice A em A.1.

## 4.2 Problema do Obstáculo Elíptico: Membrana Elástica

Os problemas que iremos abordar, serão problemas cuja solução está definida num domínio regular  $\Omega \subset \mathbb{R}^2$ . A formulação forte do problema é dada por:

$$\begin{cases} -\Delta u - f \geq 0 & , x \in \Omega \\ u \geq g & , x \in \Omega \\ (u - g)(-\Delta u - f) = 0 & , x \in \Omega \\ u = 0 & , x \in \partial\Omega \end{cases} \quad (4.3)$$

tal como visto no capítulo 2.

### 4.2.1 Discretização com Diferenças finitas

Normalmente, FDM é usado em domínios retangulares, o que nos permite fazer uma discretização uniforme em cada eixo, cobrindo todo o domínio, para  $(x, y) \in [a, b] \times [c, d]$  para quaisquer  $a, b, c, d \in \mathbb{R}$ . Mas se por exemplo o domínio  $\Omega \subset \mathbb{R}^2$  for um círculo, desenhar a malha onde iremos discretizar o nosso problema, exige mais cuidado, principalmente na região do bordo. Então precisamos de uma outra notação.

Para tal começou-se por fazer uma discretização regular dum domínio que contém o domínio  $\Omega$ . Definiu-se os pontos do interior de  $\Omega$ , como:  $\mathbf{P}_I$ , com um total de  $N_I$  pontos no interior.

Para marcar os pontos da fronteira, prolongou-se todas as linhas paralelas horizontais e verticais, que passam pelos pontos da discretização do interior, e intercetou-se com o bordo  $\partial\Omega$ , permitindo formar o conjunto:  $\mathbf{P}_B$ , com um total de  $N_B$  pontos na fronteira.

Consideremos todos os pontos da discretização, com um total de  $N = N_I + N_B$  pontos ordenados, como sendo um conjunto:

$$\mathbf{P} = \mathbf{P}_I \cup \mathbf{P}_B = \{(x_1, y_1); (x_2, y_2); \dots; (x_N, y_N)\}.$$

Considere-se  $\mathbf{i}_I$ , como sendo o conjunto de índices que indexam todos os pontos do interior, do conjunto  $\mathbf{P}$ , isto é:

$$\mathbf{P}(\mathbf{i}_I) = \mathbf{P}_I$$

e considere-se  $\mathbf{i}_B$ , como sendo o conjunto de índices que indexam todos os pontos da fronteira, do conjunto  $\mathbf{P}$ , isto é:

$$\mathbf{P}(\mathbf{i}_B) = \mathbf{P}_B.$$

Para um qualquer ponto,  $p_i = (x_i, y_i) \in \mathbf{P}$ , considere-se a solução avaliada nesse ponto  $u(p_i) = u(x_i, y_i)$ , e a sua aproximação como sendo  $u(p_i) \approx u_i$ .

Considere agora o ponto  $p_{i+\alpha_x}$  (ou sem perda de generalidade o ponto  $p_{i+\alpha_y}$ ), com  $\alpha_x \in \mathbb{Z}$ . O ponto  $p_{i+\alpha_x}$  terá todas as coordenadas iguais a  $p_i$ , exceto a coordenada do eixo  $x$ , pois o ponto  $p_{i+\alpha_x}$  corresponderá ao  $|\alpha_x|$ -ésimo ponto mais próximo de  $p_i$ , relativamente à coordenada  $x$ , no sentido positivo (à sua direita) se  $\alpha_x > 0$ , ou no sentido negativo (à sua esquerda) se  $\alpha_x < 0$ . Assim o ponto é dado por:

$$p_{i+\alpha_x} = \begin{cases} p_j = & |\alpha_x|^o \arg \min_{p_j \in P} |x_i - x_j| \\ \text{tal que} & \\ & y_j = y_i \\ & x_j > x_i, \text{ se } \alpha_x > 0 \\ & x_j < x_i, \text{ se } \alpha_x < 0 \end{cases} \quad (4.4)$$

onde a notação  $|\alpha_x|^o \arg \min$  representa o  $|\alpha_x|$ -ésimo menor argumento. Considere-se como sendo o passo no eixo  $x$ :  $h_{i,\alpha}^x = |x_i - x_{i+\alpha_x}|$ , e no eixo  $y$ :  $h_{i,\alpha}^y = |y_i - y_{i+\alpha_y}|$ . Pelas diferenças finitas, considere-se a discretização do Laplaciano:

$$\begin{aligned}\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} &\approx \frac{2}{(h_{i,-1}^x)^2 + (h_{i,-1}^x h_{i,1}^x)} u_{i-1,x} + \frac{-2}{h_{i,-1}^x h_{i,1}^x} u_i + \frac{2}{(h_{i,1}^x)^2 + (h_{i,-1}^x h_{i,1}^x)} u_{i+1,x} + \\ &+ \frac{2}{(h_{i,-1}^y)^2 + (h_{i,-1}^y h_{i,1}^y)} u_{i-1,y} + \frac{-2}{h_{i,-1}^y h_{i,1}^y} u_i + \frac{2}{(h_{i,1}^y)^2 + (h_{i,-1}^y h_{i,1}^y)} u_{i+1,y} = \quad (4.5) \\ &= (a_i^x u_{i-1,x} + b_i^x u_i + c_i^x u_{i+1,x}) + (a_i^y u_{i-1,y} + b_i^y u_i + c_i^y u_{i+1,y}) = \\ &= [a_i^x] u_{i-1,x} + [a_i^y] u_{i-1,y} + (b_i^x + b_i^y) u_i + [c_i^y] u_{i+1,y} + [c_i^x] u_{i+1,x}\end{aligned}$$

discretizando o operador, para os pontos do interior, portanto a discretização:

$$-\Delta u(p_i) - f(p_i), \forall i \in \mathbf{i}_I$$

escreve-se na forma matricial:

$$-\mathbf{A}\mathbf{u} - \mathbf{f}$$

onde  $\mathbf{u} \in \mathbb{R}^N$ , é o vetor solução:

$$\mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_N \end{bmatrix} = \begin{bmatrix} u(p_1) \\ \vdots \\ u(p_N) \end{bmatrix} \in \mathbb{R}^N \quad (4.6)$$

A matriz  $\mathbf{A} \in \mathbb{R}^{N \times N}$  é uma matriz esparsa, onde cada linha  $i$  estará associada ao valor de solução no ponto  $p_i = (x_i, y_i)$ , e nessa mesma linha  $i$ , a cada coluna  $j$ , e se temos  $a_{ij} \neq 0$ , quer dizer que  $p_i$  e  $p_j$  são pontos vizinhos, e que a discretização do laplaciano em torno de  $p_i$ , contempla o valor de  $u_j$ . No entanto repare-se que a solução  $u$  nos pontos da fronteira já é conhecidos por nós, e só não sabemos os pontos do interior, portanto podemos remover do sistema todas as linhas que não correspondem ao interior do domínio (sendo a notação ":" , que representa todas as linhas ou todas as colunas):

$$-\mathbf{A}(\mathbf{i}_I, :) \mathbf{u} - \mathbf{f}(\mathbf{i}_I)$$

mas precisamos de um sistema com uma matriz quadrada, para podermos trabalhar, e que contenha apenas todas as soluções que não sabemos:

$$-\mathbf{A}(\mathbf{i}_I, :) \mathbf{u} - \mathbf{f}(\mathbf{i}_I) = -[\mathbf{A}(\mathbf{i}_I, \mathbf{i}_I) \mathbf{u}(\mathbf{i}_I) + \mathbf{A}(\mathbf{i}_I, \mathbf{i}_B) \mathbf{u}(\mathbf{i}_B)] - \mathbf{f}(\mathbf{i}_I).$$

Isto permite-nos escrever o **problema de complementaridade** para o problema do obstáculo

$$\begin{cases} -\mathbf{A}\mathbf{u} - \mathbf{f} \geq \mathbf{0} \\ \mathbf{u} - \mathbf{g} \geq \mathbf{0} \\ (\mathbf{u} - \mathbf{g})^T(\mathbf{A}\mathbf{u} + \mathbf{f}) = 0 \end{cases} \Leftrightarrow \begin{cases} \mathbf{A}\mathbf{u} + \mathbf{f} \leq \mathbf{0} \\ \mathbf{g} - \mathbf{u} \leq \mathbf{0} \\ (\mathbf{g} - \mathbf{u})^T(\mathbf{A}\mathbf{u} + \mathbf{f}) = 0. \end{cases} \quad (4.7)$$

Como queremos apenas descobrir os pontos do interior:

$$\begin{cases} \mathbf{A}(\mathbf{i}_I, \mathbf{i}_I)\mathbf{u}(\mathbf{i}_I) + \mathbf{A}(\mathbf{i}_I, \mathbf{i}_B)\mathbf{u}(\mathbf{i}_B) + \mathbf{f}(\mathbf{i}_I) \leq \mathbf{0} \\ \mathbf{g}(\mathbf{i}_I) - \mathbf{u}(\mathbf{i}_I) \leq \mathbf{0} \\ (\mathbf{g}(\mathbf{i}_I) - \mathbf{u}(\mathbf{i}_I))^T(\mathbf{A}(\mathbf{i}_I, \mathbf{i}_I)\mathbf{u}(\mathbf{i}_I) + \mathbf{A}(\mathbf{i}_I, \mathbf{i}_B)\mathbf{u}(\mathbf{i}_B) + \mathbf{f}(\mathbf{i}_I)) = 0 \end{cases} \quad (4.8)$$

sendo  $\mathbf{x} = \mathbf{u}(\mathbf{i}) \in \mathbb{R}^{N_I}$ ,  $\mathbf{M} = \mathbf{A}(\mathbf{i}_I, \mathbf{i}_I) \in \mathbb{R}^{N_I \times N_I}$  e  $\mathbf{b} = \mathbf{A}(\mathbf{i}_I, \mathbf{i}_N) + \mathbf{f}(\mathbf{i}_I) \in \mathbb{R}^{N_I}$  e  $\mathbf{h} = \mathbf{g}(\mathbf{i}_I) \in \mathbb{R}^{N_I}$ , reescrevemos o problema de complementaridade como encontrar  $\mathbf{x}$  tal que:

$$\begin{cases} \mathbf{Mx} + \mathbf{b} \leq \mathbf{0} \\ \mathbf{h} - \mathbf{x} \leq \mathbf{0} \\ (\mathbf{h} - \mathbf{x})^T(\mathbf{Mu} + \mathbf{b}) = 0 \end{cases} \quad (4.9)$$

e ainda a sua expressão de mínimo/máximo equivalente:

$$\max\{\mathbf{Mx} + \mathbf{b}; \mathbf{h} - \mathbf{x}\} = \mathbf{0} \quad (4.10)$$

o que pode ser resolvido com o Semi-smooth Newton Method descrito em A.1.

#### 4.2.2 Discretização com Elementos Finitos

Para o problema do obstáculo discretizado com FEM, iremos usar a sua formulação com multiplicadores de Lagrange:

$$\begin{cases} -\Delta u - \lambda = f & , x \in \Omega \\ g - u \leq 0 & , x \in \Omega \\ \lambda \geq 0 & , x \in \Omega \\ (u - g)\lambda = 0 & , x \in \Omega \\ u = 0 & , x \in \partial\Omega \end{cases} \quad (4.11)$$

Os espaços de elementos finitos que iremos considerar terão como base uma triangulação de forma regular, denotada por  $\mathcal{C}_h$  de  $\Omega$ , que daqui em diante assumimos ser poligonal. Denotamos por  $\mathcal{E}_h$  as arestas internas de  $\Omega$ . E os subespaços de elementos finitos são:  $V_h \subset V$  e  $Q_h \subset Q$ .

Além disso definimos

$$\Lambda_h = \{\mu_h \in Q_h : \mu_h \geq 0 \text{ } x \in \Omega\} \subset \Lambda. \quad (4.12)$$

Quando  $Q_h$  é constituído por polinómios de grau 2 ou superior a condição  $\mu_h \geq 0$  é difícil de satisfazer na prática.

A nossa análise é baseada na seguinte condição de estabilidade. Observe que frequentemente escrevemos  $a \gtrsim b$  (ou  $a \lesssim b$ ) quando  $a \geq Cb$  (ou  $a \leq Cb$ ) para alguma constante positiva  $C$  independente da malha de elementos finitos.

**Teorema 8.** Para todo  $(v, \xi) \in V \times Q$  existe  $w \in V$  tal que

$$\mathcal{B}(c, \xi; w, -\xi) \gtrsim (\|v\|_1 + \|\xi\|_{-1})^2 \quad (4.13)$$

e

$$\|w\|_1 \lesssim \|v\|_1 + \|\xi\|_{-1} \quad (4.14)$$

**Demonstração:** Ver [10].

Para a formulação correta dos métodos iremos usar 2 métodos que adicionam estabilidade aos algoritmos que iremos derivar, estes são o *método misto* e o *método estabilizado* (cf. [9], [10] e [21]).

#### 4.2.2.1 Formulação mista

A formulação variacional mista para o problema é a seguinte: encontrar  $(u_h, \lambda_h) \in V_h \times \Lambda_h$  tal que:

$$\mathcal{B}(u_h, \lambda_h; v_h, \mu_h - \lambda_h) \leq \mathcal{L}(v_h, \mu_h - \lambda_h), \quad \forall (v_h, \mu_h) \in V_h \times \Lambda_h$$

Para este tipo de métodos, os espaços dos elementos finitos têm de satisfazer a condição *Babuska-Brezzi*:

$$\sup_{v_h \in V_h} \frac{\langle v_h, \xi_h \rangle}{\|v_h\|_V} \gtrsim \|\xi\|_{V'}, \quad \forall \xi_h \in Q_h. \quad (4.15)$$

Se a condição Babuska-Brezzi for válida, implica a seguinte estimativa de estabilidade discreta.

**Teorema 9.** Se a condição (4.15) é válida, então para todo par  $(v_h, \xi_h) \in V_h \times Q_h$ , existe  $w_h \in V_h$ , tal que

$$\mathcal{B}(v_h, \xi_h; w_h, -\xi_h) \gtrsim (\|v_h\|_V + \|\xi_h\|_{V'})^2. \quad (4.16)$$

e

$$\|w_h\|_V \lesssim \|v_h\|_V + \|\xi_h\|_{V'}. \quad (4.17)$$

**Demonstração:** Ver [10].

A estimativa de erro a priori é feita a partir da estimativa de estabilidade do Teorema 9.

**Teorema 10.** A seguinte estimativa de erro é válida

$$\|u - u_h\|_1 + \|\lambda - \lambda_h\|_1 \lesssim \inf_{v_h \in V_h} \|u - v_h\|_1 + \inf_{\mu_h \in \Lambda_h} \left( \|\lambda - \mu_h\|_{-1} + \sqrt{\langle u - g, \mu_h \rangle} \right). \quad (4.18)$$

**Demonstração:** Ver [10].

Usaremos a técnica de **funções bolha** (do inglês, *bubble functions*) para definir uma família de pares de elementos finitos estáveis. Com  $b_K \in P_3(K) \cap H_0^1(K)$  denotamos a função bolha dimensionada para ter um valor máximo de 1, e definimos

$$B_{l+1}(K) = \left\{ z \in H_0^1(K) : z = b_K w, w \in \tilde{P}_{l-2}(K) \right\}$$

onde  $\tilde{P}_{l-2}(K)$  denota o espaço de polinómios homogéneos de grau  $l-2$ . Seja  $k \geq 1$  o grau dos espaços de elementos finitos definidos por

$$V_h = \begin{cases} \{v_h \in V : v_h|_K \in P_1(K) \oplus B_3(K), \forall K \in \mathcal{C}_h\} \text{ para } k = 1 \\ \{v_h \in V : v_h|_K \in P_k(K) \oplus B_{k+1}(K), \forall K \in \mathcal{C}_h\} \text{ para } k \geq 2. \end{cases} \quad (4.19)$$

Então se  $k = 1$ , teremos  $v_h \in P_1(K) \oplus B_3(K)$ , onde  $B_3 = B_{2+1}$ , com  $l = 2$ , e portanto  $w \in \tilde{P}_0$  (sendo uma constante), e  $z = wb_K$  onde  $b_K \in P_3(K) \cap H_0^1(K)$ .

Matematicamente, o símbolo  $\oplus$  representa a soma direta de dois espaços vetoriais ou conjuntos. A soma direta de dois espaços vetoriais  $V$  e  $W$ , denotada por  $V \oplus W$ , é definida como o conjunto de todas as somas  $v + w$ , onde  $v$  pertence a  $V$  e  $w$  pertence a  $W$ , e cada elemento nessa soma é único.

No contexto da expressão  $P_1(K) \oplus B_3(K)$ , isso significa que estamos a considerar a soma direta dos espaços de funções lineares por partes de grau 1 em  $K$  (denotado por  $P_1(K)$ ) e as funções de bolha de grau 3 em  $K$  (denotadas por  $B_3(K)$ ). Essa expressão é usada na definição do espaço  $V_h$  para  $k = 1$ , indicando que as funções nesse espaço são uma combinação desses dois tipos de funções.

Quando se fala em *bubble functions*, tipicamente fala-se em adicionar um grau extra de liberdade ao baricentro de cada simplex da triangulação  $\mathcal{C}_h$  do domínio  $\Omega$ . Por exemplo para um triângulo de referência  $K$ , é definida usando as seguintes funções básicas relacionadas também ao triângulo  $K$ :

$$\Psi_1(x, y) = 1 - x - y ; \quad \Psi_2(x, y) = x ; \quad \Psi_3(x, y) = y$$

e define-se a função bolha, em duas dimensões, da seguinte forma:

$$\Psi_b(x, y) = 27\Psi_1(x, y)\Psi_2(x, y)\Psi_3(x, y) = 27 \prod_{i=1}^3 \Psi_i.$$

Geralmente é preferível trabalhar com uma base uniforme. O máximo de  $\Psi_1(x, y)\Psi_2(x, y)\Psi_3(x, y)$  é alcançado no baricentro de  $K$ , onde cada termo é  $1/3$ , e escala  $3^3 = 27$ . E generalizando numa dimensão  $d$ :

$$\Psi_b = (d+1)^{d+1} \prod_{i=1}^{d+1} \Psi_i.$$

Seja:

$$Q_h = \begin{cases} \{\xi_h \in Q : \xi_h|_K \in P_0(K), \forall K \in \mathcal{C}_h\} \text{ para } k=1 \\ \{\xi_h \in Q : \xi_h|_K \in P_{k-2}(K), \forall K \in \mathcal{C}_h\} \text{ para } k \geq 2. \end{cases} \quad (4.20)$$

Observe que as ordens de aproximação dos espaços de elementos finitos são balanceadas, ou seja,

$$\inf_{v_h \in V_h} \|u - v_h\|_1 = O(h^k) \text{ e } \inf_{\xi_h \in Q_h} \|\lambda - \xi_h\|_1 = O(h^k)$$

onde  $u \in H^{k+1}(\Omega)$  e  $\lambda \in H^{k-1}(\Omega)$ .

#### Formulação fraca discreta: Problema matricial algébrico

A área de contacto, ou seja, o subconjunto de  $\Omega$  onde a solução satisfaz  $u = g$ , é desconhecida e tem de ser determinada como parte da solução.

A formulação fraca discreta do nosso problema é descrita como: encontrar o par  $(u_h, \lambda_h) \in V_h \times \Lambda_h$  tal que:

$$\begin{cases} (\nabla u_h, \nabla v_h) - \langle v_h, \lambda_h \rangle = (f, v_h) & , \forall v_h \in V_h \\ \langle u_h - g, \mu_h - \lambda_h \rangle \geq 0 & , \forall \mu_h \in \Lambda_h \end{cases} \quad (4.21)$$

onde a segunda desigualdade do sistema, substituído por  $\mu_h = 0$  e  $\mu_h = 2\lambda_h$ , que correspondente ao sistema:

$$\begin{cases} (\nabla u_h, \nabla v_h) - \langle v_h, \lambda_h \rangle = (f, v_h) & , \forall v_h \in V_h \\ \langle u_h - g, \mu_h \rangle \geq 0 & , \forall \mu_h \in \Lambda_h \\ \langle u_h - g, \lambda_h \rangle = 0 & , \forall \mu_h \in \Lambda_h. \end{cases} \quad (4.22)$$

Consideramos o caso de elementos de ordem mais baixa, isto é com  $1 \leq k \leq 3$ .

Seja  $\xi_j, j \in \{1, \dots, M\}$  ( $M$  é nº total de bases de  $Q_h$  que será igual ao nº de elementos finitos da malha), as bases de Lagrange (nodal-nós) para  $Q_h$ . E temos

$$\Lambda_h = \left\{ \mu_h = \sum_{j=1}^M \mu_i \xi_j : \mu_j \geq 0, \forall \xi_j \text{ bases de } Q_h \right\}.$$

Sejam  $\Psi_j, j \in \{1, \dots, N\}$ , as bases do espaço  $V_h$  (com um total de  $N$  bases).

E portanto:

$$\mu_h = \sum_{j=1}^M \mu_j \xi_j$$

$$u_h = \sum_{j=1}^N u_j \Psi_j$$

que substituindo no sistema (4.22) obtemos o seguinte sistema matricial (complementary problem):

$$\begin{cases} \mathbf{A}\mathbf{u} - \mathbf{B}^T \boldsymbol{\lambda} = \mathbf{f} \\ \mathbf{B}\mathbf{u} \geq \mathbf{g} \\ \boldsymbol{\lambda}^T (\mathbf{B}\mathbf{u} - \mathbf{g}) = 0 \\ \boldsymbol{\lambda} \geq \mathbf{0} \end{cases} \quad (4.23)$$

onde

$$\mathbf{A} \in \mathbb{R}^{N \times N}, \quad (\mathbf{A})_{ij} = (\nabla \varphi_i, \nabla \varphi_j)$$

$$\mathbf{B} \in \mathbb{R}^{M \times N}, \quad (\mathbf{B})_{ij} = (\xi_i, \varphi_j)$$

$$\mathbf{f} \in \mathbb{R}^N, \quad (\mathbf{f})_i = (f, \varphi_i)$$

$$\mathbf{u} \in \mathbb{R}^N, \quad (\mathbf{u})_i = u_i$$

$$\mathbf{g} \in \mathbb{R}^M, \quad (\mathbf{g})_i = (g, \xi_i),$$

$$\boldsymbol{\lambda} \in \mathbb{R}^M, \quad (\boldsymbol{\lambda})_i = \lambda_i.$$

### Algoritmo iterativo: Primal-dual active set para o método misto

Pode-se ver a explicação deste algoritmo com detalhe em [22].

Para o nosso problema em concreto, é possível também tirar uma equivalência (entrada a entrada):

$$\begin{cases} \boldsymbol{\lambda}^T (\mathbf{B}\mathbf{u} - \mathbf{g}) = 0 \\ \mathbf{B}\mathbf{u} - \mathbf{g} \geq \mathbf{0} \\ \boldsymbol{\lambda} \geq \mathbf{0} \end{cases} \Leftrightarrow (\mathbf{0} = \mathbf{B}\mathbf{u} - \mathbf{g} \wedge \mathbf{0} \leq \boldsymbol{\lambda}) \vee (\mathbf{0} = \boldsymbol{\lambda} \wedge \mathbf{0} \leq \mathbf{B}\mathbf{u} - \mathbf{g}) \Leftrightarrow \min\{\boldsymbol{\lambda}; \mathbf{B}\mathbf{u} - \mathbf{g}\} = \mathbf{0}$$

o que é equivalente a

$$\begin{aligned} \min\{\boldsymbol{\lambda}; \mathbf{B}\mathbf{u} - \mathbf{g}\} = \mathbf{0} &\Leftrightarrow \max\{-\boldsymbol{\lambda}; -(\mathbf{B}\mathbf{u} - \mathbf{g})\} = \mathbf{0} \Leftrightarrow \max\{-\boldsymbol{\lambda}; \mathbf{g} - \mathbf{B}\mathbf{u}\} = \mathbf{0} \Leftrightarrow \\ &\Leftrightarrow (\mathbf{0} = \mathbf{g} - \mathbf{B}\mathbf{u} \wedge \mathbf{0} \geq -\boldsymbol{\lambda}) \vee (\mathbf{0} = -\boldsymbol{\lambda} \wedge \mathbf{0} \geq \mathbf{g} - \mathbf{B}\mathbf{u}) \end{aligned}$$

multiplicando agora  $\mathbf{g} - \mathbf{B}\mathbf{u}$  por uma constante  $c > 0$ , manteremos ainda assim as desigualdades

$$(\mathbf{0} = c(\mathbf{g} - \mathbf{B}\mathbf{u}) \wedge \mathbf{0} \geq -\boldsymbol{\lambda}) \vee (\mathbf{0} = -\boldsymbol{\lambda} \wedge \mathbf{0} \geq c(\mathbf{g} - \mathbf{B}\mathbf{u})) \Leftrightarrow \max\{-\boldsymbol{\lambda}; c(\mathbf{g} - \mathbf{B}\mathbf{u})\} = 0$$

e somando  $\boldsymbol{\lambda}$  em todas as entradas do sistema de maximização, conseguimos manter a equivalência e

obtemos

$$(\boldsymbol{\lambda} = \boldsymbol{\lambda} + c(\mathbf{g} - \mathbf{B}\mathbf{u}) \wedge \boldsymbol{\lambda} \geq \mathbf{0}) \vee (\boldsymbol{\lambda} = \mathbf{0} \wedge \mathbf{0} \geq c(\mathbf{g} - \mathbf{B}\mathbf{u})) \Leftrightarrow \max\{\mathbf{0}; \boldsymbol{\lambda} + c(\mathbf{g} - \mathbf{B}\mathbf{u})\} = \boldsymbol{\lambda}$$

Pode ser confuso, mas agora poderemos usar esta fórmula de uma forma iterativa, e assim atualizar  $\boldsymbol{\lambda}$  a cada iteração, ou seja:

$$\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda} + c(\mathbf{g} - \mathbf{B}\mathbf{u}) \quad (4.24)$$

E considere-se  $\mathbf{a}$  e  $\mathbf{i}$  como sendo dois vetores de índices de  $\tilde{\boldsymbol{\lambda}}$ , tal que

$$\tilde{\boldsymbol{\lambda}}(\mathbf{a}) > \mathbf{0} \rightarrow \text{índices dos elementos positivos de } \tilde{\boldsymbol{\lambda}}$$

$$\tilde{\boldsymbol{\lambda}}(\mathbf{i}) = \mathbf{0} \rightarrow \text{índices dos elementos não positivos de } \tilde{\boldsymbol{\lambda}}$$

e que  $\text{length}\{\mathbf{a}\} + \text{length}\{\mathbf{i}\} = M$ , formando os conjuntos (*Active Set* e *Inactive Set*, respectivamente):

$$\mathcal{A} = \{a : \lambda_a + c(g_a - \mathbf{B}[a,:]\mathbf{u}) > 0\}$$

$$\mathcal{I} = \{i : \lambda_i + c(g_i - \mathbf{B}[i,:]\mathbf{u}) \leq 0\}.$$

Para os elementos positivos de  $\boldsymbol{\lambda}$ , ou seja para  $\boldsymbol{\lambda}(\mathbf{a})$ , é verdade que

$$\begin{cases} \mathbf{A}\mathbf{u} - \mathbf{B}^T\boldsymbol{\lambda} = \mathbf{f} \\ \mathbf{B}\mathbf{u} - \mathbf{g} \geq \mathbf{0} \end{cases} \Rightarrow \begin{cases} \mathbf{A}\mathbf{u} - \mathbf{B}[\mathbf{a},:]^T\boldsymbol{\lambda}(\mathbf{a}) = \mathbf{f} \\ \mathbf{B}[\mathbf{a},:]^T\mathbf{u} - \mathbf{g}(\mathbf{a}) = \mathbf{0} \end{cases} \Leftrightarrow \begin{bmatrix} \mathbf{A} & -\mathbf{B}[\mathbf{a},:]^T \\ \mathbf{B}[\mathbf{a},:] & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\lambda}(\mathbf{a}) \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g}(\mathbf{a}) \end{bmatrix} \quad (4.25)$$

- o que nos permite calcular iterativamente a solução:

(i) Inicializar:  $k = 0$ ,  $\boldsymbol{\lambda}^{(0)} = \mathbf{0}$ , e resolva  $\mathbf{A}\mathbf{u}^{(0)} = \mathbf{f}$ ;

(ii) Para  $k \geq 0$ , calcular:  $\tilde{\boldsymbol{\lambda}}^{(k)} = \boldsymbol{\lambda}^{(k)} + c(\mathbf{g} - \mathbf{B}\mathbf{u}^{(k)})$ , e initialize-se o Active-Set e o Inactive-Set:

$$\mathcal{A}_k = \left\{ a : \tilde{\lambda}_a^{(k)} > 0 \right\} ; \quad \mathcal{I}_k = \left\{ i : \tilde{\lambda}_i^{(k)} \leq 0 \right\}$$

tal que  $\tilde{\boldsymbol{\lambda}}^{(k)}(\mathbf{a}^k) > 0$ , e  $\tilde{\boldsymbol{\lambda}}^{(k)}(\mathbf{i}^k) \leq 0$ .

(iii) Depois resolver

$$\begin{cases} \mathbf{A}\mathbf{u}^{(k+1)} - \mathbf{B}^T\boldsymbol{\lambda}^{(k+1)} = \mathbf{f} \\ \mathbf{B}\mathbf{u}^{(k+1)} - \mathbf{g} \geq \mathbf{0} \end{cases}, \text{ no conjunto } \mathcal{A}_k \quad (4.26)$$

e resolver

$$\boldsymbol{\lambda}^{(k+1)} \geq \mathbf{0}, \text{ no conjunto } \mathcal{I}_k \quad (4.27)$$

Isto é, resolver:

$$\begin{bmatrix} \mathbf{A} & -\mathbf{B}[\mathbf{a}^k, :]^T \\ \mathbf{B}[\mathbf{a}^k, :] & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(k+1)} \\ \boldsymbol{\lambda}^{(k+1)}(\mathbf{a}^k) \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g}(\mathbf{a}^k) \end{bmatrix}$$

$$\text{e } \boldsymbol{\lambda}^{(k+1)}(\mathbf{i}^k) = \mathbf{0}.$$

(iv) verificar condição de paragem: se  $\|\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^{(k)}\| \leq \text{TOL}$  terminar;

(v) Se não verificar condição de paragem, incrementar  $k = k + 1$  e retornar ao ponto (ii).

Para além desta explicação, o algoritmo iterativo primal-dual active set para o método misto, está também escrito como pseudocódigo em A.2.

#### 4.2.2.2 Formulação com estabilização

A partir do problema de Stokes, sabe-se que a técnica de utilização de funções bolha (adicionando mais graus de liberdade) para estabilizar o método pode ser evitada, através da adição no método de **termos de estabilização** (pelos chamados *residual-based stabilizing*).

Como vimos sendo  $U = V \times Q$  e definimos a forma bilinear  $\mathcal{B} : U \times U \rightarrow \mathbb{R}$ , e a forma linear,  $\mathcal{L} : U \rightarrow \mathbb{R}$ , como:

$$\mathcal{B}(w, \xi; v, \mu) = (\nabla w, \nabla v) - \langle v, \xi \rangle - \langle w, \mu \rangle \quad (4.28)$$

$$\mathcal{L}(v, \mu) = (f, v) - \langle g, \mu \rangle \quad (4.29)$$

Vamos agora introduzir a forma bilinear e linear  $\mathcal{S}_h$  e  $\mathcal{F}_h$ , por:

$$\mathcal{S}_h(w, \xi; v, \mu) = \sum_{K \in \mathcal{C}_h} h_K^2 (-\Delta w - \xi, -\Delta v - \mu)_K \quad (4.30)$$

$$\mathcal{F}_h(v, \mu) = \sum_{K \in \mathcal{C}_h} h_K^2 (f, -\Delta v - \mu)_K \quad (4.31)$$

e então definimos as formas  $\mathcal{B}_h$  e  $\mathcal{L}_h$  por:

$$\mathcal{B}_h(w, \xi; v, \mu) = \mathcal{B}(w, \xi; v, \mu) - \alpha \mathcal{S}_h(w, \xi; v, \mu) \quad (4.32)$$

$$\mathcal{L}_h(v, \mu) = \mathcal{L}(v, \mu) - \alpha \mathcal{F}_h(v, \mu) \quad (4.33)$$

onde  $\alpha > 0$  é um parâmetro de estabilização.

Notando que a assunção que  $f \in L^2(\Omega)$  permite que seja válido  $\Delta u + \lambda \in L^2(\Omega)$ , mesmo que  $\Delta u \notin L^2(\Omega)$  e  $\lambda \notin L^2(\Omega)$ , pois  $\Delta u + \lambda = f \in L^2(\Omega)$ . Portanto, sustenta que

$$\mathcal{S}_h(u, \lambda; v_h, \mu_h) = \mathcal{F}_h(v_h, \mu_h), \forall (v_h, \mu_h) \in V_h \times \Lambda_h \quad (4.34)$$

motivando o método de elementos finitos estabilizados (*stabilized finite element method*). Neste método, o problema passa a ser: encontrar o par  $(u_h, \lambda_h) \in V_h \times \Lambda_h$  tal que

$$\mathcal{B}_h(u_h, \lambda_h; v_h, \mu_h - \lambda_h) \leq \mathcal{L}_h(v_h, \mu_h - \lambda_h), \forall (v_h, \mu_h) \in V_h \times \Lambda_h. \quad (4.35)$$

Na nossa análise, precisamos de uma desigualdade inversa, que escrevemos da seguinte forma: existe uma constante positiva  $C_I$  tal que

$$C_I \sum_{K \in C_K} h_K^2 \|\Delta v_h\|_{0,K}^2 \leq \|\nabla v_h\|_0^2, \forall v_h \in V_h. \quad (4.36)$$

A seguinte condição de estabilidade é válida.

**Teorema 11.** Suponhamos que  $0 \leq \alpha \leq C_I$ . Então para todo  $(v_h, \xi_h) \in V_h \times Q_h$ , existe  $w_h \in V_h$  tal que

$$\mathcal{B}_h(v_h, \xi_h; w_h, -\xi_h) \gtrsim (\|v_h\|_1 + \|\xi_h\|_{-1})^2 \quad (4.37)$$

e

$$\|w_h\|_1 \lesssim \|v_h\|_1 + \|\xi_h\|_{-1}. \quad (4.38)$$

**Demonstração:** Ver [10].

Apresentamos em seguida a seguinte estimativa de erro a priori.

**Teorema 12.** A seguinte estimativa de erro é válida

$$\|u - u_h\|_1 + \|\lambda - \lambda_h\|_1 \lesssim \inf_{v_h \in V_h} \|u - v_h\|_1 + \inf_{\mu_h \in \Lambda_h} \left( \|\lambda - \mu_h\|_{-1} + \sqrt{\langle u - g, \mu_h \rangle} \right) + osc(f). \quad (4.39)$$

**Demonstração:** Ver [10].

Para o método estabilizado, definimos os espaços de elementos finitos como:

$$V_h = \{v_h \in V : v_h|_K \in P_k(K), \forall K \in \mathcal{C}_h\} \quad (4.40)$$

e

$$Q_h = \begin{cases} \{\xi_h \in Q : \xi_h|_K \in P_0(K), \forall K \in \mathcal{C}_h\} \text{ para } k = 1 \\ \{\xi_h \in Q : \xi_h|_K \in P_{k-2}(K), \forall K \in \mathcal{C}_h\} \text{ para } k \geq 2 \end{cases} \quad (4.41)$$

### Formulação fraca discreta: Problema matricial algébrico

A formulação fraca discreta do nosso problema é descrita como: encontrar o par  $(u_h, \lambda_h) \in V_h \times \Lambda_h$

tal que:

$$\begin{cases} (\nabla u_h, \nabla v_h) - \langle \lambda_h, v_h \rangle - \alpha \sum_{K \in C_h} h_K^2 (\Delta u_h + \lambda_h, \Delta v_h)_K = (f, v_h) + \alpha \sum_{K \in C_h} h_K^2 (f, -\Delta v_h)_K \\ \langle u_h - g, \mu_h - \lambda_h \rangle + \alpha \sum_{K \in C_h} h_K^2 (\Delta u_h + \lambda_h + f, \mu_h - \lambda_h)_K \geq 0 \end{cases} \quad (4.42)$$

válido para  $\forall (v_h, \mu_h) \in V_h \times \Lambda_h$ . Através de etapas semelhantes às do caso misto, chegamos ao sistema algébrico:

$$\begin{cases} \mathbf{A}_\alpha \mathbf{u} - \mathbf{B}_\alpha^T \boldsymbol{\lambda} = \mathbf{f}_\alpha \\ \mathbf{B}_\alpha \mathbf{u} + \mathbf{C}_\alpha \boldsymbol{\lambda} \geq \mathbf{g}_\alpha \\ \boldsymbol{\lambda}^T (\mathbf{B}_\alpha \mathbf{u} + \mathbf{C}_\alpha \boldsymbol{\lambda} - \mathbf{g}_\alpha) = 0 \\ \boldsymbol{\lambda} \geq \mathbf{0} \end{cases} \quad (4.43)$$

onde

$$\begin{aligned} \mathbf{A}_\alpha &\in \mathbb{R}^{N \times N}, \quad (\mathbf{A}_\alpha)_{ij} = (\nabla \varphi_i, \nabla \varphi_j) - \alpha \sum_{K \in C_h} h_K^2 (\Delta \varphi_i, \Delta \varphi_j)_K \\ \mathbf{B}_\alpha &\in \mathbb{R}^{M \times N}, \quad (\mathbf{B}_\alpha)_{ij} = (\xi_i, \varphi_j) + \alpha \sum_{K \in C_h} h_K^2 (\xi_i, \Delta \varphi_j)_K \\ \mathbf{C}_\alpha &\in \mathbb{R}^{M \times M}, \quad (\mathbf{C}_\alpha)_{ij} = \sum_{K \in C_h} h_K^2 (\xi_i, \xi_j)_K \\ \mathbf{f}_\alpha &\in \mathbb{R}^N, \quad (\mathbf{f}_\alpha)_i = (f, \varphi_i) + \alpha \sum_{K \in C_h} h_K^2 (f, \Delta \varphi_i)_K \\ \mathbf{g}_\alpha &\in \mathbb{R}^M, \quad (\mathbf{g}_\alpha)_i = (g, \xi_i) - \alpha \sum_{K \in C_h} h_K^2 (f, \xi_i)_K \\ \mathbf{u} &\in \mathbb{R}^N, \quad (\mathbf{u})_i = u_i \\ \boldsymbol{\lambda} &\in \mathbb{R}^M, \quad (\boldsymbol{\lambda})_i = \lambda_i. \end{aligned}$$

Para o nosso problema em concreto, é possível também tirar uma equivalência (entrada a entrada):

$$\begin{cases} \boldsymbol{\lambda}^T (\mathbf{B}_\alpha \mathbf{u} + \mathbf{C}_\alpha \boldsymbol{\lambda} - \mathbf{g}_\alpha) = 0 \\ \mathbf{B}_\alpha \mathbf{u} + \mathbf{C}_\alpha \boldsymbol{\lambda} \geq \mathbf{g}_\alpha \\ \boldsymbol{\lambda} \geq \mathbf{0} \end{cases} \Leftrightarrow$$

$$\Leftrightarrow (\mathbf{0} = \mathbf{B}_\alpha \mathbf{u} + \mathbf{C}_\alpha \boldsymbol{\lambda} - \mathbf{g}_\alpha \leq \boldsymbol{\lambda}) \vee (\mathbf{0} = \boldsymbol{\lambda} \leq \mathbf{B}_\alpha \mathbf{u} + \mathbf{C}_\alpha \boldsymbol{\lambda} - \mathbf{g}_\alpha \Leftrightarrow \min\{\boldsymbol{\lambda}; \mathbf{B}_\alpha \mathbf{u} + \mathbf{C}_\alpha \boldsymbol{\lambda} - \mathbf{g}_\alpha\} = \mathbf{0}).$$

Analogamente com o que vimos no método misto podemos escrever:

$$\begin{cases} \mathbf{A}_\alpha \mathbf{u} - \mathbf{B}_\alpha^T \boldsymbol{\lambda} = \mathbf{f}_\alpha \\ \boldsymbol{\lambda} - \max\{\mathbf{0}; \boldsymbol{\lambda} + c(\mathbf{g}_\alpha - \mathbf{B}_\alpha \mathbf{u} - \mathbf{C}_\alpha \boldsymbol{\lambda})\} = 0 \end{cases} \quad (4.44)$$

mas por outro lado

$$\min\{\boldsymbol{\lambda}; \mathbf{B}_\alpha \mathbf{u} + \mathbf{C}_\alpha \boldsymbol{\lambda} - \mathbf{g}_\alpha\} = \mathbf{0} \Leftrightarrow (\mathbf{0} = \mathbf{B}_\alpha \mathbf{u} + \mathbf{C}_\alpha \boldsymbol{\lambda} - \mathbf{g}_\alpha \wedge \mathbf{0} \leq \boldsymbol{\lambda}) \vee (\mathbf{0} = \boldsymbol{\lambda} \wedge \mathbf{0} \leq \mathbf{B}_\alpha \mathbf{u} + \mathbf{C}_\alpha \boldsymbol{\lambda} - \mathbf{g}_\alpha) \Leftrightarrow$$

$$\begin{aligned}
&\Leftrightarrow (-\mathbf{C}_\alpha \boldsymbol{\lambda} = \mathbf{B}_\alpha \mathbf{u} - \mathbf{g}_\alpha \wedge \mathbf{0} \leq \boldsymbol{\lambda}) \vee (\mathbf{0} = \boldsymbol{\lambda} \wedge -\mathbf{C}_\alpha \boldsymbol{\lambda} \leq \mathbf{B}_\alpha \mathbf{u} - \mathbf{g}_\alpha) \Leftrightarrow \\
&\Leftrightarrow (\boldsymbol{\lambda} = \mathbf{C}_\alpha^{-1}(\mathbf{g}_\alpha - \mathbf{B}_\alpha \mathbf{u}) \wedge \boldsymbol{\lambda} \geq \mathbf{0}) \vee (\boldsymbol{\lambda} = \mathbf{0} \wedge \boldsymbol{\lambda} \geq \mathbf{C}_\alpha^{-1}(\mathbf{g}_\alpha - \mathbf{B}_\alpha \mathbf{u})) \Leftrightarrow \\
&\Leftrightarrow \boldsymbol{\lambda} = \max\{\mathbf{C}_\alpha^{-1}(\mathbf{g}_\alpha - \mathbf{B}_\alpha \mathbf{u}); \mathbf{0}\}.
\end{aligned}$$

Para os elementos positivos de  $\boldsymbol{\lambda}$  é verdade que

$$\begin{aligned}
&\left\{ \begin{array}{l} \mathbf{A}_\alpha \mathbf{u} - \mathbf{B}_\alpha^T \boldsymbol{\lambda} = \mathbf{f}_\alpha \\ \mathbf{B}_\alpha \mathbf{u} + \mathbf{C}_\alpha \boldsymbol{\lambda} - \mathbf{g}_\alpha \geq \mathbf{0} \end{array} \right. \Rightarrow \left\{ \begin{array}{l} \mathbf{A}_\alpha \mathbf{u} - \mathbf{B}_\alpha[\mathbf{a}, :]^T \boldsymbol{\lambda}(\mathbf{a}) = \mathbf{f}_\alpha \\ \mathbf{B}_\alpha[\mathbf{a}, :] \mathbf{u} + \mathbf{C}_\alpha[\mathbf{a}, \mathbf{a}] \boldsymbol{\lambda}(\mathbf{a}) - \mathbf{g}_\alpha(\mathbf{a}) = \mathbf{0} \end{array} \right. \Leftrightarrow \\
&\Leftrightarrow \begin{bmatrix} \mathbf{A}_\alpha & -\mathbf{B}_\alpha[\mathbf{a}, :]^T \\ \mathbf{B}_\alpha[\mathbf{a}, :] & \mathbf{C}_\alpha[\mathbf{a}, \mathbf{a}] \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\lambda}(\mathbf{a}) \end{bmatrix} = \begin{bmatrix} \mathbf{f}_\alpha \\ \mathbf{g}_\alpha(\mathbf{a}) \end{bmatrix} \Leftrightarrow \\
&\Leftrightarrow \left\{ \begin{array}{l} \mathbf{A}_\alpha \mathbf{u} - \mathbf{B}_\alpha[\mathbf{a}, :]^T \boldsymbol{\lambda}(\mathbf{a}) = \mathbf{f}_\alpha \\ \boldsymbol{\lambda}(\mathbf{a}) = \mathbf{C}_\alpha^{-1}[\mathbf{a}, \mathbf{a}] (\mathbf{g}_\alpha(\mathbf{a}) - \mathbf{B}_\alpha[\mathbf{a}, :] \mathbf{u}) \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \mathbf{A}_\alpha \mathbf{u} - \mathbf{B}_\alpha[\mathbf{a}, :]^T \mathbf{C}_\alpha^{-1}[\mathbf{a}, \mathbf{a}] (\mathbf{g}_\alpha(\mathbf{a}) - \mathbf{B}_\alpha[\mathbf{a}, :] \mathbf{u}) = \mathbf{f}_\alpha \\ \boldsymbol{\lambda}(\mathbf{a}) = \mathbf{C}_\alpha^{-1}[\mathbf{a}, \mathbf{a}] (\mathbf{g}_\alpha(\mathbf{a}) - \mathbf{B}_\alpha[\mathbf{a}, :] \mathbf{u}) \end{array} \right. \Leftrightarrow \\
&\Leftrightarrow \left\{ \begin{array}{l} (\mathbf{A}_\alpha + \mathbf{B}_\alpha[\mathbf{a}, :]^T \mathbf{C}_\alpha^{-1}[\mathbf{a}, \mathbf{a}] \mathbf{B}_\alpha[\mathbf{a}, :]) \mathbf{u} = \mathbf{f}_\alpha + \mathbf{B}_\alpha[\mathbf{a}, :]^T \mathbf{C}_\alpha^{-1}[\mathbf{a}, \mathbf{a}] \mathbf{g}_\alpha \\ \boldsymbol{\lambda}(\mathbf{a}) = \mathbf{C}_\alpha^{-1}[\mathbf{a}, \mathbf{a}] (\mathbf{g}_\alpha(\mathbf{a}) - \mathbf{B}_\alpha[\mathbf{a}, :] \mathbf{u}) \end{array} \right.
\end{aligned}$$

### Algoritmo iterativo: Primal-dual active set para o método estabilizado

- (i) Inicializar:  $k = 0$ , resolver  $\mathbf{A}\mathbf{u}^{(0)} = \mathbf{f}$  e em seguida calcular  $\boldsymbol{\lambda}^{(0)} = \max\{\mathbf{C}_\alpha^{-1}(\mathbf{g}_\alpha - \mathbf{B}_\alpha \mathbf{u}^{(0)}); \mathbf{0}\}$ ;
- (ii) Para  $k \geq 0$ , inicializar o *Active-Set* e o *Inactive-Set*:

$$\mathcal{A}_k = \left\{ a : \lambda_a^{(k)} > 0 \right\} ; \quad \mathcal{I}_k = \left\{ i : \lambda_i^{(k)} \leq 0 \right\}$$

tal que  $\boldsymbol{\lambda}^{(k)}(\mathbf{a}^k) > 0$ , e  $\boldsymbol{\lambda}^{(k)}(\mathbf{i}^k) \leq 0$ .

- (iii) Para obter  $\mathbf{u}^{(k+1)}$  resolva:

$$\begin{cases} \mathbf{A}_\alpha \mathbf{u}^{(k+1)} - \mathbf{B}_\alpha^T \boldsymbol{\lambda} = \mathbf{f}_\alpha \\ \mathbf{B}_\alpha \mathbf{u}^{(k+1)} + \mathbf{C}_\alpha \tilde{\boldsymbol{\lambda}} - \mathbf{g}_\alpha \geq \mathbf{0} \end{cases}, \text{ no conjunto } \mathcal{A}_k \tag{4.45}$$

Isto é, resolver:

$$\begin{bmatrix} \mathbf{A}_\alpha & -\mathbf{B}_\alpha[\mathbf{a}^k, :]^T \\ \mathbf{B}_\alpha[\mathbf{a}^k, :] & \mathbf{C}_\alpha[\mathbf{a}^k, \mathbf{a}^k] \end{bmatrix} \begin{bmatrix} \mathbf{u}^{k+1} \\ \tilde{\boldsymbol{\lambda}}(\mathbf{a}^k) \end{bmatrix} = \begin{bmatrix} \mathbf{f}_\alpha \\ \mathbf{g}_\alpha(\mathbf{a}^k) \end{bmatrix}$$

podendo ainda ser escrito como

$$(\mathbf{A}_\alpha + \mathbf{B}_\alpha[\mathbf{a}^k, :]^T \mathbf{C}_\alpha^{-1}[\mathbf{a}^k, \mathbf{a}^k] \mathbf{B}_\alpha[\mathbf{a}^k, :]) \mathbf{u}^{(k+1)} = \mathbf{f}_\alpha + \mathbf{B}_\alpha[\mathbf{a}^k, :]^T \mathbf{C}_\alpha^{-1}[\mathbf{a}^k, \mathbf{a}^k] \mathbf{g}_\alpha$$

e para obter  $\boldsymbol{\lambda}^{(k+1)}$ :

$$\boldsymbol{\lambda}^{(k+1)} = \max\{\mathbf{C}_\alpha^{-1}(\mathbf{g}_\alpha - \mathbf{B}_\alpha \mathbf{u}^{(k+1)}); \mathbf{0}\}$$

(iv) Verificar condição de paragem: se  $\|\lambda^{(k+1)} - \lambda^{(k)}\| \leq \text{TOL}$  terminar;

(v) Se não verificar condição de paragem, incrementar  $k = k + 1$  e retornar ao ponto (ii).

Para além desta explicação, o algoritmo iterativo primal-dual active set para o método estabilizado, está também escrito como pseudocódigo nos anexos em A.3.

## 4.3 Problema Parabólico: Problema de valor final - Black Scholes para American Options

### 4.3.1 Aproximação Finita do Domínio

O domínio onde aproximamos a solução é

$$[0, T] \times [0, \infty[. \quad (4.46)$$

Para que possamos calcular a nossa aproximação numérica de um modo viável, introduzimos um limite artificial  $S_{\max}$  ( $S_{\max} \approx \infty$ ), que nos permite ter um domínio limitado, dado por:

$$[0, T] \times [0, S_{\max}]. \quad (4.47)$$

A escolha do valor adequado para  $S_{\max}$  é uma consideração importante. Um  $S_{\max}$  muito alto pode aumentar o número de pontos de discretização necessários para os cálculos, tornando o processo computacionalmente caro e demorado. Por outro lado, um  $S_{\max}$  muito baixo pode resultar em perda de precisão nos cálculos, pois pode excluir áreas importantes do espaço de preços que podem afetar o resultado final.

Assim o problema para Opções Europeias é dado pelo sistema:

$$\begin{cases} \frac{\partial V}{\partial t} + \mathcal{L}_{BS}V = 0, & \text{em } (0, S_{\max}) \times (0, T) \\ V = g, & \text{sobre } (0, S_{\max}) \times \{t = T\} \\ V = h, & \text{sobre } \partial\{(0, S_{\max})\} \times (0, T) \end{cases} \quad (4.48)$$

e para Opções Americanas:

$$\begin{cases} \frac{\partial V}{\partial t} + \mathcal{L}_{BS}V \leq 0, & \text{em } (0, S_{\max}) \times (0, T) \\ (g - V) \leq 0, & \text{em } (0, S_{\max}) \times (0, T) \\ (\frac{\partial V}{\partial t} + \mathcal{L}_{BS}V)(g - V) = 0, & \text{em } (0, S_{\max}) \times (0, T) \\ V = g, & \text{sobre } (0, S_{\max}) \times \{t = T\} \\ V = h, & \text{sobre } \partial\{(0, S_{\max})\} \times (0, T). \end{cases} \quad (4.49)$$

Recorda-se que  $\mathcal{L}_{BS}$  é o operador diferencial linear de segunda ordem:

$$\mathcal{L}_{BS}(V) := \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS \frac{\partial V}{\partial S} - rV \quad (4.50)$$

$g(S, t)$  é o retorno, dado por:

$$\begin{aligned} g_c(S, t) &= \max \{S - K; 0\}, \quad (\text{Opções de compra}) \\ g_p(S, t) &= \max \{K - S; 0\}, \quad (\text{Opções de venda}) \end{aligned} \quad (4.51)$$

e as condições de contorno, para Opções Europeias são dadas por:

$$\begin{aligned} h_c^{\text{Eur}}(S, t) &= \begin{cases} 0, & \forall (S, t) \in \{0\} \times (0, T) \\ S - Ke^{-r(T-t)}, & \forall (S, t) \in \{S_{\max}\} \times (0, T) \end{cases}, \quad (\text{Opções de compra}) \\ h_p^{\text{Eur}}(S, t) &= \begin{cases} Ke^{-r(T-t)}, & \forall (S, t) \in \{0\} \times (0, T) \\ 0, & \forall (S, t) \in \{S_{\max}\} \times (0, T) \end{cases}, \quad (\text{Opções de venda}) \end{aligned} \quad (4.52)$$

sendo que  $\partial\{(0, S_{\max})\} = \{0; S_{\max}\}$ . Finalmente, as condições de contorno, para Opções Americanas são:

$$h^{\text{Ame}}(S, t) = \max\{h^{\text{Eur}}, g\}.$$

### 4.3.2 Método das Diferenças Finitas para Black Scholes

- O ativo/ação subjacente terá preços no intervalo  $[0, \infty[$ , mas após introduzir um limite artificial  $S_{\max}$ , passamos a considerar o intervalo  $[0, S_{\max}]$ , que será dividido em  $M_S + 1$  sub-intervalos de igual comprimento  $h_S$ :

$$h_S = \frac{S_{\max} - S_{\min}}{M_S + 1} \Leftrightarrow S_{\max} = S_{\min} + h_S \cdot (M_S + 1)$$

no total, com  $M_S + 2$  pontos, da variável  $S$ . Portanto:

$$S_m = S_{\min} + m \cdot h_S, \quad m = \{0, 1, \dots, M_S, M_S + 1\}$$

- O intervalo  $[0, T]$  também é dividido em  $N_t + 1$  sub-intervalos de igual comprimento  $h_t$ :

$$h_t = \frac{t_{\max} - t_{\min}}{N_t + 1} \Leftrightarrow T = t_{\min} + h_t \cdot N_t$$

no total, com  $N_t + 2$  pontos, da variável tempo, pelo que:

$$t_n = n \cdot \Delta t, \quad n = \{0, 1, \dots, N_t, N_t + 1\}$$

- Assim, o espaço  $[0, T] \times [0, S_{\max}]$  é aproximado por uma grelha retangular, onde a função:

$$V(S, t) = V(m \cdot \Delta S, n \cdot \Delta t) \approx V_m^n$$

é o valor da opção no tempo  $t$  e no valor do ativo subjacente  $S$ . Para simplificar a equação de Black-Scholes, considere-se os seguintes parâmetros:

$$\begin{cases} a(S, t) = \frac{1}{2}\sigma^2 S^2 \\ b(S, t) = rS \\ c(S, t) = -r \end{cases} \quad (4.53)$$

que nos permite reescrever o operador parabólico da seguinte maneira:

$$\frac{\partial V}{\partial t} + \mathcal{L}_{BS}V = \frac{\partial V}{\partial t} + a(S, t)\frac{\partial^2 V}{\partial S^2} + b(S, t)\frac{\partial V}{\partial S} + c(S, t)V$$

Comecemos por discretizar as derivadas relativas à variável  $S$ , em que para a derivada parcial de primeira e segunda ordem usamos a aproximação de diferenças centradas:

$$\frac{\partial V}{\partial S} = \frac{V_{m+1} - V_{m-1}}{2h_S} + O(h_S^2)$$

$$\frac{\partial^2 V}{\partial S^2} = \frac{V_{m+1} - 2V_m + V_{m-1}}{h_S^2} + O(h_S^2).$$

Substituindo no modelo de Black-Scholes

$$\frac{\partial}{\partial t} V_m(t) + a_m(t) \frac{V_{m+1}(t) - 2V_m(t) + V_{m-1}(t)}{h_S^2} + b_m(t) \frac{V_{m+1}(t) - V_{m-1}(t)}{2h_S} + c_m(t)V_m(t)$$

colocando  $V$  em evidencia:

$$\frac{\partial}{\partial t} V_m(t) + \left[ \frac{a_m(t)}{h_S^2} - \frac{b_m(t)}{2h_S} \right] V_{m-1}(t) + \left[ c_m(t) - \frac{2a_m(t)}{h_S^2} \right] V_m(t) + \left[ \frac{a_m(t)}{h_S^2} + \frac{b_m(t)}{2h_S} \right] V_{m+1}(t).$$

Definem-se os termos

$$\begin{cases} A_m(t) = \frac{a_m(t)}{h_S^2} - \frac{b_m(t)}{2h_S} = \frac{\sigma^2 S_m^2}{2h_S^2} - \frac{rS_m}{2h_S} = \frac{S_m}{2h_S} \left( \frac{\sigma^2 S_m}{h_S} - r \right) \\ B_m(t) = c_m(t) - \frac{2a_m(t)}{h_S^2} = -r - \frac{\sigma^2 S_m^2}{h_S^2} \\ C_m(t) = \frac{a_m(t)}{h_S^2} + \frac{b_m(t)}{2h_S} = \frac{\sigma^2 S_m^2}{2h_S^2} + \frac{rS_m}{2h_S} = \frac{S_m}{2h_S} \left( \frac{\sigma^2 S_m}{h_S} + r \right). \end{cases}$$

Juntando todas as operações diferenciais relativas ao interior do domínio  $[0, S_{\max}]$ , ou seja para  $m = 1, \dots, M_S$ , num total de  $M_S$  linhas, estes operadores discretizados no espaço, podem ser vistos, de

uma forma matricial como:

$$\frac{\partial}{\partial t} \mathbf{V}_r(t) + \mathbf{L}(t) \mathbf{V}(t) \quad (4.54)$$

O vetor  $\mathbf{V}(t)$ , é um vetor de dimensão  $\dim\{\mathbf{V}\} = M_s + 2$ :

$$\mathbf{V}(t) = \begin{bmatrix} V_0(t) \\ V_1(t) \\ V_2(t) \\ \vdots \\ V_{M_s}(t) \\ V_{M_s+1}(t) \end{bmatrix} \in \mathbb{R}^{M_s+2} \quad (4.55)$$

e  $\mathbf{V}_r(t)$ , é um vetor de dimensão  $\dim\{\mathbf{V}\} = M_s$ , que é composto tomando apenas os pontos interiores do domínio:

$$\mathbf{V}_r(t) = \begin{bmatrix} V_1(t) \\ V_2(t) \\ \vdots \\ V_{M_s-1}(t) \\ V_{M_s}(t) \end{bmatrix} \quad (4.56)$$

A matriz  $\mathbf{L} \in \mathbb{R}^{(M_s) \times (M_s+2)}$ , é uma matriz tridiagonal não quadrada, com a seguinte configuração:

$$\mathbf{L}(t) = \begin{bmatrix} A_1(t) & B_1(t) & C_1(t) & & & & \\ A_2(t) & B_2(t) & C_2(t) & & & & \\ & A_3(t) & B_3(t) & C_3(t) & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & A_{M_s}(t) & B_{M_s}(t) & C_{M_s}(t) & \end{bmatrix} \in \mathbb{R}^{(M_s) \times (M_s+2)}. \quad (4.57)$$

Note-se que, no nosso caso,  $\mathbf{L}(t)$  não depende de  $t$ , portanto podemos escrever  $\mathbf{L}$ . Sendo a matriz  $\mathbf{L} \in \mathbb{R}^{M_s \times (M_s+2)}$  uma matriz não quadrada, e como para resolver um sistema precisamos de a inverter, vamos transformar esse sistema de matrizes não quadradas em matrizes quadradas. Tomamos:

$$\mathbf{L}(t) \mathbf{V}(t) = \mathbf{L}_r(t) \mathbf{V}_r(t) + \mathbf{r}(t)$$

em que a matriz  $\mathbf{L}_r \in \mathbb{R}^{M_s \times M_s}$  tem a seguinte disposição:

$$\mathbf{L}_r(t) = \begin{bmatrix} B_1(t) & C_1(t) & & & & & \\ A_2(t) & B_2(t) & C_2(t) & & & & \\ & A_3(t) & B_3(t) & C_3(t) & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & A_{M_s-1}(t) & B_{M_s-1}(t) & C_{M_s-1}(t) & \\ & & & A_{M_s}(t) & B_{M_s}(t) & & \end{bmatrix}. \quad (4.58)$$

Como os valores  $V_0(t)$  e  $V_{M_s+1}(t)$  são conhecidos, pois são relativos à condição de fronteira do nosso problema, corta-se a primeira e última coluna da matriz  $\mathbf{L}$  e multiplica-se respectivamente por  $V_0(t)$  e

$V_{M_S+1}(t)$ , e junta-se tudo num vetor  $\mathbf{r}(t) \in \mathbb{R}^{M_S \times M_S}$ :

$$\mathbf{r}(t) = \begin{bmatrix} V_0(t)A_1(t) \\ 0 \\ \vdots \\ 0 \\ V_{M_S+1}(t)C_{M_S}(t) \end{bmatrix} \quad (4.59)$$

Ficamos com uma semi-discretização do operador diferencial,

$$\frac{\partial}{\partial t} \mathbf{V}_r(t) + \mathbf{L}_r(t) \mathbf{V}_r(t) + \mathbf{r}(t) \quad (4.60)$$

- Para a derivada parcial com respeito ao tempo, usamos a seguinte aproximação:

$$\frac{\partial V}{\partial t} = \frac{V^{n+1} - V^n}{h_t} + O(h_t)$$

substituindo no operador parabólico do modelo de Black-Scholes

$$\frac{V_m^{n+1} - V_m^n}{h_t} + a_m(t) \frac{V_{m+1}(t) - 2V_m(t) + V_{m-1}(t)}{h_S^2} + b_m(t) \frac{V_{m+1} - V_{m-1}}{2h_S} + c_m(t)V_m(t) + O(h_t) + O(h_S^2)$$

Assim, obtemos a discretização

$$\frac{\mathbf{V}_r^{n+1} - \mathbf{V}_r^n}{h_t} + \mathbf{L}_r(t) \mathbf{V}_r(t) + \mathbf{r}(t). \quad (4.61)$$

### 4.3.3 Esquema Explícito (Puro)

O esquema explícito ir-nos-á permitir escrever a solução *explicitamente* sem grandes cálculos adicionais, a partir das iterações anteriores. Uma vez que se trata de um problema de valor final (e não de valor inicial), teremos de resolver o problema *para trás* no tempo, portanto queremos descobrir, solução  $V^n$  através dos dados da iteração anterior em  $n+1$ , para  $n = N_t, \dots, 1, 0$ , usando a discretização:

$$\frac{\mathbf{V}_r^{n+1} - \mathbf{V}_r^n}{h_t} + \mathbf{L}_r^{n+1} \mathbf{V}_r^{n+1} + \mathbf{r}^{n+1}. \quad (4.62)$$

### 4.3.4 Esquema Implícito (Puro)

No esquema implícito, ao contrário do esquema explícito, para encontrar a solução já será necessário resolver um sistema de equações, envolvendo assim cálculos extra, para encontrar a solução  $V^n$ , para  $n = N_t, \dots, 1, 0$ , obtendo o esquema:

$$\frac{\mathbf{V}_r^{n+1} - \mathbf{V}_r^n}{h_t} + \mathbf{L}_r^n \mathbf{V}_r^n + \mathbf{r}^n. \quad (4.63)$$

### 4.3.5 Esquema- $\theta$

O esquema- $\theta$  consiste numa combinação convexa entre o esquema explícito e o esquema implícito, dado  $\theta \in [0, 1]$ :

$$\text{Esquema-}\theta = \theta \text{ Implícito} + (1 - \theta) \text{ Explícito} \quad (4.64)$$

sendo que para  $\theta = 1$  temos o esquema implícito, para  $\theta = 0$  o esquema explícito, e para  $\theta = 1/2$  o esquema de *Crank-Nicolson*. Portanto

$$\begin{aligned} & \theta \left( \frac{\mathbf{V}_r^{n+1} - \mathbf{V}_r^n}{h_t} + \mathbf{L}_r^n \mathbf{V}_r^n + \mathbf{r}^n \right) + (1 - \theta) \left( \frac{\mathbf{V}_r^{n+1} - \mathbf{V}_r^n}{h_t} + \mathbf{L}_r^{n+1} \mathbf{V}_r^{n+1} + \mathbf{r}^{n+1} \right) = \\ &= \frac{\mathbf{V}_r^{n+1} - \mathbf{V}_r^n}{h_t} + \theta [\mathbf{L}_r^n \mathbf{V}_r^n + \mathbf{r}^n] + (1 - \theta) [\mathbf{L}_r^{n+1} \mathbf{V}_r^{n+1} + \mathbf{r}^{n+1}] = \\ &= \left[ \theta \mathbf{L}_r^n - \frac{1}{h_t} \mathbf{I} \right] \mathbf{V}_r^n + \left[ (1 - \theta) \mathbf{L}_r^{n+1} + \frac{1}{h_t} \mathbf{I} \right] \mathbf{V}_r^{n+1} + (1 - \theta) \mathbf{r}^{n+1} + \theta \mathbf{r}^n = \\ &= \mathbf{A}^n \mathbf{V}_r^n + \mathbf{B}^n \mathbf{V}_r^{n+1} + \mathbf{c}^n \end{aligned}$$

para  $n = N_t, \dots, 1, 0$ , onde

$$\mathbf{A}^n = \left[ \theta \mathbf{L}_r^n - \frac{1}{h_t} \mathbf{I} \right] ; \quad \mathbf{B}^n = \left[ (1 - \theta) \mathbf{L}_r^{n+1} + \frac{1}{h_t} \mathbf{I} \right] ; \quad \mathbf{c}^n = (1 - \theta) \mathbf{r}^{n+1} + \theta \mathbf{r}^n.$$

Impondo a condição de valor final  $V_m^{N_t+1} = g_m$  para  $m = 0, \dots, M_S + 1$ , matricialmente:

$$\mathbf{V}^{N_t+1} = \mathbf{g} \quad (4.65)$$

e as condições de contorno  $V_0^n = h_0^n$  e  $V_{M_S+1}^n = h_{M_S+1}^n$  para  $n = 0, \dots, N_t + 1$ , matricialmente:

$$\begin{cases} V_0^n = h_0^n \\ V_{M_S+1}^n = h_{M_S+1}^n \end{cases} \quad \text{para } n = 0, \dots, N_t + 1. \quad (4.66)$$

#### 4.3.5.1 Esquema- $\theta$ para Opções Europeias

Iterativamente calcule-se  $\mathbf{V}_r^n$ , como solução do sistema:

$$\begin{aligned} & \mathbf{A}^n \mathbf{V}_r^n + \mathbf{B}^n \mathbf{V}_r^{n+1} + \mathbf{c}^n = 0 \Leftrightarrow \\ & \Leftrightarrow \mathbf{V}_r^n = [-\mathbf{A}^n]^{-1} (\mathbf{B}^n \mathbf{V}_r^{n+1} + \mathbf{c}^n) \end{aligned} \quad (4.67)$$

para  $n = N_t, \dots, 1, 0$ . Este algoritmo iterativo, está também escrito como pseudocódigo nos anexos em A.4.

### 4.3.5.2 Esquema- $\theta$ para American Black-Scholes

A discretização do problema corresponde a encontrar  $\mathbf{V}_r^n$ , tal que:

$$\begin{cases} \mathbf{A}^n \mathbf{V}_r^n + \mathbf{B}^n \mathbf{V}_r^{n+1} + \mathbf{c}^n \leq \mathbf{0} \\ (\mathbf{g}_r - \mathbf{V}_r^n) \leq \mathbf{0} \\ (\mathbf{A}^n \mathbf{V}_r^n + \mathbf{B}^n \mathbf{V}_r^{n+1} + \mathbf{c}^n)^T (\mathbf{g}_r - \mathbf{V}_r^n) = 0 \\ \mathbf{V}^{N_t+1} = \mathbf{g} \\ V_0^n = h_0^n \\ V_{M_S+1}^n = h_{M_S+1}^n \end{cases}, \text{ para } n = N_t, \dots, 1, 0. \quad (4.68)$$

Repare-se que a partir da 1<sup>a</sup>, 2<sup>a</sup> e 3<sup>a</sup> linha do sistema, podemos tirar uma condição equivalente:

$$\mathbf{0} = \max\{\mathbf{A}^n \mathbf{V}_r^n + \mathbf{B}^n \mathbf{V}_r^{n+1} + \mathbf{c}^n; (\mathbf{g}_r - \mathbf{V}_r^n)\} \quad (4.69)$$

o que pode ser resolvido iterativamente para cada  $n$ , com SSNM. Este algoritmo iterativo, está também escrito como pseudocódigo nos anexos em A.6.

### 4.3.6 Método dos Elementos Finitos para Black-Scholes

#### 4.3.6.1 Formulação fraca

O nosso problema parabólico vai ser governado pelo operador:

$$u_t + \mathcal{L}_{BS} u. \quad (4.70)$$

Seja  $V = H^1(\Omega)$  um espaço de Hilbert, com  $\Omega = (0, S_{max})$ . Assumindo  $u \in L^2(0, T; V)$  é tal que  $u_t \in L^2(0, T; V')$ , multiplicamos (4.70) por uma função teste  $v \in V$ , e integramos em  $\Omega$ :

$$\begin{aligned} & \int_{\Omega} v u_t dx + \int_{\Omega} v \mathcal{L}_{BS} u dx = \\ & = P_d[u_t, v] + B[u, v] \end{aligned} \quad (4.71)$$

onde  $B[., .]$  é um operador bilinear:

$$B[u, v] = -\frac{1}{2} \sigma^2 \int_{\Omega} x^2 \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} dx + (r - \sigma^2) \int_{\Omega} x \frac{\partial u}{\partial x} v dx - r \int_{\Omega} u v dx + \frac{1}{2} \sigma^2 \int_{\partial\Omega} x^2 u_x v dx \quad (4.72)$$

e o produto de dualidade, ou seja, a identificação usual entre um elemento  $v \in V$ , com um elemento  $u_t \in V'$ , do dual:

$$P_d[u_t, v] = \langle u_t, v \rangle_{V' \times V} = \int_{\Omega} u_t v dx. \quad (4.73)$$

#### 4.3.6.2 Malhagem do domínio: Mesh

Fazendo uma discretização do domínio  $\bar{\Omega} = [0, S_{\max}]$ , num total de  $M_x + 2$  pontos:

$$0 = x_0 < x_1 < \dots < x_{M_x} < x_{M_x+1} = S_{\max}$$

Ficamos com um total de  $M_x + 1$  intervalos, da forma:

$$I_m = [x_m, x_{m+1}], \quad m = 0, \dots, M_x$$

cada um com comprimento ou passo:

$$h_m = x_{m+1} - x_m, \quad m = 0, \dots, M_x.$$

Tome-se  $\mathcal{I}_h$  como a malhagem (triangularização, por abuso de notação) do domínio  $\bar{\Omega}$ , e é portanto a família finita de elementos  $I_m$ , que satisfaz:

$$\bar{\Omega} = \bigcup_{I \in \mathcal{I}_h} I$$

onde

$$I_i \cap I_j = \emptyset, \quad \forall I_i, I_j \in \mathcal{I}_h, \quad I_i \neq I_j$$

que dizemos  $\{I, \mathbb{P}_k(I), \Sigma_I\}$  é um elemento finito de Lagrange de grau  $k$ , onde  $I \in \mathcal{I}_h$ , e  $\Sigma_I$  é o conjunto de nós de interpolação de  $I$  e  $\mathbb{P}_k(I)$  é o espaço vetorial de funções polinomiais de grau menor ou igual que  $k$  em  $I$ .

#### 4.3.6.3 Espaço de dimensão finita

Vamos definir  $V_h$ , como um subespaço vetorial de dimensão finita de  $V = H^1(\Omega)$ , por:

$$V_h = \{v_h \in C^0(\bar{\Omega}) : v_h|_I \in \mathbb{P}_1(I), \quad \forall I \in \mathcal{I}_h\} \subset H^1(\Omega) = V. \quad (4.74)$$

Neste caso o número de bases globais coincidirá com o número de nós da discretização. Assim a solução  $u_h \in V_h$ , formada pela combinação linear de funções base  $\{\phi_0, \dots, \phi_{M_S+1}\}$ , ou seja:

$$u_h = \sum_{j=0}^{M_S+1} \phi_j(x) U_j(t).$$

Segue-se que

$$\frac{d}{dt}u_h = \sum_{j=0}^{Ms+1} \phi_j(x)U'_j(t).$$

As funções teste  $v_h \in V_h$  são definidas por:

$$v_h = \sum_{i=0}^{Ms+1} \phi_i(x)\alpha_i$$

onde os coeficientes  $\alpha$  não dependem de  $t$ .

#### 4.3.6.4 Sistema Matricial

Discretizando a expressão (4.71), através da substituição das funções  $v_h$  e  $u_h$  obtemos:

$$\begin{aligned} P_d \left[ \frac{du_h}{dt}, v_h \right] + B[u_h, v_h] &= \\ &= P_d \left[ \sum_{j=0}^{M_x+1} U'_j(t) \cdot \phi_j(x), v_h \right] + B \left[ \sum_{j=0}^{M_x+1} U_j(t) \cdot \phi_j(x), v_h \right] \\ &= \sum_{j=0}^{M_x+1} U'_j(t) \cdot P_d [\phi_j(x), v_h] + \sum_{j=0}^{M_x+1} U_j(t) \cdot B [\phi_j(x), v_h] \end{aligned}$$

Escolhendo sucessivamente as funções teste compostas apenas por uma função base

$$v_h = \phi_i(x), \quad i = 0, 1, \dots, M_s, M_s + 1$$

tem-se

$$\sum_{j=0}^{M_x+1} U'_j(t) \cdot P_d [\phi_j(x), \phi_i(x)] + \sum_{j=0}^{M_x+1} U_j(t) \cdot B [\phi_j(x), \phi_i(x)], \quad i = 0, \dots, M_s + 1$$

o que pode ser visto na forma matricial como

$$\mathbf{P}\mathbf{U}'(t) + \mathbf{B}\mathbf{U}(t) \tag{4.75}$$

seja  $b_{ij} = B[\phi_j, \phi_i]$  e  $p_{ij} = P_d[\phi_j, \phi_i]$ ,  $\mathbf{B} \in \mathbb{R}^{(M_S+2) \times (M_S+2)}$  é uma matriz tridiagonal quadrada com a seguinte configuração:

$$\mathbf{B} = \begin{bmatrix} b_{0,0} & b_{0,1} & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ b_{1,0} & b_{1,1} & b_{1,2} & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & b_{2,1} & b_{2,2} & b_{2,3} & 0 & \dots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & b_{k,k-1} & b_{k,k} & b_{k,k+1} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 & b_{M_x-1,M_x-2} & b_{M_x-1,M_x-1} & b_{M_x-1,M_x} & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & b_{M_x,M_x-1} & b_{M_x,M_x} & b_{M_x,M_x+1} \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 & b_{M_x+1,M_x} & b_{M_x+1,M_x+1} \end{bmatrix} \in \mathbb{R}^{(M_S+2) \times (M_S+2)} \quad (4.76)$$

em que  $b_{ij} = B[\phi_j, \phi_i]$ .  $\mathbf{P} \in \mathbb{R}^{(M_S+2) \times (M_S+2)}$  é também uma matriz tridiagonal quadrada, com uma configuração igual a  $\mathbf{B}$ , em que  $p_{ij} = P_d[\phi_j, \phi_i]$ . Os vetores  $\mathbf{U}(t), \mathbf{U}'(t) \in \mathbb{R}^{M_S+2}$  são dados por:

$$\mathbf{U}(t) = \begin{bmatrix} U_0(t) \\ U_1(t) \\ U_2(t) \\ \vdots \\ U_{M_S}(t) \\ U_{M_S+1}(t) \end{bmatrix} \in \mathbb{R}^{M_S+2} ; \quad \mathbf{U}'(t) = \begin{bmatrix} U'_0(t) \\ U'_1(t) \\ U'_2(t) \\ \vdots \\ U'_{M_S}(t) \\ U'_{M_S+1}(t) \end{bmatrix} \in \mathbb{R}^{M_S+2}. \quad (4.77)$$

No operador bilinear, temos o seguinte integral de fronteira, que convém ser analisado à parte:

$$\int_{\partial\Omega} x^2 u_h v \, dx = \left[ x^2 \frac{\partial u_h(x, t)}{\partial x} v_h(x) \right]_{x_I=x_0}^{x_F=x_{M_S+1}} = x_F^2 u_x(x_F) v(x_F) - x_I^2 u_x(x_I) v(x_I).$$

Para calcular a derivada no espaço, avaliada nos pontos do extremos do intervalo  $[a, b]$ :

$$\frac{\partial u}{\partial x}(a), \quad \frac{\partial u}{\partial x}(b)$$

aproximar (ou estender) a derivada no espaço, por continuidade, nos seus extremos,

$$\frac{\partial u_h}{\partial x}(a^+) = \sum_{k=0}^{M_S+1} \phi'_k(a^+) U_k(t) = \phi'_0(a^+) U_0 + \phi'_1(a^+) U_1 = \frac{-1}{h_0} U_0 + \frac{1}{h_0} U_1$$

$$\frac{\partial u_h}{\partial x}(b^-) = \sum_{k=0}^{M_S+1} \phi'_k(b^-) U_k(t) = \phi'_{M_S}(b^-) U_{M_S} + \phi'_{M_S+1}(b^-) U_{M_S+1} = \frac{-1}{h_{M_S}} U_{M_S} + \frac{1}{h_{M_S}} U_{M_S+1}$$

ou então por diferenças finitas, que irá dar o mesmo:

$$\frac{\partial u}{\partial x}(x_I) = \frac{u(x_I + h_0) - u(x_I)}{h_0} = \frac{-1}{h_0} U_0 + \frac{1}{h_0} U_1$$

$$\frac{\partial u}{\partial x}(x_F) = \frac{u(x_F) - u(x_F - h_{M_s})}{h_{M_s}} = \phi'_{M_s}(x_F^-)U_{M_s} + \phi'_{M_s+1}(x_F^-)U_{M_s+1} = \frac{-1}{h_{M_s}}U_{M_s} + \frac{1}{h_{M_s}}U_{M_s+1}$$

dando um vetor na forma:

$$\begin{aligned} \left[ x^2 \frac{\partial u_h(x, t)}{\partial x} v_h(x) \right]_{a=x_0}^{b=x_{M_s+1}} &= x_{M_s+1}^2 \frac{\partial u_h(x_{M_s+1}, t)}{\partial x} v_h(x_{M_s+1}) - x_0^2 \frac{\partial u_h(x_0, t)}{\partial x} v_h(x_0) = \\ &= \left[ \frac{-x_{M_s+1}}{h_{M_s}} U_{M_s} + \frac{x_{M_s+1}}{h_{M_s}} U_{M_s+1} \right] v_h(x_{M_s+1}) + \left[ \frac{x_0}{h_0} U_0 - \frac{x_0}{h_0} U_1 \right] v_h(x_0) \end{aligned}$$

o que gera um sistema:

$$\begin{bmatrix} \frac{x_0}{h_0} & -\frac{x_0}{h_0} & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & -\frac{x_{M_s+1}}{h_{M_s}} & \frac{x_{M_s+1}}{h_{M_s}} \end{bmatrix} \begin{bmatrix} U_0 \\ U_1 \\ \vdots \\ U_{M_s} \\ U_{M_s+1} \end{bmatrix} = \mathbf{B}_4 \mathbf{U}(t)$$

onde  $\mathbf{B}_4 \in \mathbb{R}^{(M_s+2) \times (M_s+2)}$ .

- Repare-se que  $U_0(t)$  e  $U_{M_s+1}(t)$  são conhecidos para todo o  $t$ , apenas não sabemos os pontos do interior do domínio, ou seja  $U_i(t)$  para  $i = 1, \dots, M_s$ , pelo que a primeira linha do sistema não é necessária, e podemos escrever:

$$\mathbf{P}_R \mathbf{U}'(t) + \mathbf{B}_R \mathbf{U}(t) \quad (4.78)$$

onde  $\mathbf{P}_R \in \mathbb{R}^{(M_s) \times (M_s+2)}$  e  $\mathbf{B}_R \in \mathbb{R}^{(M_s) \times (M_s+2)}$ , têm as mesma configuração de  $\mathbf{P} \in \mathbb{R}^{(M_s+2) \times (M_s+2)}$  e  $\mathbf{B} \in \mathbb{R}^{(M_s+2) \times (M_s+2)}$ , respetivamente, só que sem a primeira e última linha.

$$\mathbf{B}_R = \begin{bmatrix} b_{1,0} & b_{1,1} & b_{1,2} & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & b_{2,1} & b_{2,2} & b_{2,3} & 0 & \dots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & b_{k,k-1} & b_{k,k} & b_{k,k+1} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 & b_{M_x-1,M_x-2} & b_{M_x-1,M_x-1} & b_{M_x-1,M_x} & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & b_{M_x,M_x-1} & b_{M_x,M_x} & b_{M_x,M_x+1} \end{bmatrix} \in \mathbb{R}^{(M_s+2) \times (M_s+2)}. \quad (4.79)$$

### 4.3.7 Diferenças finitas no tempo: Esquema- $\theta$

Depois de termos discretizado o problema no espaço, agora temos de lidar com a variável no tempo, para isso recorreremos às diferenças finitas.

- O intervalo  $[0, T]$  é dividido em  $N_t + 1$  sub-intervalos de igual comprimento com comprimento  $h_t$ :

$$h_t = \frac{t_{\max} - t_{\min}}{N_t + 1} \Leftrightarrow T = t_{\min} + h_t \cdot N_t$$

no total, com  $N_t + 2$  pontos, da variável tempo:

$$t(n) = t_n = n \cdot \Delta t, \quad n = \{0, 1, \dots, N_t, N_t + 1\}.$$

Portanto

$$\begin{aligned} & \theta \left( \frac{1}{h_t} \mathbf{P}_R (\mathbf{U}^{n+1} - \mathbf{U}^n) + \mathbf{B}_R \mathbf{U}^n \right) + (1 - \theta) \left( \frac{1}{h_t} \mathbf{P}_R (\mathbf{U}^{n+1} - \mathbf{U}^n) + \mathbf{B}_R \mathbf{U}^{n+1} \right) = \\ &= \frac{1}{h_t} \mathbf{P}_R (\mathbf{U}^{n+1} - \mathbf{U}^n) + \theta \mathbf{B}_R \mathbf{U}^n + (1 - \theta) \mathbf{B}_R \mathbf{U}^{n+1} = \\ &= \left[ \theta \mathbf{B}_R - \frac{1}{h_t} \mathbf{P}_R \right] \mathbf{U}^n + \left[ (1 - \theta) \mathbf{B}_R + \frac{1}{h_t} \mathbf{P}_R \right] \mathbf{U}^{n+1} = \\ &= \mathbf{E}_R \mathbf{U}^n + \mathbf{H}_R \mathbf{U}^{n+1} = \\ &= \mathbf{E}_r \mathbf{U}_r^n + \mathbf{r}_E^n + \mathbf{H}_R \mathbf{U}^{n+1} \end{aligned} \tag{4.80}$$

onde  $\mathbf{E}_R, \mathbf{H}_R \in \mathbb{R}^{(M_S) \times (M_S+2)}$ . Como os valores  $U_0(t)$  e  $U_{M_S+1}(t)$  já são conhecidos, pois são relativos à condição de fronteira do nosso problema, corta-se a primeira e última coluna da matriz  $\mathbf{E}_R$ , transformando-a numa matriz quadrada  $\mathbf{E}_r$ , e multiplica-se respetivamente a primeira e última coluna de  $\mathbf{E}_R$  por  $U_0(t)$  e  $U_{M_S+1}(t)$ , e soma-se tudo num vetor  $\mathbf{r}_E(t) \in \mathbb{R}^{M_S}$ :

$$\mathbf{r}_E(t) = \begin{bmatrix} E_{1,0} U_0(t) \\ 0 \\ \vdots \\ 0 \\ E_{M_S, M_S+1} U_{M_S+1}(t) \end{bmatrix} \in \mathbb{R}^{M_S} \tag{4.81}$$

Obtemos assim o esquema- $\theta$  para  $n = N_t, \dots, 1, 0$ :

$$\mathbf{E}_r \mathbf{U}_r^n + \mathbf{r}_E^n + \mathbf{H}_R \mathbf{U}^{n+1} \tag{4.82}$$

Impondo a condição de valor final  $U_m^{N_t+1} = g_m$  para  $m = 0, \dots, M_S + 1$ , matricialmente:

$$\mathbf{U}^{N_t+1} = \mathbf{g} \tag{4.83}$$

e as condições de contorno  $U_0^n = h_0^n$  e  $U_{M_S+1}^n = h_{M_S+1}^n$  para  $n = 0, \dots, N_t + 1$ , matricialmente:

$$\begin{cases} U_0^n = h_0^n \\ U_{M_S+1}^n = h_{M_S+1}^n \end{cases} \quad \text{para } n = 0, \dots, N_t + 1. \tag{4.84}$$

#### 4.3.7.1 Esquema- $\theta$ para European Black-Scholes

Iterativamente calcule-se  $\mathbf{U}_r^n$ , como solução do sistema:

$$\begin{aligned} \mathbf{E}_r \mathbf{U}_r^n + \mathbf{r}_E^n + \mathbf{H}_R \mathbf{U}^{n+1} &= 0 \Leftrightarrow \\ \Leftrightarrow \mathbf{U}_r^n &= [-\mathbf{E}_r]^{-1}(\mathbf{r}_E^n + \mathbf{H}_R \mathbf{U}^{n+1}) \end{aligned} \quad (4.85)$$

para  $n = N_t, \dots, 1, 0$ . Este algoritmo iterativo, está também escrito como pseudocódigo nos anexos em A.5.

#### 4.3.7.2 Esquema- $\theta$ para American Black-Scholes

A discretização do problema corresponde a encontrar  $\mathbf{U}_r^n$ , tal que:

$$\left\{ \begin{array}{l} \mathbf{E}_r \mathbf{U}_r^n + \mathbf{r}_E^n + \mathbf{H}_R \mathbf{U}^{n+1} \leq \mathbf{0} \\ (\mathbf{g}_r - \mathbf{U}_r^n) \leq \mathbf{0} \\ (\mathbf{E}_r \mathbf{U}_r^n + \mathbf{r}_E^n + \mathbf{H}_R \mathbf{U}^{n+1})(\mathbf{g}_r - \mathbf{V}_r^n) = \mathbf{0} \\ \mathbf{U}^{N_t+1} = \mathbf{g} \\ U_0^n = h_0^n \\ U_{M_S+1}^n = h_{M_S+1}^n \end{array} \right. , \text{ para } n = N_t, \dots, 1, 0. \quad (4.86)$$

Repare-se que a partir da 1<sup>a</sup>, 2<sup>a</sup> e 3<sup>a</sup> linha do sistema, podemos tirar uma condição equivalente ao problema:

$$\mathbf{0} = \max\{\mathbf{E}_r \mathbf{U}_r^n + \mathbf{r}_E^n + \mathbf{H}_R \mathbf{U}^{n+1}; (\mathbf{g}_r - \mathbf{U}_r^n)\}, \text{ para } n = N_t, \dots, 1, 0 \quad (4.87)$$

o que pode ser resolvido iterativamente para cada  $n$ , com SSNM. Este algoritmo iterativo, está também escrito como pseudocódigo nos anexos em A.7.

#### 4.3.8 Consistência, Estabilidade e Convergência de FDM e FEM com o esquema- $\theta$ no tempo

As questões de consistência, estabilidade e convergência para métodos numéricos aplicados à equação de Black-Scholes estão bem estabelecidas na literatura académica. De acordo com o capítulo 18 do livro [23], também referido em [24], esses métodos demonstraram ser robustos e confiáveis numa ampla gama de aplicações financeiras.

Para o caso do esquema implícito (quando  $\theta = 1$ ), temos garantia de estabilidade. Portanto, optaremos por esse método nas implementações futuras, assegurando soluções numéricas estáveis e precisas para problemas financeiros.

# Capítulo 5

## Resolução Numérica e resultados

Neste capítulo implementamos os métodos do capítulo anterior aos problemas do obstáculo, a problemas práticos concretos, e analisamos os resultados.

### 5.1 Problema Elíptico: Problema Estacionário - Membrana Elástica

#### 5.1.1 Problema prático: Domínio Circular

Considere o seguinte problema (ver [10]):

Seja  $\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 4\}$  (domínio circular), e considere o sistema

$$\begin{cases} -\Delta u \geq f & , \text{em } \Omega \\ u - g \geq 0 & , \text{em } \Omega \\ (u - g)(\Delta u - f) = 0 & , \text{em } \Omega \\ u = 0 & , \text{sobre } \partial\Omega \end{cases} \quad (5.1)$$

com os seguintes dados:

$$\begin{cases} g(r) = \begin{cases} \sqrt{1-r^2}, & r < b \\ c_1r + c_2, & c.c. \end{cases} \\ f(x, y) = -1 \end{cases} \quad (5.2)$$

onde  $r = \sqrt{x^2 + y^2}$  é a distância à origem, e  $c_1$  e  $c_2$  são escolhidos tal que o obstáculo é  $g \in C^1(\Omega)$ , e a constante  $b = 0.9$ .

#### 5.1.2 Solução Analítica

Introduzindo o multiplicador de Lagrange, o problema é reformulado como:

$$\begin{cases} -\Delta u - \lambda = f & , \text{em } \Omega \\ g - u \leq 0 & , \text{em } \Omega \\ \lambda \geq 0 & , \text{em } \Omega \\ (u - g)\lambda = 0 & , \text{em } \Omega \\ u = 0 & , \text{sobre } \partial\Omega. \end{cases} \quad (5.3)$$

A simetria radial reduz o problema a uma ODE:

$$\begin{cases} \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \lambda = 1, & x \in \Omega \\ g - u \leq 0, & x \in \Omega \\ \lambda \geq 0, & x \in \Omega \\ (u - g)\lambda = 0, & x \in \Omega \\ u = 0, & x \in \partial\Omega \end{cases}$$

A primeira e segunda derivadas da função  $g(r)$  (obstáculo):

$$g'(r) = \begin{cases} \frac{-r}{(1-r^2)^{1/2}}, & r < b \\ c_1, & c.c. \end{cases} \quad (5.4)$$

$$g''(r) = \begin{cases} \frac{-1}{(1-r^2)^{1/2}} + \frac{r^2}{(1-r^2)^{3/2}}, & r < b \\ 0, & c.c. \end{cases} \quad (5.5)$$

sendo  $g \in C^1(\Omega)$ , podemos tirar as constantes  $c_1$  e  $c_2$  pela continuidade da sua derivada:

$$\begin{cases} \lim_{r \rightarrow b^-} g'(r) = \lim_{r \rightarrow b^+} g'(r) \\ \lim_{r \rightarrow b^-} g(r) = \lim_{r \rightarrow b^+} g(r) \end{cases} \Leftrightarrow \begin{cases} \frac{-b}{(1-b^2)^{1/2}} = c_1 \\ (1-b^2)^{1/2} = c_1 b + c_2 \end{cases} \Leftrightarrow \begin{cases} c_1 = \frac{-b}{(1-b^2)^{1/2}} \\ c_2 = (1-b^2)^{1/2} + \frac{b^2}{(1-b^2)^{1/2}} \end{cases}$$

e solução analítica:

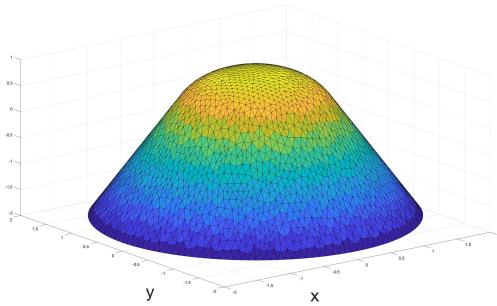
$$u(r) = \frac{(-a^2 + 4g(a) + r^2) \cdot \log(2) + (4 - r^2) \cdot \log(a) + (a^2 - 4 - 4g(a)) \cdot \log(r)}{4(\log(2) - \log(a))} \quad (5.6)$$

e a sua derivada:

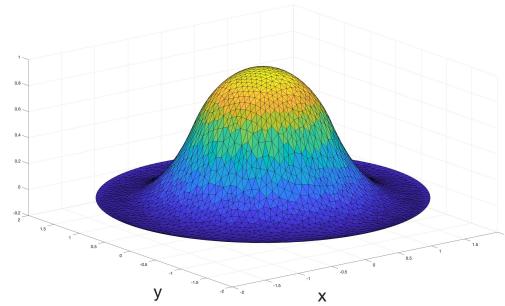
$$u'(r) = \frac{2r(\log(2) - \log(a)) + (a^2 - 4 - 4g(a)) \cdot \frac{1}{r}}{4(\log(2) - \log(a))} \quad (5.7)$$

com um valor aproximadamente de  $a \approx 0.8294$ .

Apresentamos nas figuras 5.1 e 5.2, respectivamente a solução e o obstáculo do problema.



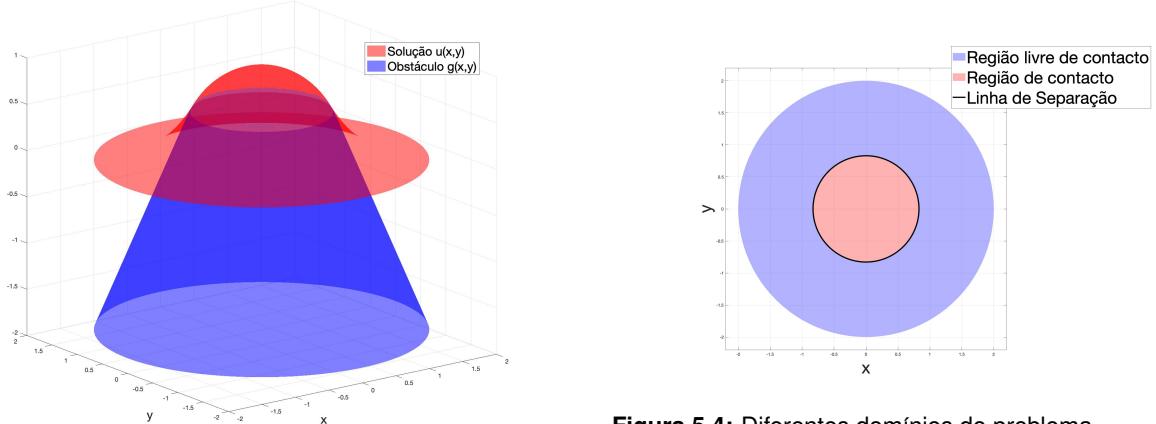
**Figura 5.1:** Função obstáculo  $g(x, y)$



**Figura 5.2:** Solução  $u(x, y)$

Na figura 5.3 apresentamos a membrana sobreposta sobre o obstáculo, e na figura 5.4 apresenta-

mos a curva de separação (fronteira livre) entre a região de contacto e a região livre de contacto.

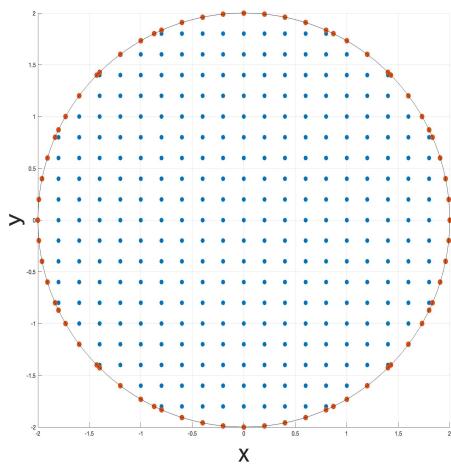


**Figura 5.4:** Diferentes domínios do problema.

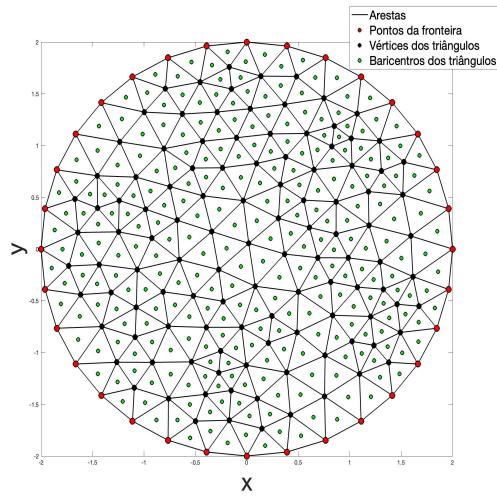
**Figura 5.3:** Funções  $u(x,y)$  e  $g(x,y)$  sobrepostas.

### 5.1.3 Implementação dos Algoritmos

Para resolver este problema numericamente foram implementados três métodos diferentes, um deles usam diferenças finitas e os outros dois elementos finitos. São dois métodos distintos, cuja geometria de discretização do domínio, gerada por cada um deles é diferente. Na figuras 5.5 e 5.6 mostramos respetivamente a discretização para cada um dos métodos.



**Figura 5.5:** Grelha da discretização do domínio circular  $\Omega$ , para diferenças finitas



**Figura 5.6:** Malha triangular do domínio circular  $\Omega$ , usado para elementos finitos

### 5.1.3.1 Discretização com diferenças finitas e implementação do SSNM para o LCP

Com o objetivo de discretizar diretamente a formulação forte do problema com diferenças finitas apresentamos o domínio discretizado na figura 5.5. Tal como descrito na secção 4.2.1, sendo  $\mathbf{x} = \mathbf{u}(\mathbf{i}) \in \mathbb{R}^{N_I}$ ,  $\mathbf{M} = \mathbf{A}(\mathbf{i}_I, \mathbf{i}_I) \in \mathbb{R}^{N_I \times N_I}$  e  $\mathbf{b} = \mathbf{A}(\mathbf{i}_I, \mathbf{i}_N) + \mathbf{f}(\mathbf{i}_I) \in \mathbb{R}^{N_I}$  e  $\mathbf{g}(\mathbf{i}_I) = \mathbf{h} \in \mathbb{R}^{N_I}$ , reescrevemos o problema de complementariedade como encontrar  $\mathbf{x}$  tal que:

$$\begin{cases} \mathbf{M}\mathbf{x} + \mathbf{b} \leq \mathbf{0} \\ \mathbf{h} - \mathbf{x} \leq \mathbf{0} \\ (\mathbf{h} - \mathbf{x})^T(\mathbf{M}\mathbf{u} + \mathbf{b}) = 0 \end{cases} \quad (5.8)$$

Baseando-se na sua expressão de mínimo/máximo equivalente:

$$\max\{\mathbf{M}\mathbf{x} + \mathbf{b}; \mathbf{h} - \mathbf{x}\} = \mathbf{0} \quad (5.9)$$

resolvemos o problema através do método iterativo de SSNM descrito em anexo A.1.

Apresentam-se na tabela 5.1, para diferentes valores de  $h$ , os erros na norma do máximo entre a solução numérica e solução analítica, uma coluna com os graus de liberdade DOF (*degrees of freedom*), e uma coluna com o número de iterações, e finalmente uma coluna para o tempo que o algoritmo demorou.

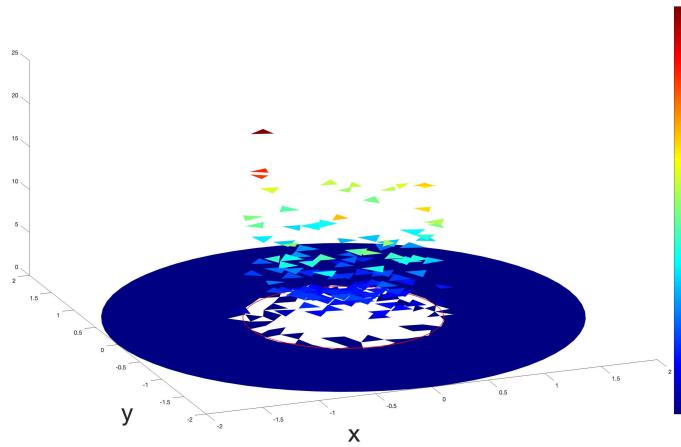
$h_{max}$	DOF	iter	$\ e_h\ _\infty$	tempo (s)
0.8	16	4	7.6336e-02	0.001767
0.4	69	5	2.2057e-02	0.001698
0.2	305	9	1.5153e-02	0.006905
0.1	1245	15	4.2743e-03	0.069570
0.05	5013	29	2.0576e-03	4.644646
0.025	20069	54	3.5914e-04	282.017176

**Tabela 5.1:** Discretização com diferenças finitas, e resolução do LCP com SSNM, com tolerância  $TOL = 10^{-12}$

### 5.1.3.2 Formulação com multiplicadores de Lagrange: elementos finitos mistos

Neste método foram usadas elementos  $\mathbb{P}_1$ -bolha/ $\mathbb{P}_0$ , ou seja elementos de Lagrange de primeira ordem enriquecidos de funções bolha, para as funções base que compõem as funções teste  $v_h$  e a solução  $u_h$ , e polinómios de grau zero para as bases que compõem  $\lambda_h$ . As funções de bolha são polinómios de grau três que se anulam nos três lados do triângulo e tomam o valor um no baricentro do triângulo, tal como descrito na secção 4.2.2.1.

Para determinar a solução numérica para este problema com este tipo de discretização foi implementado o algoritmo denominado como *Primal-dual active set*, descrito em anexo A.2. Na figura 5.7 apresenta-se o valor do multiplicador de Lagrange,  $\lambda_h$ , ao longo do domínio.



**Figura 5.7:** Valores de  $\lambda_h$  obtidos para cada elemento com o Primal-dual active set para o método misto com  $h = 0.2$

Apresentamos uma tabela 5.2, onde para diferentes malhas, temos uma coluna com o número de graus de liberdade, uma coluna para o número de iterações, e três colunas para o erro de  $u_h$ , uma na norma de  $L^2$ , outra na norma de  $H^1$ , e ou outra na norma do máximo, e ainda uma coluna com o tempo que o algoritmo demorou a correr.

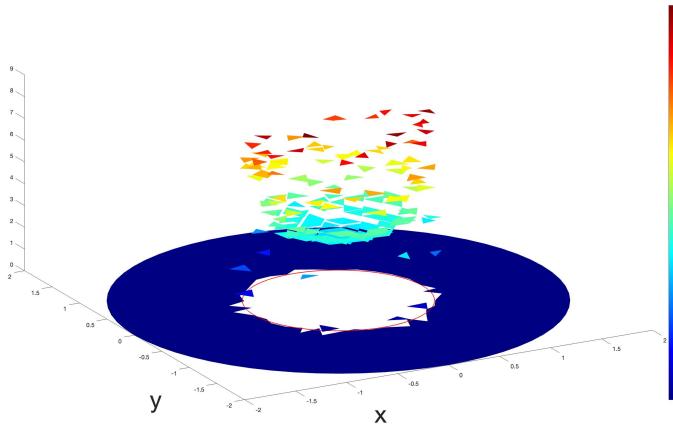
$h_{max}$	DOF	iter	$\ e_h\ _{L^2}$	$\ e_h\ _{H^1}$	$\ e_h\ _\infty$	tempo (s)
0.8	189	5	2.2430e-01	1.1999	1.7276e-01	0.015605
0.4	874	7	5.3611e-02	6.0700e-01	4.6393e-02	0.014635
0.2	3546	10	1.5437e-02	4.3343e-01	1.7621e-02	0.031837
0.1	18325	15	3.4128e-03	3.8091e-01	4.7721e-03	0.375181
0.05	76514	23	4.5392e-04	3.7121e-01	2.0865e-03	1.095855
0.025	333090	39	1.2857e-04	3.6651e-01	3.3101e-04	68.359351

**Tabela 5.2:** Resultados numéricos para o método misto, com  $TOL = 10^{-32}$

### 5.1.3.3 Formulação com multiplicadores de lagrange: elementos finitos com método de estabilização

Neste método foram usadas elementos  $\mathbb{P}_1-\mathbb{P}_0$ . E como visto na secção 4.2.2.2, com este tipo de discretização foi implementado também o algoritmo *Primal-dual active set*, aplicado ao método estabilizado, algortimo este descrito em A.3.

Na figura 5.8 apresenta- se o valor do multiplicador de Lagrange,  $\lambda_h$ , ao longo do domínio, após aplicar o método.



**Figura 5.8:** Valores de  $\lambda$  obtidos para cada elemento com o Primal-dual active set para o método estabilizado com  $h = 0.2$

E na tabela 5.3, implementámos o algoritmo para várias discretizações, onde apresentamos os erros obtidos e mais alguns dados de interesse tal como no método anterior.

$h_{max}$	DOF	iter	$\ e_h\ _{L^2}$	$\ e_h\ _{H^1}$	$\ e_h\ _\infty$	tempo (s)
0.8	77	4	2.8823e-01	9.7053e-01	9.8367e-02	0.011604
0.4	366	7	6.5068e-02	5.4807e-01	3.2832e-02	0.016358
0.2	1502	9	1.7144e-02	4.2510e-01	1.0823e-02	0.021827
0.1	7813	14	3.0908e-03	3.7831e-01	2.5104e-03	0.153261
0.05	32710	22	7.7914e-04	3.6860e-01	6.6608e-04	1.006936
0.025	142582	39	1.9219e-04	3.6621e-01	1.6308e-04	8.954543

**Tabela 5.3:** Resultados numéricos para o método estabilizado, com  $TOL = 10^{-32}$

#### 5.1.4 Análise e comparação dos resultados

Embora os valores de  $h_{max}$  sejam iguais nas tabelas 5.1, 5.2 e 5.3, é crucial notar que a geometria da discretização é diferente, para FDM e para o FEM. No FDM,  $h_{max}$  representa o espaçamento uniforme entre os nós de uma malha, sendo esta uma grelha geralmente regular (excepto no bordo) e cada ponto da grelha está equidistante dos seus vizinhos. Assim, no FDM,  $h_{max}$  é a distância máxima entre os pontos da grelha na discretização do domínio. No FEM,  $h_{max}$  refere-se ao maior comprimento de aresta dos elementos na malha, que pode ser não uniforme e composta de diferentes tipos de elementos.

A análise dos três métodos (diferenças finitas, elementos finitos mistos e estabilizados) mostra que todos convergem para a solução exata com o refinamento da malha, comprovado pela diminuição consistente dos erros ( $\|e_h\|_{L^2}$ ,  $\|e_h\|_{H^1}$ ,  $\|e_h\|_\infty$ ).

Comparando os métodos, observamos que o método com diferenças finitas atinge uma precisão

razoável com menos DOF para malhas grosseiras, mas os DOF aumentam rapidamente, tornando-o ineficiente para malhas finas devido ao alto custo computacional. O método misto oferece alta precisão, mas exige um número muito maior de DOF e, consequentemente, maior tempo de execução, especialmente para malhas finas. O método estabilizado, por outro lado, requer menos DOF que o método misto para alcançar precisão semelhante, resultando em tempos de execução menores.

Concluindo, o método estabilizado é geralmente o mais eficiente em termos de uso de recursos computacionais, oferecendo uma boa combinação de precisão e tempo de execução. O método misto pode ser preferido quando a precisão é extremamente crítica e há disponibilidade de recursos computacionais. O método com diferenças finitas é adequado para soluções rápidas em malhas grosseiras, mas acaba por tornar-se menos prático para malhas finas devido ao aumento exponencial de DOF e tempo de execução.

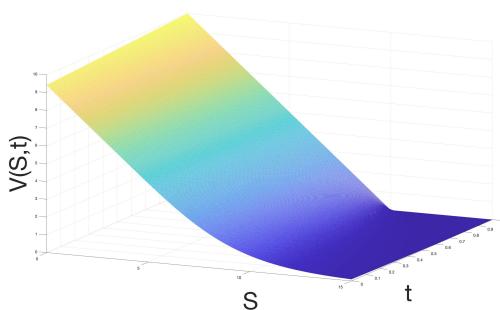
Para problemas onde o tempo de execução é uma preocupação crucial, o Stabilized FEM pode ser a escolha preferida devido à sua capacidade de fornecer alta precisão com menos DOF e tempos de execução menores comparados ao Mixed FEM.

## 5.2 Modelo de Black-Scholes

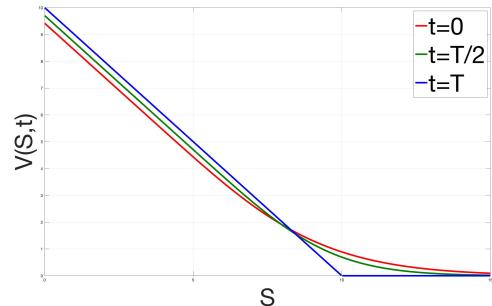
O objetivo principal na precificação de opções é determinar o valor presente da opção, dado por  $V(S_0, t_0)$ , onde  $S_0$  é o preço inicial do ativo subjacente e  $t_0$  é o tempo inicial. Mas para além da relevância prática, avaliar o erro em  $V(S_0, t_0)$  evita as oscilações e a distorção causada por valores extremos da função, o que permite uma análise mais consistente e significativa do erro naquilo que realmente importa: o valor presente da opção.

### 5.2.1 Opcões Europeias

Implementou-se o esquema implícito no tempo ( $\theta = 1$ ) para os dois tipos de discretização feitas no espaço, uma com elementos finitos de grau 1 e a outra com diferenças finitas, correspondendo respectivamente aos algoritmos A.5 e A.4. Para Opções Europeias de venda, definiu-se  $T = 1$ ,  $r = 0.06$ ,  $\sigma = 0.3$ ,  $K = 10$ ,  $S_{max} = 200$ , e apresentamos nas figuras 5.9 e 5.10 a respetiva solução para  $V(S, t)$ .



**Figura 5.9:** Superfície  $V(S, t)$  para opções de venda Europeias, obtida com:  $T = 1$ ,  $r = 0.06$ ,  $\sigma = 0.3$ ,  $K = 10$



**Figura 5.10:**  $V(S, t)$  para opções de venda Europeias, obtida com  $T = 1$ ,  $r = 0.06$ ,  $\sigma = 0.3$ ,  $K = 10$ , avaliada em três momentos diferentes:  $t = T$ ,  $t = T/2$  e  $t = 0$

De modo a avaliar os algoritmos, calculou-se  $V(S_0, 0)$ , para o valor  $S_0 = 8$ , e comparou-se com o valor da solução analítica  $V_{ana}(S_0, 0) \approx 1.8956$ , calculando o respetivo erro, dado pelo módulo da diferença entre o valor analítico e o valor obtido com os respetivos métodos  $e = |V_{ana}(S_0, 0) - V(S_0, 0)|$ . Apresenta-se na tabela 5.4 e 5.5, os erros calculados com diferentes tipos de discretização, para os dois algoritmos.

		$h_t$					
		0.1	0.05	0.025	0.0125	0.00625	0.003125
$h_S$	0.5	3.9798e-03	3.5290e-03	3.2412e-03	3.0811e-03	2.9970e-03	2.9540e-03
	0.25	1.8822e-03	1.3860e-03	1.0777e-03	9.0806e-04	8.1930e-04	7.7393e-04
	0.125	1.3564e-03	8.4942e-04	5.3629e-04	3.6433e-04	2.7447e-04	2.2856e-04
	0.0625	1.2248e-03	7.1522e-04	4.0089e-04	2.2838e-04	1.3824e-04	9.2196e-05
	0.03125	1.1919e-03	6.8167e-04	3.6704e-04	1.9439e-04	1.0418e-04	5.8105e-05
	0.015625	1.1837e-03	6.7328e-04	3.5858e-04	1.8589e-04	9.5668e-05	4.9582e-05

**Tabela 5.4:** Erros para opções de venda Europeias, com FDM no espaço e esquema implícito no tempo, em  $V(S_0, 0)$ , com:  $T = 1$ ,  $r = 0.06$ ,  $\sigma = 0.3$ ,  $K = 10$ ,  $S_0 = 8$ ,  $S_{max} = 200$

		$h_t$					
		0.1	0.05	0.025	0.0125	0.00625	0.003125
$h_S$	0.5	4.0493e-03	3.4940e-03	3.1558e-03	2.9712e-03	2.8751e-03	2.8260e-03
	0.25	1.8990e-03	1.3774e-03	1.0568e-03	8.8109e-04	7.8937e-04	7.4253e-04
	0.125	1.3606e-03	8.4727e-04	5.3107e-04	3.5763e-04	2.6703e-04	2.2075e-04
	0.0625	1.2259e-03	7.1469e-04	3.9959e-04	2.2671e-04	1.3638e-04	9.0247e-05
	0.03125	1.1922e-03	6.8154e-04	3.6672e-04	1.9397e-04	1.0372e-04	5.7617e-05
	0.015625	1.1838e-03	6.7325e-04	3.5850e-04	1.8579e-04	9.5552e-05	4.9460e-05

**Tabela 5.5:** Erros para opções de venda Europeias, com FEM no espaço e esquema implícito no tempo, em  $V(S_0, 0)$ , com:  $T = 1$ ,  $r = 0.06$ ,  $\sigma = 0.3$ ,  $K = 10$ ,  $S_0 = 8$ ,  $S_{max} = 200$

Destes dados pode-se estimar a ordem de convergência:

$$|e_h| \approx C|h|^p$$

aplicando o logaritmo temos

$$\log_2 |e_h| \approx \log_2 C + p \log_2 |h|$$

onde a estimativa da ordem de convergência será dada pela formula:

$$p \sim \frac{\log_2 |e_h| - \log_2 |e_{h/2}|}{\log_2 |h| - \log_2 |h/2|} = \log_2 \frac{|e_h|}{|e_{h/2}|}$$

fixando um passo  $h_S = 0.015625$ , e para diferentes discretizações no tempo obteve-se a convergência para os métodos implementados:

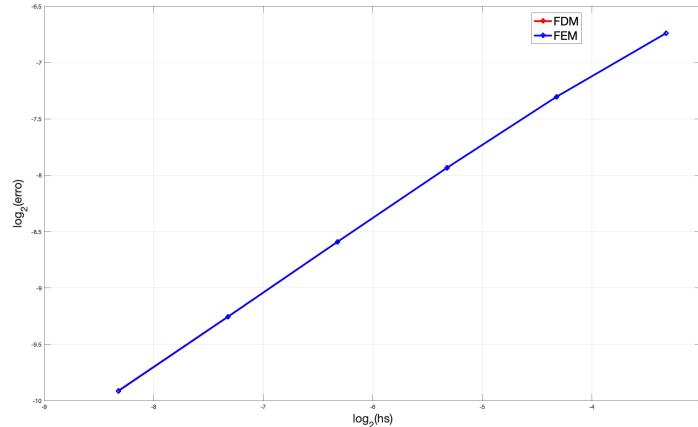
	$h_t$	$h_t/2$	$h_t/4$	$h_t/8$	$h_t/16$	$h_t/32$
$ e_h $	1.1837e-03	6.7328e-04	3.5858e-04	1.8589e-04	9.5668e-05	4.9582e-05
$p$	0.8140	0.9089	0.9478	0.9583	0.9482	-

**Tabela 5.6:** FDM com  $h_t = 0.1$ ,  $h_S = 0.015625$

	$h_t$	$h_t/2$	$h_t/4$	$h_t/8$	$h_t/16$	$h_t/32$
$ e_h $	1.1838e-03	6.7325e-04	3.5850e-04	1.8579e-04	9.5552e-05	4.9460e-05
$p$	0.8142	0.9092	0.9483	0.9593	0.9500	-

**Tabela 5.7:** FEM com  $h_t = 0.1$ ,  $h_S = 0.015625$

e trançando o gráfico em escala logarítmica, presente na figura 5.11



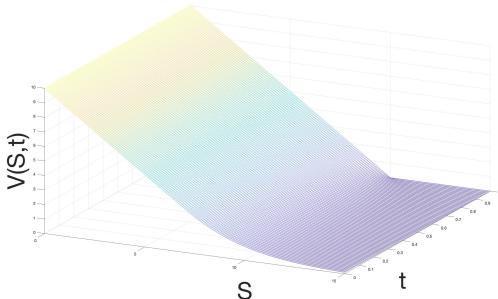
**Figura 5.11:** Convergência do erro na escala logarítmica

A análise dos resultados dos métodos de diferenças finitas e elementos finitos para o cálculo do valor presente das opções europeias mostra que ambos os métodos convergem de maneira consistente, com a ordem de convergência aproximada de 1, tal como esperado teoricamente. A comparação dos erros calculados com diferentes discretizações no tempo e no espaço demonstra que os erros diminuem de forma previsível à medida que os passos  $h_S$  e  $h_t$  são reduzidos. Isso sugere que os métodos implementados são eficazes e precisos na solução do problema de precificação de opções europeias, validando a escolha dos algoritmos utilizados.

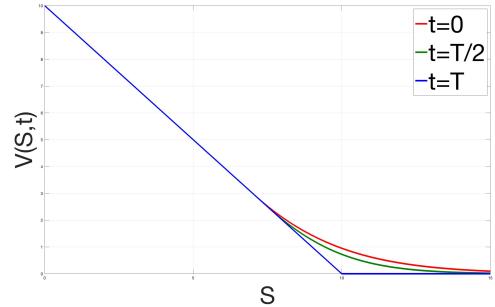
### 5.2.2 Opções Americanas

Para finalmente resolver o problema do obstáculo parabólico, derivado das Opções Americanas, tal como nas opções Europeias, implementou-se o esquema implícito no tempo ( $\theta = 1$ ) para os dois tipos de discretização feitas no espaço, uma com elementos finitos de grau 1 e a outra com diferenças finitas, correspondendo respetivamente aos algoritmos, onde para cada iteração no tempo temos de resolver um problema de complementaridade, que também será resolvido iterativamente com o Semi-Smooth Newton Method (SSNM),correspondendo aos algoritmos A.6 e A.7.

E tal como nas opções Europeias, escolheu-se implementar os algoritmos para opções de venda (Put Options), definiu-se  $T = 1$ ,  $r = 0.06$ ,  $\sigma = 0.3$ ,  $K = 10$ ,  $S_{max} = 200$ , e uma tolerância para o SSNM de  $TOL = 10^{-12}$  e apresentamos nas figuras 5.12 e 5.13 a respetiva solução para  $V(S, t)$ , obtidas com o algoritmo A.7.

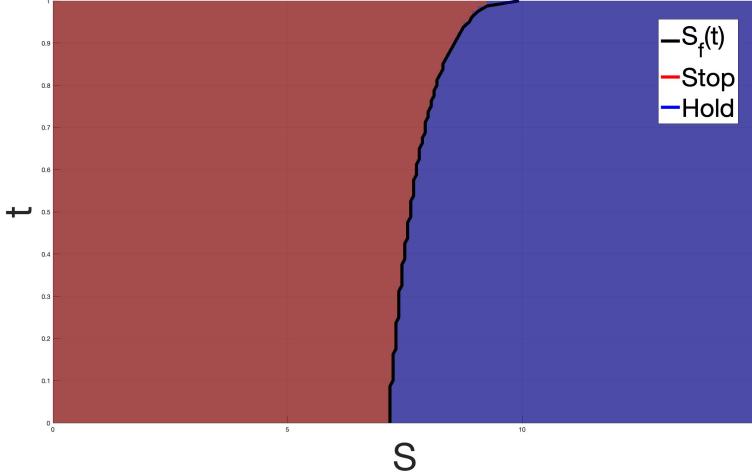


**Figura 5.12:** Superfície  $V(S, t)$  para opções de venda Americanas, obtida com:  $T = 1$ ,  $r = 0.06$ ,  $\sigma = 0.3$ ,  $K = 10$



**Figura 5.13:**  $V(S, t)$  para opções de venda Americanas, obtida com  $T = 1$ ,  $r = 0.06$ ,  $\sigma = 0.3$ ,  $K = 10$ , avaliada em três momentos diferentes:  $t = T$ ,  $t = T/2$  e  $t = 0$

E apresentamos ainda na figura 5.14 que mostra a curva  $S_f(t)$ , que separa as duas regiões distintas, uma onde a solução coincide com a função obstáculo, chamada de região de espera (hold region), e a região de exercício da opção (ou região de paragem - stop region).



**Figura 5.14:** Curva de separação  $S_f(t)$  da região de paragem (stop) e da região de continuação (hold), para uma opção americana de venda obtida com os parâmetros:  $r = 0.06$ ,  $\sigma = 0.3$ ,  $T = 1$  e  $K = 10$ ,  $S_{max} = 15$ .

Para os 2 algoritmos implementados, apresenta-se nas tabelas 5.8 e 5.9 o valor da opção  $V(S_0, 0)$  para diferentes valores de  $S_0$ , e para diferentes discretizações no espaço e no tempo ( $h_S$  e  $h_t$ ).

		$(h_S, h_t)$	$(h_S, h_t)/2$	$(h_S, h_t)/4$	$(h_S, h_t)/8$
$S_0$	6	6	6	6	6
	7	4	4	4	4
	8	2.091784	2.098640	2.101516	2.102763
	10	0.930211	0.942267	0.947756	0.950406
	12	0.381158	0.388353	0.391762	0.393449
	14	0.151940	0.153277	0.153922	0.154257
	16	0.061117	0.059861	0.059193	0.058859

**Tabela 5.8:** Valor  $V(S_0, 0)$  para diferentes  $S_0$  de uma opção de venda Americana, com FDM no espaço e esquema implícito no tempo, e SSNM para cada iteração, com:  $T = 1$ ,  $r = 0.06$ ,  $\sigma = 0.3$ ,  $K = 10$ ,  $S_{max} = 200$ ,  $TOL = 10^{-12}$ ,  $h_S = 0.5$ ,  $h_t = 0.1$  (Algoritmo A.6).

		$(h_S, h_t)$	$(h_S, h_t)/2$	$(h_S, h_t)/4$	$(h_S, h_t)/8$
$S_0$	6	6	6	6	6
	7	4	4	4	4
	8	2.093145	2.099071	2.101635	2.102794
	10	0.933621	0.943157	0.947992	0.950467
	12	0.383015	0.388861	0.391901	0.393485
	14	0.152263	0.153387	0.153957	0.154267
	16	0.060892	0.059810	0.059183	0.058857

**Tabela 5.9:** Valor  $V(S_0, 0)$  para diferentes  $S_0$  de uma opção de venda Americana, com FEM no espaço e esquema implícito no tempo, e SSNM para cada iteração, com:  $T = 1$ ,  $r = 0.06$ ,  $\sigma = 0.3$ ,  $K = 10$ ,  $S_{max} = 200$ ,  $TOL = 10^{-12}$ ,  $h_S = 0.5$ ,  $h_t = 0.1$  (Algoritmo A.7).

De modo a comparar os dois métodos, uma vez que não existe solução analítica para este problema,

calculamos os erros relativos entre os 2 algoritmos,  $e_{\text{rela}} = |V_{\text{FEM}}(S_0, 0) - V_{\text{FDM}}(S_0, 0)|$ , que apresentamos na tabela 5.10

		$(h_S, h_t)$	$(h_S, h_t)/2$	$(h_S, h_t)/4$	$(h_S, h_t)/8$
$S_0$	6	0	0	0	0
	7	0	0	0	0
	8	1.3610e-03	4.3100e-04	1.1900e-04	3.1000e-05
	10	3.4100e-03	8.9000e-04	2.3600e-04	6.1000e-05
	12	1.8570e-03	5.0800e-04	1.3900e-04	3.6000e-05
	14	3.2300e-04	1.1000e-04	3.5000e-05	1.0000e-05
	16	2.2500e-04	5.1000e-05	1.0000e-05	2.0000e-06

**Tabela 5.10:** Erros relativos entre os dois métodos implementados para diferentes valores de  $S_0$ ,  $e_{\text{rela}} = |V_{\text{FEM}}(S_0, 0) - V_{\text{FDM}}(S_0, 0)|$ , para valores  $V(S_0, 0)$  das tabelas 5.8 e 5.9

Os resultados obtidos demonstram a convergência dos métodos de diferenças finitas e elementos finitos para resolver o problema das Opções Americanas. À medida que a discretização no espaço e no tempo é refinada, os valores de  $V(S_0, 0)$  para diferentes  $S_0$  convergem consistentemente para os mesmos resultados, como evidenciado pelas tabelas 5.8 e 5.9. A tabela 5.10 mostra que os erros relativos entre os dois métodos diminuem significativamente com o refinamento da malha, indicando que ambos os métodos convergem para a mesma solução. Esta convergência validada pelos erros relativos confirma a precisão e robustez dos algoritmos implementados.

# **Capítulo 6**

## **Conclusão e trabalho futuro**

Esta tese apresentou um estudo abrangente sobre a aplicação de métodos numéricos para resolver problemas de obstáculos, em específico o problema da membrana e da valoração de opções no contexto da matemática financeira.

Começámos por explorar o problema clássico do obstáculo, utilizando a membrana elástica como exemplo. Estabelecemos os fundamentos matemáticos, abordando a existência e regularidade das soluções e fornecendo uma base teórica sólida para as abordagens numéricas subsequentes.

Nos métodos numéricos de discretização, descrevemos detalhadamente os métodos das diferenças finitas e dos elementos finitos, essenciais para converter problemas contínuos em formas discretas solucionáveis. Abordámos também as técnicas de integração numérica necessárias para a implementação eficaz desses métodos.

A aplicação prática focou-se em problemas específicos, como o modelo de Black-Scholes para opções europeias e americanas. Implementámos várias estratégias algorítmicas, incluindo o Método de Newton Semi-Suave (SSNM), para resolver o problema de complementaridade linear (LCP). A análise da consistência, estabilidade e convergência assegurou a fiabilidade e robustez dos métodos.

Os resultados demonstraram a eficácia dos métodos numéricos na resolução de problemas financeiros complexos, validando as abordagens desenvolvidas. Confirmámos que tanto os métodos das diferenças finitas quanto os dos elementos finitos são eficazes na resolução do modelo de Black-Scholes para a valoração de opções.

Para trabalhos futuros, sugere-se a exploração de métodos adaptativos de refinamento de malha para melhorar a precisão das soluções em regiões críticas do domínio. Seria interessante aplicar, no caso parabólico, o método dos elementos finitos também na variável do tempo. Já que no nosso caso utilizámos um sistema híbrido de elementos finitos e diferenças finitas, apesar da aparente simplicidade pode não ser o mais robusto. Além disso poderíamos aplicar os métodos de estabilização, tal como fizemos no caso elíptico.

Finalmente, validar estas abordagens com dados financeiros reais permitirá ajustar os modelos para aplicações práticas robustas e precisas.

Em síntese, esta tese, através do desenvolvimento de métodos numéricos, aproximou a solução numérica para problemas de obstáculos com sucesso. Estes resultados, incluindo aplicações aos problemas de obstáculos da membrana, fornecem uma base sólida para futuras pesquisas e aplicações práticas.

# Bibliografia

- [1] L. Evans, *Partial Differential Equations*. American Mathematical Society, 2010.
- [2] N. Kikuchi and J. Oden, *Contact Problems in Elasticity*. Society for Industrial and Applied Mathematics, 1988.
- [3] J. Rodrigues, *Obstacle Problems in Mathematical Physics*. Elsevier Science, 1987.
- [4] A. Gonçalves, “Resolução numérica de problemas de obstáculo com aplicações à matemática financeira,” Master’s thesis, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal, 2018.
- [5] A. Sorsimo, “Solution of the inequality constrained reynolds equation by the finite element method,” Master’s thesis, Department of Mathematics and Systems Analysis, Espoo, Finland, 2012.
- [6] K. Atkinson and W. Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*. Springer, 2009, vol. 39.
- [7] R. Glowinski and G. Vijayasundaram, *Lectures on Numerical Methods for Non-Linear Variational Problems*. Springer, 2008.
- [8] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*. Society for Industrial and Applied Mathematics (SIAM), 1980.
- [9] T. Gustafsson, R. Stenberg, and J. Videman, “On finite element formulations for the obstacle problem – mixed and stabilised methods,” *Computational Methods in Applied Mathematics*, vol. 17, 2017.
- [10] ——, “Mixed and stabilized finite element methods for the obstacle problem,” *SIAM Journal on Numerical Analysis*, vol. 55, no. 6, pp. 2718–2744, 2017.
- [11] J. Haslinger, I. Hlaváček, and J. Nečas, “Numerical methods for unilateral problems in solid mechanics,” in *Finite Element Methods (Part 2), Numerical Methods for Solids (Part 2)*, ser. Handbook of Numerical Analysis. Elsevier, 1996, vol. 4, pp. 313–485.

- [12] C. Oosterlee and L. Grzelak, *Mathematical Modeling And Computation In Finance: With Exercises And Python And Matlab Computer Codes*. World Scientific, 2019.
- [13] D. Lamberton and B. Lapeyre, *Introduction to Stochastic Calculus Applied to Finance, Second Edition*. Taylor & Francis, 2008.
- [14] P. Antunes, *A Short Introduction to Computational Methods in Finance, Lecture Notes*. Departamento de Matemática, Instituto Superior Técnico, 2023.
- [15] F. Black and M. Scholes, “The pricing of options and corporate liabilities,” in *The Journal of Political Economy*, Vol. 81, No. 3. The University of Chicago Press, 1973, pp. 637–654.
- [16] H. Brezis, “Problèmes unilatéraux,” *J. Math. Pures Appl.*, vol. 51, pp. 1–168, 1972.
- [17] A. Petrosyan and H. Shahgholian, “Parabolic obstacle problems applied to finance. a free-boundary-regularity approach,” *Contemp. Math.*, vol. 439, 2007.
- [18] C. Alves, *Análise Numérica de Equações Diferenciais Parciais (uma introdução)*, *Lecture Notes*. Departamento de Matemática, Instituto Superior Técnico, 2020.
- [19] P. Antunes, *A Short Introduction to Numerical Methods for Partial Differential Equations, Lecture Notes*. Departamento de Matemática, Instituto Superior Técnico, 2024.
- [20] R. Carrington, “Speed comparison of solution methods for the obstacle problem,” Master Thesis, McGill University, Montréal, Canada, 2017.
- [21] T. Gustafsson, “Finite element methods for contact problems,” Ph.D. dissertation, Aalto University, Department of Mathematics and Systems Analysis, Finland, 2018.
- [22] M. Hintermüller, K. Ito, and K. Kunisch, “The primal-dual active set strategy as a semismooth newton method,” *SIAM Journal on Optimization*, vol. 13, no. 3, pp. 865–888, 2002.
- [23] P. Forsyth and K. Vetzal, “Numerical methods for nonlinear pdes in finance,” in *Handbook of Computational Finance*. Springer, 2012, pp. 503–528.
- [24] T. Chihaluca, “Aproximação numérica de equações diferenciais parciais não lineares com aplicações em finanças,” Tese de Mestrado, Universidade da Beira Interior, 2021.

hapappApêndice

# Apêndice A

## Algoritmos

---

**Algoritmo A.1:** Iterative Solution Algorithm: Semi-smooth Newton Method (SSNM) for Linear Complementarity Problem (LCP)

---

Dado o problema de encontrar  $\mathbf{u} \in \mathbb{R}^N$ , tal que:

$$\min\{-\mathbf{A}\mathbf{u} - \mathbf{f}; \mathbf{u} - \mathbf{g}\} = \mathbf{0}$$

onde  $\mathbf{A} \in \mathbb{R}^{N \times N}$  e  $\mathbf{f}, \mathbf{g} \in \mathbb{R}^N$  são sabidos.

(i) Inicialize:  $k = 0$ ,  $\mathbf{u}^{(0)}$ , por exemplo  $\mathbf{u}^{(0)} = \mathbf{g}$ ;

(ii) Para  $k \geq 0$ , itera-se:

$$\mathbf{u}^{k+1} = \mathbf{u}^k - [\nabla_u G(\mathbf{u}^k)]^{-1} G(\mathbf{u}^k)$$

onde

$$G(\mathbf{u}^{(k)}) := \min\{-\mathbf{A}\mathbf{u}^{(k)} + \mathbf{b}; \mathbf{u}^{(k)} - \mathbf{g}\} = \min\{\mathbf{a}^{(k)}; \mathbf{b}^{(k)}\}$$

$$\nabla_u G(\mathbf{u}) = -\mathbf{D}_a^{(k)} \mathbf{A} + \mathbf{D}_b^{(k)}$$

$\mathbf{D}_a$  e  $\mathbf{D}_b$  serão matrizes diagonais, onde a sua diagonal irá assumir os valores de 0 ou 1, dependendo de que entrada o mínimo corresponde. Para uma entrada genérica  $(i, i)$  da diagonal:

$$D_{a_{ii}} = \begin{cases} 1 & , \min\{a_i, b_i\} = a_i \\ 0 & , \min\{a_i, b_i\} = b_i \end{cases} ; \quad D_{b_{ii}} = \begin{cases} 0 & , \min\{a_i, b_i\} = a_i \\ 1 & , \min\{a_i, b_i\} = b_i \end{cases}$$

(iii) Verificar condição de paragem: se  $\|\mathbf{u}^{(k+1)} - \mathbf{u}^{(k)}\| \leq \text{TOL}$  parar;

(iv) Caso contrário, incrementa-se  $k = k + 1$  e volta-se ao ponto (ii).

---

---

**Algoritmo A.2:** Iterative Solution Algorithm: Primal-dual active set for Mixed Method

---

(i) Inicialize:  $k = 0$ ,  $\lambda^{(0)} = \mathbf{0}$ , e resolva  $\mathbf{A}\mathbf{u}^{(0)} = \mathbf{f}$ ;

(ii) Para  $k \geq 0$ , calcule-se:  $\tilde{\lambda}^{(k)} = \lambda^{(k)} + c(\mathbf{g} - \mathbf{B}\mathbf{u}^{(k)})$ , e initialize-se o Active-Set e o Inactive-Set:

$$\mathcal{A}_k = \left\{ a : \tilde{\lambda}_a^{(k)} > 0 \right\} ; \quad \mathcal{I}_k = \left\{ i : \tilde{\lambda}_i^{(k)} \leq 0 \right\}$$

tal que  $\tilde{\lambda}^{(k)}(\mathbf{a}^k) > 0$ , e  $\mathbf{s}^{(k)}(\mathbf{i}^k) \leq 0$ ;

(iii) Depois resolva

$$\begin{cases} \mathbf{A}\mathbf{u}^{(k+1)} - \mathbf{B}^T \lambda^{(k+1)} = \mathbf{f} \\ \mathbf{B}\mathbf{u}^{(k+1)} - \mathbf{g} \geq \mathbf{0} \end{cases}, \text{ no conjunto } \mathcal{A}_k \quad (\text{A.1})$$

e resolva

$$\lambda^{(k+1)} \geq \mathbf{0}, \text{ no conjunto } \mathcal{I}_k \quad (\text{A.2})$$

Isto é, resolver:

$$\begin{bmatrix} \mathbf{A} & -\mathbf{B}[\mathbf{a}^k, :]^T \\ \mathbf{B}[\mathbf{a}^k, :] & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(k+1)} \\ \lambda^{(k+1)}(\mathbf{a}^k) \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g}(\mathbf{a}^k) \end{bmatrix}$$

e  $\lambda^{(k+1)}(\mathbf{i}^k) = \mathbf{0}$ .

(iv) Se a condição de paragem:  $\|\lambda^{(k+1)} - \lambda^{(k)}\| \leq \text{TOL}$  se verificar, parar o algoritmo;

(v) Caso contrário, incrementa-se  $k = k + 1$  e volta-se ao ponto (ii).

---

---

**Algoritmo A.3:** Iterative Solution Algorithm: Primal-dual active set for Stabilized FEM

---

- (i) Inicialize:  $k = 0$ , resolva  $\mathbf{A}\mathbf{u}^{(0)} = \mathbf{f}$  e em seguida calcule  $\boldsymbol{\lambda}^{(0)} = \max\{\mathbf{C}_\alpha^{-1}(\mathbf{g}_\alpha - \mathbf{B}_\alpha\mathbf{u}^{(0)}); \mathbf{0}\}$ ;
- (ii) Para  $k \geq 0$ , inicialize-se o Active-Set e o Inactive-Set:

$$\mathcal{A}_k = \left\{ a : \lambda_a^{(k)} > 0 \right\} ; \quad \mathcal{I}_k = \left\{ i : \lambda_i^{(k)} \leq 0 \right\}$$

tal que  $\boldsymbol{\lambda}^{(k)}(\mathbf{a}^k) > 0$ , e  $\boldsymbol{\lambda}^{(k)}(\mathbf{i}^k) \leq 0$ ;

- (iii) Para obter  $\mathbf{u}^{(k+1)}$  resolva:

$$\begin{bmatrix} \mathbf{A}_\alpha & -\mathbf{B}_\alpha[\mathbf{a}^k, :]^T \\ \mathbf{B}_\alpha[\mathbf{a}^k, :] & \mathbf{C}_\alpha[\mathbf{a}^k, \mathbf{a}^k] \end{bmatrix} \begin{bmatrix} \mathbf{u}^{k+1} \\ \tilde{\boldsymbol{\lambda}}(\mathbf{a}^k) \end{bmatrix} = \begin{bmatrix} \mathbf{f}_\alpha \\ \mathbf{g}_\alpha(\mathbf{a}^k) \end{bmatrix}$$

podendo ainda ser escrito como

$$(\mathbf{A}_\alpha + \mathbf{B}_\alpha[\mathbf{a}^k, :]^T \mathbf{C}_\alpha^{-1}[\mathbf{a}^k, \mathbf{a}^k] \mathbf{B}_\alpha[\mathbf{a}^k, :]) \mathbf{u}^{(k+1)} = \mathbf{f}_\alpha + \mathbf{B}_\alpha[\mathbf{a}^k, :]^T \mathbf{C}_\alpha^{-1}[\mathbf{a}^k, \mathbf{a}^k] \mathbf{g}_\alpha$$

- (iv) Para obter  $\boldsymbol{\lambda}^{(k+1)}$ :

$$\boldsymbol{\lambda}^{(k+1)} = \max\{\mathbf{C}_\alpha^{-1}(\mathbf{g}_\alpha - \mathbf{B}_\alpha\mathbf{u}^{(k+1)}); \mathbf{0}\}$$

- (v) Se a condição de paragem:  $\|\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^{(k)}\| \leq \text{TOL}$  se verificar, parar o algoritmo;
  - (vi) Caso contrário, incrementa-se  $k = k + 1$  e volta-se ao ponto (ii).
- 

---

**Algoritmo A.4:** Diferenças finitas no espaço e esquema- $\theta$  no tempo para Opções Europeias

---

- (i) Defina:  $T, r, \sigma, h_S, h_t, S_{max}$  e  $\theta$ ;
- (ii) Inicialize:  $n = N_T + 1$ , a condição de valor final  $\mathbf{V}^{(N_t+1)} = \mathbf{g}$ , e as condições de contorno  $V_0^n = h_0^n$  e  $V_{M_S+1}^n = h_{M_S+1}^n$  para  $n = 0, \dots, N_t + 1$ ;
- (iii) Para  $n = N_t, \dots, 1, 0$  calcular:

$$\mathbf{V}_r^n = [-\mathbf{A}]^{-1}(\mathbf{B}\mathbf{V}_r^{n+1} + \mathbf{c}^n)$$

para  $\mathbf{A}, \mathbf{B}$  e  $\mathbf{c}^n$  definidas tal como na secção 4.3.5.

---

---

**Algoritmo A.5:** Elementos finitos no espaço e esquema- $\theta$  no tempo para Opções Europeias

---

- (i) Defina:  $T, r, \sigma, h_S, h_t, S_{max}$  e  $\theta$ ;
- (ii) Initialize:  $n = N_T + 1$ , a condição de valor final  $\mathbf{U}^{(N_t+1)} = \mathbf{g}$ , e as condições de contorno  $U_0^n = h_0^n$  e  $U_{M_S+1}^n = h_{M_S+1}^n$  para  $n = 0, \dots, N_t + 1$ ;
- (iii) Para  $n = N_t, \dots, 1, 0$  calcular:

$$\mathbf{U}_r^n = [-\mathbf{E}_r]^{-1}(\mathbf{r}_E^n + \mathbf{H}_R \mathbf{U}^{n+1})$$

para  $\mathbf{E}_r$ ,  $\mathbf{H}_R$  e  $\mathbf{r}_E^n$  definidas tal como definidas na secção 4.3.7.

---

---

**Algoritmo A.6:** Diferenças finitas no espaço, esquema- $\theta$  no tempo, com SSNM para o problema de complementaridade, para Opções Americanas

---

- (i) Defina:  $T, r, \sigma, h_S, h_t, S_{max}$  e  $\theta$ .
- (ii) Initialize:  $n = N_T + 1$ , a condição de valor final  $\mathbf{V}^{(N_t+1)} = \mathbf{g}$ , e as condições de contorno  $V_0^n = h_0^n$  e  $V_{M_S+1}^n = h_{M_S+1}^n$  para  $n = 0, \dots, N_t + 1$ ;
- (iii) Para  $n = N_t, \dots, 1, 0$  calcular  $\mathbf{V}_r^n$ , através da seguinte condição de minímo:

$$\max\{\mathbf{A}^n \mathbf{V}_r^n + \mathbf{B}^n \mathbf{V}_r^{n+1} + \mathbf{c}^n; (\mathbf{g}_r - \mathbf{V}_r^n)\} = \mathbf{0} \Leftrightarrow \min\{-\mathbf{A}^n \mathbf{V}_r^n - \mathbf{B}^n \mathbf{V}_r^{n+1} - \mathbf{c}^n; (\mathbf{V}_r^n - \mathbf{g}_r)\} = \mathbf{0}$$

para  $\mathbf{A}$ ,  $\mathbf{B}$  e  $\mathbf{c}^n$  definidas tal como na secção 4.3.5.

- (iv) Aplicar o SSNM à igualdade anterior de modo a calcular iterativamente a solução para cada  $n$ :
  - (1) Initialize:  $k = 0$ ,  $\mathbf{V}_r^{n(0)}$ , por exemplo  $\mathbf{V}_r^{n(0)} = \mathbf{g}_r$ ;
  - (2) Para  $k \geq 0$ , itera-se:

$$\mathbf{V}_r^{n(k+1)} = \mathbf{V}_r^{n(k)} - [\nabla_u G(\mathbf{V}_r^{n(k)})]^{-1} G(\mathbf{V}_r^{n(k)})$$

onde

$$G(\mathbf{V}_r^{n(k)}) := \min\{-\mathbf{A}^n \mathbf{V}_r^n - \mathbf{B}^n \mathbf{V}_r^{n+1} - \mathbf{c}^n; (\mathbf{V}_r^n - \mathbf{g}_r)\} = \min\{\mathbf{a}^{(k)}; \mathbf{b}^{(k)}\}$$

$$\nabla_u G(\mathbf{u}) = -\mathbf{D}_a^{(k)} \mathbf{A}^n + \mathbf{D}_b^{(k)}$$

$\mathbf{D}_a$  e  $\mathbf{D}_b$  serão matrizes diagonais, onde a sua diagonal irá assumir os valores de 0 ou 1, dependendo de que entrada o mínimo corresponde. Para uma entrada genérica  $(i, i)$  da diagonal:

$$D_{a_{ii}} = \begin{cases} 1 & , \min\{a_i, b_i\} = a_i \\ 0 & , \min\{a_i, b_i\} = b_i \end{cases}; \quad D_{b_{ii}} = \begin{cases} 0 & , \min\{a_i, b_i\} = a_i \\ 1 & , \min\{a_i, b_i\} = b_i \end{cases}$$

- (3) Verificar condição de paragem: se  $\|\mathbf{V}_r^{n(k+1)} - \mathbf{V}_r^{n(k)}\| \leq \text{TOL}$  parar;
  - (4) Caso contrário, incrementa-se  $k = k + 1$  e volta-se ao ponto (ii).
-

---

**Algoritmo A.7:** Elementos finitos no espaço, esquema- $\theta$  no tempo, com SSNM para o problema de complementaridade, para Opções Americanas

---

- (i) Defina:  $T, r, \sigma, h_S, h_t, S_{max}$  e  $\theta$ .
- (ii) Inicialize:  $n = N_T + 1$ , a condição de valor final  $\mathbf{U}^{(N_t+1)} = \mathbf{g}$ , e as condições de contorno  $U_0^n = h_0^n$  e  $U_{M_S+1}^n = h_{M_S+1}^n$  para  $n = 0, \dots, N_t + 1$ ;
- (iii) Para  $n = N_t, \dots, 1, 0$  calcular  $\mathbf{U}_r^n$ , através da seguinte condição de mínimo:

$$\min\{-\mathbf{E}_r \mathbf{U}_r^n - \mathbf{r}_E^n - \mathbf{H}_R \mathbf{U}^{n+1}; (\mathbf{U}_r^n - \mathbf{g}_r)\} = \mathbf{0}$$

para  $\mathbf{E}_r$ ,  $\mathbf{r}_E^n$  e  $\mathbf{H}_r$  definidas tal como na secção 4.3.5.

- (iv) Aplicar o SSNM à igualdade anterior de modo a calcular iterativamente a solução para cada  $n$ :
- (1) Inicialize:  $k = 0$ ,  $\mathbf{U}_r^{n(0)}$ , por exemplo  $\mathbf{U}_r^{n(0)} = \mathbf{g}_r$ ;
- (2) Para  $k \geq 0$ , itera-se:

$$\mathbf{U}_r^{n(k+1)} = \mathbf{U}_r^{n(k)} - [\nabla G(\mathbf{U}_r^{n(k)})]^{-1} G(\mathbf{V}_r^{n(k)})$$

onde

$$G(\mathbf{U}_r^{n(k)}) := \min\{-\mathbf{E}_r \mathbf{U}_r^{n(k)} - \mathbf{r}_E^n - \mathbf{H}_R \mathbf{U}^{n+1}; \mathbf{U}_r^{n(k)} - \mathbf{g}_r\} = \min\{\mathbf{a}^{(k)}; \mathbf{b}^{(k)}\}$$

$$\nabla_u G(\mathbf{u}) = -\mathbf{D}_a^{(k)} \mathbf{E}_r + \mathbf{D}_b^{(k)}$$

$\mathbf{D}_a$  e  $\mathbf{D}_b$  serão matrizes diagonais, onde a sua diagonal irá assumir os valores de 0 ou 1, dependendo de que entrada o mínimo corresponde. Para uma entrada genérica  $(i, i)$  da diagonal:

$$D_{a_{ii}} = \begin{cases} 1 & , \min\{a_i, b_i\} = a_i \\ 0 & , \min\{a_i, b_i\} = b_i \end{cases} ; \quad D_{b_{ii}} = \begin{cases} 0 & , \min\{a_i, b_i\} = a_i \\ 1 & , \min\{a_i, b_i\} = b_i \end{cases}$$

- (3) Verificar condição de paragem: se  $\|\mathbf{U}_r^{n(k+1)} - \mathbf{U}_r^{n(k)}\| \leq \text{TOL}$  parar;
  - (4) Caso contrário, incrementa-se  $k = k + 1$  e volta-se ao ponto (2).
-