Introdução aos Métodos de Decomposição de Domínio

Publicações Matemáticas

Introdução aos Métodos de Decomposição de Domínio

Juan Galvis
Texas A&M University

impa

27º Colóquio Brasileiro de Matemática

Impresso no Brasil / Printed in Brazil

Capa: Noni Geiger / Sérgio R. Vaz

27º Colóquio Brasileiro de Matemática

- A Mathematical Introduction to Population Dynamics - Howard Weiss
- Algebraic Stacks and Moduli of Vector Bundles - Frank Neumann
- An Invitation to Web Geometry - Jorge Vitório Pereira e Luc Pirio
- Bolhas Especulativas em Equilíbrio Geral - Rodrigo Novinski e Mário Rui Páscoa
- C*-algebras and Dynamical Systems - Jean Renault
- Compressive Sensing - Adriana Schulz, Eduardo A. B. da Silva e Luiz Velho
- Differential Equations of Classical Geometry, a Qualitative Theory - Ronaldo Garcia e Jorge Sotomayor
- Dynamics of Partial Actions - Alexander Arbieto e Carlos Morales
- Introduction to Evolution Equations in Geometry - Bianca Santoro
- Introduction to Intersection Homology and Perverse Sheaves - Jean-Paul Brasselet
- Introdução à Análise Harmônica e Aplicações - Adán J. Corcho Fernandez e Marcos Petrúcio de A. Cavalcante
- **Introdução aos Métodos de Decomposição de Domínio - Juan Galvis**
- Problema de Cauchy para Operadores Diferenciais Parciais - Marcelo Rempel Ebert e José Ruidival dos Santos Filho
- Simulação de Fluidos sem Malha: Uma Introdução ao Método SPH - Afonso Paiva, Fabiano Petronetto, Geovan Tavares e Thomas Lewiner
- Teoria Ergódica para Autômatos Celulares Algébricos - Marcelo Sobottka
- Uma Iniciação aos Sistemas Dinâmicos Estocásticos - Paulo Ruffino
- Uma Introdução à Geometria de Contato e Aplicações à Dinâmica Hamiltoniana - Umberto L. Hryniewicz e Pedro A. S. Salomão
- Viscosity Solutions of Hamilton-Jacobi Equations - Diogo Gomes

ISBN: 978-85-244-0300-2

Distribuição: IMPA
Estrada Dona Castorina, 110
22460-320 Rio de Janeiro, RJ
E-mail: ddic@impa.br
http://www.impa.br

Prefácio

Decomposição de Domínio refere-se a um conjunto de técnicas do tipo *"divisão e conquista"* usadas para achar soluções numéricas de equações diferenciais parciais, principalmente equações *elípticas* e *parabólicas*. Estas técnicas são baseadas numa *decomposição* ou partição em *subdomínios* do domínio onde a equação diferencial é formulada. A ideia geral é usar a solução da equação nos subdomínios para obter a solução (ou aproximações da solução) no domínio original. Por exemplo, no estudo do fluxo de petróleo ou água em meios porosos, a solução direta da equação diferencial no domínio original requer a solução de um sistema linear gigantesco. Resolver diretamente este sistema linear é impraticável em muitos casos devido ao alto custo computacional (e/ou muito tempo de processamento). Neste caso, uma abordagem de decomposição de domínio requer somente resolver a equação diferencial nos subdomínios *várias vezes*, em vez de resolver o problema diretamente no domínio original. As soluções da equação nos subdomínios podem ser obtidas resolvendo sistemas lineares menores que requerem baixo custo computacional e pouco tempo de processamento. *Computação paralela* pode ser usada para reduzir ainda mais o tempo total de processamento.

O objetivo geral do minicurso "Introdução aos métodos de decomposição de domínio" é introduzir a teoria clássica dos métodos de decomposição de domínio. Em particular, as ideias básicas de decomposição de domínio aplicadas a solução de sistemas lineares que aparecem na aproximação numérica da solução de uma equação diferencial parcial elíptica.

Iniciamos com uma introdução a modelagem de fluxo de fluidos em *meios porosos heterogêneos*. Depois apresentamos o *método dos elementos finitos* para aproximar a solução de equações diferenciais elípticas. Em seguida estudamos o método do *gradiente conjugado precondicionado* para a solução de sistemas lineares. Na parte principal do curso apresentamos vários *precondicionadores* de decomposição de domínio. O objetivo desta parte será entender a construção e implementação destes precondicionadores. Apresentar-se-á uma introdução curta à análise teórica envolvida. O minicurso será finalizado ilustrando rapidamente outros precondicionadores de decomposição de domínio e mencionando vários outros modelos e aplicações onde as técnicas estudadas podem ser aplicadas.

Observamos aqui que a construção de precondicionadores é somente uma subárea de decomposição de domínios. Muitos outros problemas encontrados na análise numérica de equações diferenciais parciais podem ser abordados usando ideias de decomposição de domínios, veja [31, 24, 30, 27].

O curso está orientado para estudantes de iniciação científica ou primeiro ano de mestrado. É requerido familiaridade com os conceitos básicos de álgebra linear e noções básicas de equações diferenciais parciais em duas dimensões.

Gostaria de agradecer a Martha Miranda pela ajuda com o texto e ao Duilio Conceição por revisar o conteúdo e o texto do minicurso e também pelas suas contribuições assim como as discussões científicas.

Conteúdo

1 Introdução **1**

1.1 Problemas elípticos em uma e duas dimensões 1

1.2 Simulação de fluidos em meios porosos 3

2 Elementos finitos em 1D **8**

2.1 Introdução 8

2.2 Formulação fraca 9

2.2.1 Espaços de funções 10

2.2.2 Exemplo: a equação de Laplace 11

2.2.3 Exemplo: equação elíptica básica 13

2.2.4 Existência de soluções fracas 14

2.3 Formulação de Galerkin 15

2.3.1 O espaço de funções lineares por partes 18

2.3.2 O sistema linear obtido 26

2.4 Erro de aproximação 27

2.5 Experimentos numéricos 28

3 Elementos finitos em 2D **32**

3.1 Introdução 32

3.2 Espaços de funções 34

3.3 Formulação fraca 35

3.3.1 Exemplo: a equação de Laplace 36

3.3.2 Exemplo: equação elíptica básica 37

3.4 Formulação de Galerkin 38

3.4.1 O espaço de funções lineares por partes 39

3.4.2 O sistema linear obtido 51

3.4.3 Erro de aproximação 52

4 O método do gradiente conjugado 53
4.1 O método do gradiente conjugado 53
4.2 Contagem de número de iterações 63
4.3 Experimentos numéricos 64
4.4 O gradiente conjugado precondicionado 69

5 Métodos com superposição em 1D 73
5.1 Decomposição com e sem sobreposição 73
5.2 Espaços locais e operadores de restrição 75
5.3 Precondicionador aditivo de um nível 77
5.4 Experimentos numéricos 79
5.5 Precondicionador de dois níveis 80
5.5.1 Espaços grossos 83
5.5.2 Número de condição 84
5.6 Experimentos numéricos 85
5.7 Introdução à análise 87

6 Métodos com superposição em 2D 99
6.1 Decomposição com e sem sobreposição 99
6.2 Espaços locais e operadores de restrição 102
6.3 Precondicionador aditivo de um nível 103
6.4 Experimentos numéricos 104
6.5 Precondicionador de dois níveis 107
6.5.1 Espaços grossos 108
6.6 Experimentos numéricos 112
6.7 Introdução à análise 113

7 Comentários finais 115
7.1 Introdução aos métodos sem sobreposição 115
7.1.1 O complemento de Schur 117
7.1.2 Precondicionadores 118
7.2 Outras equações diferenciais parciais 119
7.3 Bibliografia recomendada 121

Bibliografia 122

Capítulo 1

Introdução

1.1 Problemas elípticos em uma e duas dimensões

Neste capítulo introduzimos as equações diferenciais usadas neste minicurso. Seja $D \subset \mathbb{R}^2$ um domínio aberto e limitado e ∂D seu bordo. Consideramos equações diferenciais da forma

$$\begin{array}{l} \textit{Achar } u : D \subset \mathbb{R}^2 \to \mathbb{R} \textit{ tal que:} \\ \left\{ \begin{array}{rcll} -\mathrm{div}(\kappa(x)\nabla u(x)) = & f(x) & \text{para todo } x = (x_1, x_2) \in D, \\ u(x) = & g(x) & \text{para todo } x = (x_1, x_2) \in \partial D \end{array} \right. \end{array} \quad (1.1)$$

onde ∇u denota o vetor *linha* gradiente de u dado por $\nabla u = \left[\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2} \right]$ e o tensor de permeabilidade κ é da forma

$$\kappa(x) = \begin{bmatrix} \kappa_{11}(x) & \kappa_{12}(x) \\ \kappa_{21}(x) & \kappa_{22}(x) \end{bmatrix}.$$

O operador diferencial envolvido é

$$\mathrm{div}(\kappa(x)\nabla u(x)) = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{\partial}{\partial x_1} \left(\kappa_{ij}(x) \frac{\partial}{\partial x_j} u(x) \right).$$

Dizemos que a equação diferencial (1.1), ou que o operador $\mathrm{div}(\kappa\nabla(\cdot))$, é elíptico se existem $\kappa_{\min}$ e $\kappa_{\max}$ tais que para todo $x \in D$ temos

$$0 < \kappa_{\min} \leq \mu_{\min}(x) \leq \mu_{\max}(x) \leq \kappa_{\max}, \tag{1.2}$$

onde, para cada $x \in D$, $\mu_{\min}(x)$ e $\mu_{\max}(x)$ são o menor e maior autovalor da matriz $\kappa(x)$.

Note que a versão em uma dimensão da equação diferencial (1.1) é dada por

$$\begin{array}{l} \textit{Achar } u : (a,b) \to \mathbb{R} \textit{ tal que:} \\ \left\{ \begin{array}{rll} -(\kappa(x)u'(x))' = & f(x) & \text{para todo } x \in (a,b), \\ u(x) = & g(x) & \text{para } x = a \text{ e } x = b, \end{array} \right. \end{array} \tag{1.3}$$

que é elíptica quando existem $\kappa_{\max}$ e $\kappa_{\min}$ tais que

$$0 < \kappa_{\min} \leq \kappa(x) \leq \kappa_{\max} \qquad \text{para todo } x \in (a,b). \tag{1.4}$$

O objetivo deste minicurso é o uso de técnicas de decomposição de domínios para obter de forma eficiente a solução numérica de (1.3) e (1.1). Para poder apresentar as técnicas de decomposição de domínio precisamos primeiro introduzir uma *discretização* para aproximar as soluções das equações (1.3) e (1.1). Podemos eleger dentro muitas discretizações possíveis, veja por exemplo [4]. Neste minicurso usamos o método dos elementos finitos para obter uma aproximação numérica destas equações elípticas. Usando elementos finitos a solução direta da equação diferencial no domínio original requer resolver um sistema linear *gigantesco*, *mal condicionado* e *esparso*. Resolver diretamente este sistema linear é impraticável em muitos casos devido ao alto custo computacional e/ou muito tempo de processamento. Uma abordagem de decomposição de domínio introduz uma decomposição do domínio original em subdomínios e requer resolver a equação diferencial somente nos subdomínios *várias vezes*, em vez de resolver o problema diretamente no domínio original. As soluções da equação nos subdomínios podem ser obtidas resolvendo sistemas lineares menores que requerem baixo custo computacional e pouco tempo de processamento. *Computação paralela* pode ser usada para reduzir ainda mais o tempo total de processamento.

Equações diferenciais elípticas mais gerais podem também ser estudadas, sendo a escolha de (1.3) e (1.1) motivada unicamente pela simplicidade da notação e exposição de ideias. Por exemplo pode-se considerar a equação

$$\begin{array}{l} \textit{Achar } u : D \subset \mathbb{R}^2 \to \mathbb{R} \textit{ tal que:} \\ \left\{ \begin{array}{rrl} -\mathrm{div}(\kappa(x)\nabla u(x)) + \boldsymbol{B}(x)\nabla u + c(x)u = & f(x) & \forall x \in D \\ u(x) = & g(x) & \forall x \in \partial D \end{array} \right. \end{array}$$

com hipóteses adequadas para $\boldsymbol{B}$ e c. Também pode-se usar outras condições de contorno. Veja [31, 24, 27, 30].

1.2 Simulação de fluidos em meios porosos

O tratamento numérico de (1.3) e (1.1) depende do tipo de coeficiente $\kappa(x)$ usado. Definimos o contraste do coeficiente por

$$\mathrm{contraste}(\kappa) = \kappa_{\max}/\kappa_{\min}.$$

Em geral, entre maior o contraste, mais difícil é obter computacionalmente uma aproximação numérica da equação elíptica (1.3) ou (1.1). Veja os experimentos numéricos na Seções 2.5 e 4.3.

Quando as equações elípticas acima modelam fenômenos de escoamento de fluidos em meios porosos, o coeficiente κ representa, de algum jeito, a permeabilidade do meio poroso. A permeabilidade do meio poroso mede a facilidade do meio de deixar o fluido escoar e depende de muitos fatores. Sendo difícil descrever a permeabilidade do meio poroso, mencionamos três classes de coeficiente que podem aparecer dependendo da informação sobre o meio poroso e de como é processada esta informação:

1) Coeficiente constante por partes

Em algumas situações o meio poroso é particionado em blocos de tamanho apropriado e tem-se informação da permeabilidade "média" em cada um destes blocos. Temos então que nas equações (1.1) e (1.3) o coeficiente κ é uma função constante por partes. Note que neste caso as segundas derivadas em (1.1) podem não existir no sentido clássico. Uma interpretação adequada desta equação é requerida.

2) Coeficiente oscilatório

Outro jeito de modelar a permeabilidade de um meio poroso é usando coeficientes oscilatórios, isto é, coeficientes baseados em funções periódicas com diferentes períodos. A idéia é que quando misturamos variações em diferentes períodos, obtemos uma representação aceitável da física do meio poroso. Neste caso o coeficiente nas equações acima envolve funções periódicas com períodos variando em diferentes escalas. O uso de coeficientes oscilatórios reflete as múltiplas escalas presentes no meio poroso: tamanho do poro, escala geológica e tamanho do reservatório entre outras. Veja [1, 5, 13].

3) Coeficiente aleatório

Pode-se também assumir que a permeabilidade é impossível de descrever exatamente. Neste caso usamos uma permeabilidade aleatória, isto é, o coeficiente da equação elíptica acima é uma variável aleatória para cada $x \in D$. Este tipo de modelo é adequado em muitos casos. Quando o coeficiente $\kappa(x)$ de (1.1) é aleatório pode-se usar métodos tipo Monte Carlo para calcular uma aproximação numérica. Os métodos tipo Monte Carlo requerem a solução numérica de uma equação elíptica várias vezes.

Pode-se também usar coeficientes que combinem estes três tipos acima, por exemplo um coeficiente calculado como a soma de uma função constante por partes e uma função oscilatória. Em todos os casos, quando uma aproximação numérica de uma equação diferencial elíptica é requerida, temos que resolver um sistema linear gigantesco e esparso. A matriz obtida é uma matriz mal condicionada. Entre maior é o contraste do meio, pior a condição da matriz. Entre maior é a precisão requerida, maior a dimensão do sistema linear e pior a condição da matriz.

Levando em conta os comentários acima, apresentamos agora uma situação prática. Consideramos o modelo de escoamento bifásico (água e óleo) em meios porosos. Neste modelo em particular, a quantidade que queremos aproximar depende do tempo e para poder obter uma simulação com sucesso do tempo inicial até o tempo final é

necessário resolver a equação elíptica (1.1) muitas vezes. Para poder aplicar este modelo em problemas reais no reservatório de petróleo, precisamos, pelo menos, uma forma *eficiente* de obter soluções *acuradas* das equações elípticas do tipo (1.1) ou (1.3).

A modelagem do fluxo de fluidos em meios porosos aparece em uma ampla gama de aplicações, incluindo áreas como a Engenharia de Petróleo, Ciências Ambientais, Biologia, Hidrologia e Geologia, entre outras. O objetivo ate o final do capítulo é apresentar ideias gerais da modelagem de fluidos em meios porosos e identificar a necessidade do uso de métodos adequados na solução dos sistemas lineares que aparecem como parte deste processo.

A modelagem do fluxo de fluidos em meios porosos envolve muitas questões práticas e teóricas importantes e é o epicentro de muitos projetos de pesquisa no mundo acadêmico/científico. As ferramentas matemáticas usadas tornam-se então fundamentais para lidar eficientemente com as dificuldades inerentes à modelagem, como por exemplo o tratamento das heterogeneidades do meio poroso e outras questões relacionadas com a permeabilidade. Em particular, os métodos numéricos para fazer frente a estas dificuldades tem que ser eficientes e aproveitar o incremento constante do poder computacional dos computadores modernos.

Para fixar ideias consideramos a modelagem numérica de fluidos (possivelmente multifásico) num reservatório de petróleo. Na modelagem em meios porosos uma das equações diferenciais mais usadas é a equação de Darcy, ou lei de Darcy. Para um fluido em um meio poroso modelado pelo domínio $D \subset \mathbb{R}^2$, esta lei pode ser escrita como

$$\boldsymbol{u}(x) = -\frac{\kappa(x)}{\mu}\nabla p(x) + \boldsymbol{F}(x), \tag{1.5}$$

onde $\boldsymbol{u} = (u_1(x), u_2(x)) : D \to \mathbb{R}^2$ é a velocidade do escoamento, a função $p : D \to \mathbb{R}$ é a pressão do fluido, o parâmetro μ é a viscosidade do fluido e a função $\boldsymbol{F} : D \to \mathbb{R}^2$ representa um força externa, gravidade por exemplo. O coeficiente κ é a permeabilidade do meio poroso. Este coeficiente representa a capacidade do meio poroso de deixar o fluido escoar. Em geral κ é um objeto complicado e existem

várias opções para modelar este coeficiente e para trata-lo numericamente.

Na modelagem de fluidos em meios porosos a equação (1.5), ou equações derivadas dela, precisam ser resolvidas numericamente. Por exemplo, o modelo matemático para o transporte de fluido bifásico (água e petróleo) no processo de recuperação de petróleo, temos que resolver numericamente um sistema de equações que tem uma equação de *transporte* para a saturação (quantidade relativa de água ou óleo) acoplada com a equação de Darcy para a velocidade junto com uma condição de incompressibilidade do fluido; veja [1]. Se assumimos que não existem fontes nem sumidouros e sem considerar a gravidade, o sistema é,

$$\begin{aligned} \mathbf{u} &= -\lambda(s)\kappa\nabla p \\ -\mathrm{div}\cdot\mathbf{u} &= 0 \\ \frac{\partial s}{\partial t} + \nabla\cdot(F(s)\mathbf{u}) &= 0. \end{aligned} \tag{1.6}$$

Aqui s é a saturação de água. As funções $\lambda(s)$ e $F(s)$ representam a *mobilidade total* e o *fluxo fracionário*, respectivamente. Para aproximar este sistema numericamente temos que introduzir uma discretização no tempo (t) e uma discretização na variável espacial ($x \in D$). Para a maioria das escolhas para a discretização temporal, na aproximação numérica de (1.6), ainda temos que lidar, em cada passo de tempo, com uma equação da forma

$$-\mathrm{div.}\,(\tilde{\kappa}\nabla p) = f \text{ in } D, \tag{1.7}$$

onde $\tilde{\kappa} = \lambda(s)\kappa$ e o lado direito f depende da aproximação no passo do tempo anterior. Esta equação resulta de substituir a primeira equação em (1.6) na segunda. A aproximação de (1.7) pode ser feita utilizando várias *discretizações*; veja [4, 7, 8, 21]. Podemos utilizar por exemplo, uma discretização de elementos finitos, a qual requer o cálculo da solução de um sistema linear muito grande, esparso, e muito mal condicionado. A solução desta classe de sistemas lineares requer muito tempo de CPU e muitos recursos de memória do computador. Em geral, calcular a solução numérica de (1.7) é o gargalo computacional do processo de aproximação do sistema (1.6).

Para poder aplicar este modelo em problemas reais no reservatório de petróleo, precisamos, então, de um meio de obter soluções acuradas e eficientes das equações elípticas do tipo (1.1) ou (1.3).

Outros exemplos como o caso de fluido trifásico (água, petróleo e gás) e outros problemas associados a modelagem em meios porosos compartilham características similares. Veja por exemplo [1, 33, 5, 11]. Por último mencionamos que no lugar da equação elíptica (1.1), equações como (1.5) podem requerer uma aproximação. Neste caso é também possível usar o método dos elementos finitos para obter uma aproximação e as técnicas de decomposição de domínio calcular eficientemente esta aproximação numérica. Veja [31, 27, 24, 30].

Capítulo 2

O método dos elementos finitos em uma dimensão

Neste capítulo apresentamos uma introdução curta ao método dos elementos finitos em uma dimensão. O leitor interessado pode consultar [21, 23] e as referências ali citadas.

2.1 Introdução

Na hora de resolver numericamente equações diferenciais parciais o *método dos elementos finitos* é um dos mais usados na prática, especialmente para aproximar equações elípticas, parabólicas e sistemas de equações tipo Navier-Stokes. Em geral, o método dos elementos finitos para obter a aproximação numérica de uma equação diferencial requer basicamente três passos:

1. O primeiro passo consiste em construir uma *formulação fraca* ou *formulação variacional* da equação diferencial, isto é, o problema deve ser posto num espaço de funções adequado, usualmente um espaço de *Hilbert*. Verifica-se aqui a existência de *soluções fracas* para o problema estudado.

2. O segundo passo é introduzir espaços de elementos finitos, o problema obtido no primeiro passo é aproximado por um problema posto num espaço de dimensão finita. Depois de introduzir bases adequadas, achar a aproximação de elementos finitos é equivalente a resolver um sistema linear. Na prática, a matriz deste sistema linear é muito grande, esparsa e mal condicionada.

3. O terceiro passo é resolver o sistema linear obtido no passo anterior. Para resolver este sistema linear, usa-se algum método iterativo. Quando o sistema linear é definido positivo, o método preferido é o método de gradiente conjugado precondicionado.

Vamos estudar agora os dois primeiros passos acima mencionados do método dos elementos finitos para o caso de uma dimensão. No Capítulo 3 apresentamos uma introdução ao caso de duas dimensões. Depois, no Capítulo 4 estudamos o método do gradiente conjugado precondicionado para resolver o sistema linear obtido (terceiro passo). Nos Capítulos 5, 6 e 7 apresentamos os precondicionadores de decomposição de domínios usados no método do gradiente conjugado precondicionado.

2.2 Formulação fraca

Para deduzir a formulação fraca de uma equação diferencial (e.g. (1.3)) no intervalo (a, b) temos que

1. Supor que existe uma solução u de nossa equação diferencial, multiplicar os dois lados da equação por uma *função teste* $v \in C_0^\infty(a, b)$ e tomar integrais nos dois lados da igualdade.

2. Usar a fórmula de integração por partes e a condição de contorno para ficar com expressões que nessecitem somente de derivadas da mais baixa ordem possível. Por exemplo, geralmente, uma expressão requerendo calcular somente primeiras derivadas é preferida a uma que envolve o cálculo de alguma segunda derivada.

3. Trocar $v \in C_0^\infty(a,b)$ por $v \in V$ onde o espaço de funções teste V é o maior possível. Neste passo também escolhemos o espaço de funções U tal que a solução $u \in U$.

Usamos os exemplos introduzidos no Capítulo 1 para ilustrar o procedimento acima.

2.2.1 Espaços de funções

Agora vamos a definir o espaço $H^1(a,b)$ que será o espaço de funções apropriado para todos os problemas elípticos considerados neste minicurso; veja [23, 9]. Definimos o espaço $L^2(a,b)$ como o conjunto das funções de quadrado integrável. Também definimos

$$H^1(a,b) := \left\{ v \in L^2(a,b) \mid \quad v' \in L^2(a,b) \right\}. \tag{2.1}$$

O espaço $H^1(a,b)$ é um espaço de Hilbert com o produto interno

$$\begin{aligned} (u,v)_{H^1(a,b)} &= \int_a^b u(x)v(x)dx + \int_a^b u'(x)v'(x)dx \\ &= (u,v)_{L^2(a,b)} + (u',v')_{L^2(a,b)} \end{aligned}$$

e norma

$$\begin{aligned} \|v\|^2_{H^1(a,b)} &= \int_a^b |v(x)|^2 dx + \int_a^b |v'(x)|^2 dx \\ &= \|v\|^2_{L^2(a,b)} + \|v'\|^2_{L^2(a,b)}. \end{aligned}$$

Também usaremos o espaço de funções $H_0^1(a,b) \subset H^1(a,b)$ definido por

$$H_0^1(a,b) = \{v \in H^1(a,b) \mid \ v(a) = 0 \text{ e } v(b) = 0\}. \tag{2.2}$$

Finalmente definimos o espaço $H^2(a,b)$ por

$$H^2(a,b) := \left\{ v \in L^2(a,b) \mid \quad v', v'' \in L^2(a,b) \right\}$$

com a norma

$$\begin{aligned} \|v\|^2_{H^2(a,b)} &= \int_a^b |v(x)|^2 dx + \int_a^b |v'(x)|^2 dx + \int_a^b |v''(x)|^2 dx \\ &= \|v\|^2_{L^2(a,b)} + \|v'\|^2_{L^2(a,b)} + \|v''\|^2_{L^2(a,b)}. \end{aligned}$$

Note que $H^2(a,b) \subset H^1(a,b) \subset L^2(a,b)$.

2.2.2 Exemplo: a equação de Laplace

Considere a seguinte equação diferencial parcial elíptica de Laplace,

$$\begin{array}{l} \textit{Achar } u : (a,b) \to \mathbb{R} \textit{ tal que:} \\ \left\{ \begin{array}{rll} -u''(x) = & f(x), & \text{para } x \in (a,b) \\ u(x) = & g(x), & \text{para } x = a, x = b. \end{array} \right. \end{array} \tag{2.3}$$

Para obter a formulação fraca de (2.3), primeiro multiplicamos os dois lados da primeira igualdade em (2.3) por $v \in C_0^\infty(a,b)$ fixa mas arbitrária, depois integramos os dois lados da equação e obtemos

$$\begin{array}{l} \textit{Achar } u : (a,b) \to \mathbb{R} \textit{ tal que:} \\ \left\{ \begin{array}{rll} -\displaystyle\int_a^b u''(x)v(x)dx = & \displaystyle\int_a^b f(x)v(x)dx, & \forall v \in C_0^\infty(a,b) \\ u(x) = & g(x), & \text{para } x = a, x = b, \end{array} \right. \end{array} \tag{2.4}$$

em seguida, usando a fórmula de integração por partes e o fato $v(a) = v(b) = 0$ para toda $v \in C_0^\infty(a,b)$ obtemos

$$\begin{aligned} -\int_a^b u''(x)v(x)dx &= \int_a^b u'(x)v'(x)dx - [u'(b)v(b) - u'(a)v(a)] \\ &= \int_a^b u'(x)v'(x)dx - [u'(b)0 - u'(a)0] \\ &= \int_a^b u'(x)v'(x)dx. \end{aligned}$$

A equação (2.4) pode ser escrita então como

$$\begin{array}{l} \textit{Achar } u : [0,1] \to \mathbb{R} \textit{ tal que:} \\ \left\{ \begin{array}{rll} \displaystyle\int_a^b u'(x)v'(x)dx = & \displaystyle\int_a^b f(x)v(x)dx, & \forall v \in C_0^\infty(a,b) \\ u(x) = & g(x), & \text{para } x = a, x = b. \end{array} \right. \end{array} \tag{2.5}$$

A formulação fraca de (2.3) é quase (2.5), pois ainda temos que trocar os espaços de funções envolvidos. Isto é necessário pois queremos que o problema seja posto num espaço de *Hilbert* e $C_0^\infty(a,b)$ não é um espaço de Hilbert. Temos que escolher um espaço U onde procurar a solução u, i.e., $u \in U$, e também temos que escolher um espaço

V para as funções teste v, i.e., $v \in V$. Queremos $C_0^\infty(a,b) \subset V$. Em termos gerais, os espaços U e V devem ser tais que:

- Todas as integrais na possível formulação fraca obtida devem estar bem definidas. No nosso exemplo da equação de Laplace estamos considerando as integrais na primeira equação em (2.5), isto é, as derivadas u', v' e a integral $\int_a^b u'(x)v'(x)dx$ devem estar bem definidas para toda $u \in U$ e $v \in V$. Note que $\int_a^b u'(x)v'(x)dx$ está bem definida se u' e v' são funções de quadrado integrável, i.e., funções em $L^2(a,b)$.
  Também precisamos que a integral $\int_a^b f(x)v(x)dx$ esteja bem definida para toda $v \in V$. Note que se $v \in L^2(a,b)$ e $f \in L^2(a,b)$ temos que esta integral está bem definida.

- Para toda $u \in U$ tem que fazer sentido a segunda igualdade em (2.5). Lembramos que no caso geral esta igualdade corresponde a imposição da condição de contorno do problema estudado. No exemplo da equação de Laplace com condição de contorno de Dirichlet esta equação envolve os valores da função u nos pontos fronteira do intervalo (a,b), i.e., em a e b. Por exemplo, a escolha $U = L^2(a,b)$ não é boa neste sentido pois se $u \in L^2(a,b)$ os valores $u(a)$ e $u(b)$ podem não estar bem definidos.

- A escolha de U e V pode ser feita independente uma da outra. Pode-se também escolher $U = V$. Em todas as formulações fracas deste livro usamos $U = V$, i.e., o espaço onde procuramos a solução é o mesmo espaço de funções teste. Esta escolha tem a vantagem de ser mais fácil obter sistemas lineares *simétricos*.

- Os espaços de funções mais adequados considerando os três items acima são os *Espaços de funções tipo Sobolev* definidos em (a,b). Uma definição geral destes espaços requer o conhecimento de um pouco de *medida e integração* e de *derivadas generalizadas*. Na Seção 2.2.1 fizemos uma revisão rápida destes espaços. Veja [22] por exemplo.

Escolhemos $V = H_0^1(a,b)$ e $U = H^1(a,b)$. Para facilitar a escrita introduzimos a seguinte notação

$$\mathcal{A}(u,v) = \int_a^b u'(x)v'(x)dx \quad \text{para toda } v \in V \text{ e } u \in U \tag{2.6}$$

e

$$\mathcal{F}(v) \quad = \quad \int_a^b f(x)v(x)dx \quad \text{para toda } v \in V. \tag{2.7}$$

Note que $\mathcal{A} : U \times V \to \mathbb{R}$ é uma forma bilinear e $\mathcal{F} : V \to \mathbb{R}$ é um funcional linear. A formulação fraca de (2.3) é então a formulação (2.5) posta nos espaços U e V definidos acima:

$$\begin{array}{l} \textit{Achar } u \in U \textit{ tal que:} \\ \left\{ \begin{array}{rl} \mathcal{A}(u,v) = & \mathcal{F}(v), \quad \forall v \in V \\ u(x) = & g(x) \qquad \text{para } x = a, x = b. \end{array} \right. \end{array} \tag{2.8}$$

Observação 1. *A formulação fraca (2.8) pode ser reduzida ao caso $g(x) = 0$ em $x = a$ e $x = b$. Com efeito, se u_g é uma função suave qualquer tal que $u_g(a) = g(a)$ e $u_g(b) = g(b)$ podemos escrever $u = u^* + u_g$, note que $u^* \in V = H_0^1(a,b)$ e para achar u^* podemos resolver o problema*

$$\begin{array}{l} \textit{Achar } u^* \in V \textit{ tal que:} \\ \left\{ \begin{array}{rl} \mathcal{A}(u^*,v) = & \mathcal{F}(v) - \mathcal{A}(u_g,v) \quad \forall v \in V. \end{array} \right. \end{array} \tag{2.9}$$

2.2.3 Exemplo: equação elíptica básica em meios heterogêneos

Considere a seguinte equação diferencial parcial elíptica,

$$\begin{array}{l} \textit{Achar } u : (a,b) \to \mathbb{R} \textit{ tal que:} \\ \left\{ \begin{array}{rl} -(\kappa(x)u'(x))' = & f(x) \quad \text{para } x \in (a,b) \\ u(x) = & g(x) \quad \text{para } x = a, x = b, \end{array} \right. \end{array} \tag{2.10}$$

onde supomos que existem $\kappa_{\min}$ e $\kappa_{\max}$ tais que

$$0 < \kappa_{\min} \leq \kappa(x) \leq \kappa_{\max} \qquad \text{para todo } x \in (a,b). \tag{2.11}$$

Para obter a formulação fraca de (2.10), primeiro multiplicamos os dois lados da primeira igualdade por $v \in C_0^\infty(a,b)$ fixa mas arbitrária, depois integramos os dois lados da equação, aplicamos integração por

partes e obtemos

$$\begin{aligned} -\int_a^b (\kappa(x)u'(x))'v(x)dx &= \int_a^b \kappa(x)u'(x)v'(x)dx \\ &\quad - [\kappa(b)u'(b)v(b) - \kappa(a)u'(a)v(a)] \\ &= \int_a^b \kappa(x)u'(x)v'(x)dx - 0 \\ &= \int_a^b \kappa(x)u'(x)v'(x)dx. \end{aligned}$$

A formulação fraca de (2.10) pode ser escrita como

$$\begin{array}{l} \textit{Achar } u \in U \textit{ tal que:} \\ \left\{ \begin{array}{rcll} \mathcal{A}(u,v) = & \mathcal{F}(v) & \forall v \in V \\ u(x) = & g(x) & \text{para } x = a, x = b, \end{array} \right. \end{array} \tag{2.12}$$

onde a forma bilinear $\mathcal{A}$ é definida por

$$\mathcal{A}(u,v) = \int_a^b \kappa(x)u'(x)v'(x)dx \quad \text{para toda } v \in V \text{ e } u \in U \tag{2.13}$$

e o funcional linear F é como antes

$$\mathcal{F}(v) \quad = \quad \int_a^b f(x)v(x)dx \quad \text{para toda } v \in V. \tag{2.14}$$

Usando o mesmo argumento na Observação 1, a formulação (2.12) pode ser reduzida a

$$\begin{array}{l} \textit{Achar } u^* \in V \textit{ tal que:} \\ \left\{ \begin{array}{rcll} \mathcal{A}(u^*,v) = & \mathcal{F}(v) - \mathcal{A}(u_g,v) & \forall v \in V \end{array} \right. \end{array} \tag{2.15}$$

2.2.4 Existência de soluções fracas

A solução de uma formulação fraca é conhecida como solução fraca da equação original. Por exemplo, a solução de (2.12) é uma solução fraca da equação diferencial (2.10). Uma solução de (2.10) é dita solução forte, se as duas igualdades em (2.10) são satisfeitas para todo $x \in D$. Neste caso temos que poder calcular as duas derivadas na equação.

Observação 2 (Veja [21])**.** *Toda solução forte de uma equação diferencial é solução fraca. Se os coeficientes da equação são* regulares *e uma solução fraca é* regular*, então ela é solução forte. Aqui* regular *significa que a solução fraca tem derivadas contínuas até a ordem da equação diferencial. No caso de (2.10) precisamos de duas derivadas contínuas.*

Observação 3. *O método dos elementos finitos é usado para aproximar soluções fracas de equações diferenciais parciais.*

Para provar a existência de soluções fracas usa-se resultados como o Lema de Lax-Milgram. Nestas notas assumiremos que para o tipo de coeficiente em (2.11), temos existência de soluções fracas. Veja [21].

Lema 4. *Se κ satisfaz (2.11), então existe uma única solução fraca para o problema (2.12).*

2.3 Formulação de Galerkin

Todas as formulações fracas da Seção 2.2 são da forma,

$$\begin{array}{l} \textit{Achar } u^* \in V \textit{ tal que:} \\ \left\{ \begin{array}{l} \mathcal{A}(u^*, v) = \quad \mathcal{F}(v) \quad \text{para toda } v \in V \end{array} \right. \end{array} \tag{2.16}$$

onde V é o espaço de dimensão infinita $V = H_0^1(a, b)$. O segundo passo na aplicação do método dos elementos finitos consiste em trocar os espaços de dimensão infinita na formulação fraca (2.16) por espaços de dimensão finita, i.e., espaços de elementos finitos.

Considere V^h, $h > 0$, subespaços de dimensão finita $V^h \subset V$. O parâmetro h indica o tamanho do espaço, quanto o h, maior a dimensão de V^h. A idéia geral é que V^h aproxime o espaço V no limite quando $h \to 0$. Dado $V^h \subset V$ de dimensão finita, a formulação de Galerkin de (2.16) em V^h é

$$\begin{array}{l} \textit{Achar } u_h^* \in V^h \textit{ tal que:} \\ \left\{ \begin{array}{l} \mathcal{A}(u_h^*, v_h) = \quad \mathcal{F}(v_h) \quad \text{para toda } v_h \in V^h. \end{array} \right. \end{array} \tag{2.17}$$

O espaço V^h é de dimensão finita, podemos então considerar uma base de V^h denotada por

$$\{\phi_1, \phi_2, \ldots, \phi_{N_h}\} \quad \text{onde } N_h \text{ é a dimensão de } V^h. \tag{2.18}$$

Podemos escrever u_h^* como combinação linear desta base,

$$u_h^* = x_1\phi_1 + x_2\phi + \cdots + x_{N_h}\phi_{N_h} = \sum_{i=1}^{N_h} x_i\phi_i.$$

Como $\mathcal{A}$ é uma forma bilinear temos que para toda $v_h \in V^h$ vale

$$\mathcal{A}(u_h^*, v_h) = \mathcal{A}\left(\sum_{i=1}^{N_h} x_i\phi_i, v_h\right) = \sum_{i=1}^{N_h} x_i\mathcal{A}(\phi_i, v_h).$$

Sendo $\mathcal{A}$ uma forma bilinear e $\mathcal{F}$ um funcional linear, é suficiente verificar (2.17) somente para as funções teste $v_h = \phi_i$, $i = 1, \ldots, N_h$, na base (2.18) de V^h, i.e, a formulação de Galerkin (2.17) é equivalente a formulação

$$\begin{aligned} &\textit{Achar } u_h^* = \sum_{i=1}^{N_h} x_i\phi_i \textit{ tal que:} \\ &\begin{cases} \displaystyle\sum_{i=1}^{N_h} x_i\mathcal{A}(\phi_i, \phi_j) = \quad \mathcal{F}(\phi_j), \quad j = 1, \ldots, N_h. \end{cases} \end{aligned} \tag{2.19}$$

Introduzimos a representação matricial da forma bilinear $\mathcal{A}$, isto é, a matriz $A = [a_{ij}]$ de dimensão $N_h \times N_h$ com entradas

$$a_{ij} = \mathcal{A}(\phi_i, \phi_j), \quad i, j = 1, \ldots, N_h. \tag{2.20}$$

Também definimos os vetores

$$\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{N_h} \end{bmatrix} \in \mathbb{R}^{N_h} \quad \mathrm{b}_j = \mathcal{F}(\phi_j), \quad j = 1, \ldots, N_h. \tag{2.21}$$

e o vetor $\boldsymbol{u}_h = [x_1, \ldots, x_{N_h}]^T \in \mathbb{R}^{N_h}$.

Com a nova notação (2.17) e (2.19) são equivalentes ao problema

$$\begin{aligned} &\textit{Achar } u_h^* = \sum_{i=1}^{N_h} x_i \phi_i \textit{ tal que:} \\ &\left\{ \sum_{i=1}^{N_h} a_{ij} x_i = \quad b_j, \quad j = 1, \ldots, N_h \right. \end{aligned} \tag{2.22}$$

que em notação matricial é o sistema linear

$$\textit{Achar } \boldsymbol{u}_h \in \mathbb{R}^{N_h} \textit{ tal que: } A\boldsymbol{u}_h = \boldsymbol{b} \tag{2.23}$$

onde A e $\boldsymbol{b}$ foram definidos em (2.20) e (2.21) e a solução do problema (2.17) é dada pela combinação linear das funções base (2.18) com os pesos nas coordenadas do vetor $\boldsymbol{u}_h$ solução de (2.23), i.e.,

$$u_h^* = \sum_{i=1}^{N_h} x_i \phi_i.$$

O seguinte lema é muito útil na hora de estudar as propriedades da matriz de elementos finitos A definida em (2.20).

Lema 5. *Sejam $\boldsymbol{v}_h, \boldsymbol{w}_h \in \mathbb{R}^{N_h}$ as coordenadas das funções de elementos finitos $v_h, w_h \in V^h$. Então*

$$\boldsymbol{v}_h^T A \boldsymbol{w}_h = \mathcal{A}(v_h, w_h).$$

Agora vamos escolher um espaço V^h adequado. Em geral existem duas formas de definir V_h. O espaço V_h pode depender ou ser independente de uma malha, grade, partição ou triangulação. No caso de uma escolha de V_h dependente de uma malha, esta pode depender ou não do domínio de definição da equação diferencial parcial. A escolha do espaço V_h e a sua base são muito importantes. Estas escolhas tem implicações diretas na *condição* e *padrão de esparsidade* da matriz A definida em (2.20). A idéia geral do método dos elementos finitos é usar funções bases com *suporte pequeno* de tal forma que a condição da matriz *esparsa* resultante se mantenha moderada; entre outras vantagens do método. Neste minicurso usamos somente o espaço de funções lineares por partes baseado numa triangulação (partição) do domínio da equação diferencial.

2.3.1 O espaço de elementos finitos de funções lineares por partes

Suponha que o domínio da equação diferencial é o intervalo (a, b). Seja $\mathcal{T}^h$ uma *triangulação* ou partição do intervalo (a, b). Temos que a partição ou triangulação $\mathcal{T}^h$ é formada por intervalos pequenos, chamados *elementos*, da forma $K_i = (x_i, x_{i+1})$, $i = 1, \ldots, M-1$; onde os *vértices* $\{x_i\}_{i=1}^M$ satisfazem

$$a = x_1 < x_2 < \cdots < x_{M-1} < x_M = b$$

e os diâmetros dos elementos são de tamanho proporcional a $h > 0$,

$$\text{diâmetro}(K_i) = |x_{i+1} - x_i| = O(h) \quad \text{para todo } i = 1, \ldots, M-1.$$

Os vértices são divididos em, vértices *fronteira* $\{x_1, x_M\}$ e vértices *interiores* $\{x_i\}_{i=2}^{M-1}$. Definimos o espaço de funções contínuas lineares por partes associado a triangulação $\mathcal{T}^h$,

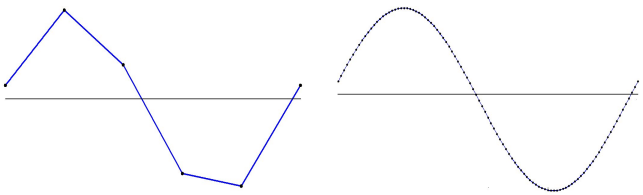
$$\mathbb{P}^1(\mathcal{T}^h) = \left\{ v \in C(a,b) \quad : \quad \begin{array}{l} v\big|_K \text{ é polinômio de grau 1 para todo} \\ \text{elemento } K \text{ da triangulação } \mathcal{T}^h \end{array} \right\}$$

onde $v\big|_K$ denota a restrição da função v ao elemento K. Na Figura 2.1 a) e b) vemos duas funções lineares por partes numa triangulação uniforme com $h = \frac{1}{5}$ e $h = \frac{1}{100}$, respectivamente. Também definimos $\mathbb{P}_0^1(\mathcal{T}^h)$ como o espaço de funções lineares por partes com valor zero nos pontos extremos (fronteira) do intervalo (a, b), isto é,

$$\mathbb{P}_0^1(\mathcal{T}^h) = \left\{ v \in \mathbb{P}^1(\mathcal{T}^h) : \quad v(x_1) = v(x_M) = 0 \right\}.$$

Lema 6. *O conjunto de função lineares por partes $\mathbb{P}^1(\mathcal{T}^h)$ é subconjunto do espaço $H^1(a, b)$. A derivada de uma função linear por partes é uma função constante por partes em $\mathcal{T}^h$.*

Dados $h > 0$ e uma triangulação do intervalo (a, b) pode-se então usar $V = \mathbb{P}^1(\mathcal{T}^h)$ em (2.17). A solução obtida em (2.17) é a aproximação de elementos finitos $\mathbb{P}^1(\mathcal{T}^h)$ da solução de (2.16). Denotamos por u^h esta aproximação.



a) b)

Figura 2.1: Exemplo de funções lineares por partes: a) com $h = 1/5$, b) com $h = 1/100$.

Note que se $v^h \in \mathbb{P}^1(\mathcal{T}^h)$, então

$$v^h(x) = \sum_{i=1}^{M} v^h(x_i)\phi_i(x) \tag{2.24}$$

onde as funções base ou funções chapéu ϕ_i, $i = 1, \ldots, M$, estão definidas por

$$\phi_i(x) = \begin{cases} 1, & \text{se } x = x_i, \text{ (1 no vértice } x_i) \\ 0, & \text{se } x = x_j, j \neq i, \text{(0 nos outros vértices)} \\ \text{extensão linear}, & \text{se } x \text{ não é vértice.} \end{cases} \tag{2.25}$$

Veja a Figura 2.2 para um exemplo de função base. Da equação (2.24) concluímos que $\mathbb{P}^1(\mathcal{T}^h)$ é gerado pelas M funções $\{\phi_i\}_{i=1}^{M}$. Observe também que $\mathbb{P}_0^1(\mathcal{T}^h)$ é gerado pelas $M-2$ funções $\{\phi_i\}_{i=2}^{M-1}$.

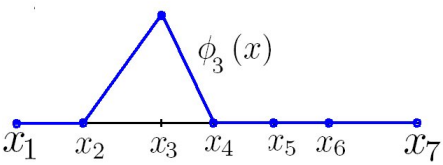


Figura 2.2: Função base numa triangulação não uniforme.

Lema 7. *Seja $K_i = (x_i, x_{i+1})$ um elemento da triangulação $\mathcal{T}^h$ de (a, b). Temos que*

$$\phi_i(x) = \frac{x_{i+1} - x}{x_{i+1} - x_i}, \quad \phi_{i+1}(x) = \frac{x - x_i}{x_{i+1} - x_i}, \quad \text{para todo } x \in K_i.$$

Observação 8. *De acordo com o Lema 7 dizemos que o método dos elementos finitos lineares por partes tem dois graus de liberdade por cada elemento.*

Da equação (2.24) vemos que o vetor $\boldsymbol{u}^h \in \mathbb{R}^M$ que representa as coordenadas da função de elementos finitos $u^h \in \mathbb{P}_0^h(\mathcal{T}^h)$ é dado pelos valores da função u^h nos vértices da triangulação $\mathcal{T}^h$, isto é,

$$\boldsymbol{u}^h = [u^h(x_1), u^h(x_2), \ldots, u^h(x_M)]^T \in \mathbb{R}^M. \tag{2.26}$$

O sistema linear obtido com funções do espaço de elementos finitos de funções lineares por partes é então

$$A\boldsymbol{u}^h = \boldsymbol{b} \tag{2.27}$$

onde $\boldsymbol{u}^h$ é como em (2.26), $\boldsymbol{b}$ é dado por (2.21) e a matriz A é calculada como em (2.20).

Observação 9 (Matrizes de Neumann e de Dirichlet)**.** *Se usamos todos os vértices (i.e., o espaço $\mathbb{P}^1(\mathcal{T}^h)$), a matriz obtida é de dimensão $M \times M$ e é chamada de Matriz de Neumann. Se usamos somente os vértices interiores (i.e., o espaço $\mathbb{P}_0^1(\mathcal{T}^h)$) obtemos uma matriz de dimensão $(M-2) \times (M-2)$ conhecida como matriz de Dirichlet.*

Comentaremos mais uma propriedade da matriz de elementos finitos. Esta propriedade é útil na hora de montar a matriz de elementos finitos. Consideramos somente o caso da forma bilinear definida em (2.13) e funções de elementos finitos lineares por partes.

Lema 10. *Seja $\mathcal{A}$ a forma bilinear definida em (2.13) e $\mathcal{T}^h$ uma triangulação de (a, b). Sejam $K_i = (x_i, x_{i+1})$, $i = 1, \ldots, M-1$ os elementos da triangulação. Defina R_i como a matriz $2 \times M$ de restrição ao elemento K_i, i.e., a matriz R_i tem todas as entradas nulas com excessão das posições $(1, i)$ (correspondente ao vértice x_i)*

e na posição $(2, i+1)$ *(do vértice* x_{i+1}*), onde tem o valor um. Seja* A_{K_i} *a matriz local definida por*

$$A_{K_i} = \begin{bmatrix} \mathcal{A}_{K_i}(\phi_i, \phi_i) & \mathcal{A}_{K_i}(\phi_i, \phi_{i+1}) \\ \mathcal{A}_{K_i}(\phi_{i+1}, \phi_i) & \mathcal{A}_{K_i}(\phi_{i+1}, \phi_{i+1}) \end{bmatrix} \tag{2.28}$$

onde $\mathcal{A}_{K_i}$*, a restrição da forma bilinear* $\mathcal{A}$ *ao elemento* K_i*, é dada por*

$$\mathcal{A}_{K_i}(v, w) = \int_{K_i} \kappa(x) v'(x) w'(x) dx. \tag{2.29}$$

Finalmente seja $\boldsymbol{b}_{K_i}$ *o lado direito local,*

$$\boldsymbol{b}_{K_i} = \begin{bmatrix} \mathcal{F}_{K_i}(\phi_i) \\ \mathcal{F}_{K_i}(\phi_{i+1}) \end{bmatrix} \tag{2.30}$$

onde $\mathcal{F}_{K_i}$*, a restrição de* $\mathcal{F}$ *ao elemento* K_i*, é dada por*

$$\mathcal{F}_{K_i}(v) = \int_{K_i} f(x) v(x) dx.$$

Temos então que

$$A = \sum_{i=1}^{M-1} R_i^T A_{K_i} R_i \tag{2.31}$$

e

$$\boldsymbol{b} = \sum_{i=1}^{M-1} R_i^T \boldsymbol{b}_{K_i} \tag{2.32}$$

O lema anterior permite montar a matriz A usando as contribuições locais de cada elemento. Isto representa uma grande vantagem na hora da implementação numérica. Na igualdade (2.31) o papel das matrizes R_i é somente colocar a contribuição local no lugar certo na matriz *global* A, desse modo, as matrizes R_i não precisam ser calculadas. Note também que as matrizes locais (2.29) são calculadas em cada elemento K_i separadamente.

Exemplo: a equação de Laplace

Suponha que queremos aproximar da solução a equação de Laplace (2.3) usando elementos finitos. Queremos resolver

$$\begin{array}{l} \textit{Achar } u : [0,1] \to \mathbb{R} \textit{ tal que:} \\ \begin{cases} -u''(x) = -1, & 0 < x < 1 \\ u(0) = 1, u(1) = 1. \end{cases} \end{array} \tag{2.33}$$

Vamos agora introduzir ideias básicas úteis na implementação computacional do método dos elementos finitos.

Formulação fraca

A formulação fraca de (2.33) foi construída na Seção 2.2.2. Lembramos a definição da forma bilinear $\mathcal{A}$ em (2.6) e do funcional linear $\mathcal{F}$ definido em (2.7).

Triangulação

Neste exemplo queremos usar a triangulação uniforme com quatro vértices,

$$\mathcal{T}^h = \{x_1 = 0, x_2 = \frac{1}{3}, x_3 = \frac{2}{3}, x_4 = 1\}.$$

Nesta triangulação temos três elementos, $K_1 = (0, \frac{1}{3}), K_2 = (\frac{1}{3}, \frac{2}{3})$ e $K_3 = (\frac{2}{3}, 1)$.

Montagem da matriz

Depois de definir a triangulação, montamos a matriz A usando (2.31), isto é, juntando as contribuições locais de cada elemento. Para isto precisamos construir, elemento por elemento, as matrizes locais A_{K_i} em (2.28) e as matrizes de restrição R_{K_i}, $i = 1, 2, 3$.

O primeiro elemento é $K_1 = (x_1, x_2) = (0, \frac{1}{3})$. As funções base em K_1 são $\phi_1(x) = \frac{x_2 - x}{x_2 - x_1} = 3(\frac{1}{3} - x)$ e $\phi_2 = \frac{x - x_1}{x_2 - x_1} = 3x$, $x \in (x_1, x_2)$. Com estas duas funções base calculamos as entradas da matriz local

A_{K_1} em (2.28) como segue,

$$\begin{aligned}
\mathcal{A}_{K_1}(\phi_1,\phi_1) &= \int_{K_1}\phi_1'\phi_1' = \int_{x_1}^{x_2}|\phi_1'|^2 = (-3)^2\frac{1}{3} = 3,\\
\mathcal{A}_{K_1}(\phi_1,\phi_2) &= \mathcal{A}_{K_1}(\phi_2,\phi_1) = \int_{K_1}\phi_0'\phi_1' = \int_{x_1}^{x_2}(-3)(3) = -3,\\
\mathcal{A}_{K_1}(\phi_2,\phi_2) &= \int_{K_1}\phi_2'\phi_2' = \int_{x_1}^{x_2}|\phi_2'|^2 = 3^2\frac{1}{3} = 3,
\end{aligned}$$

e portanto $A_{K_1} = 3\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$. Note que $R_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$ e

$$R_1^T A_{K_1} R_1 = 3\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Analogamente pode-se calcular as matrizes locais A_{K_i} e de restrição R_i para $i = 2$ e $i = 3$. Obtemos

$$A_{K_i} = 3\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad i = 1, 2, 3.$$

Temos finalmente que a matriz (Neumann) global é dada por

$$A = \sum_{i=1}^{3} R_i^T A_{K_i} R_i = 3\begin{bmatrix} (1) & -1 & 0 & 0 \\ -1 & (1+1) & -1 & 0 \\ 0 & -1 & (1+1) & -1 \\ 0 & 0 & -1 & (1) \end{bmatrix},$$

ou seja,

$$A = 3\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}.$$

Observamos que computacionalmente não é necessário criar as matrizes R_i e R_i^T, mas sim adicionar a matriz A_{K_i} nas posições corretas na matriz global A.

Montagem do lado direito

Agora montamos o lado direito $\boldsymbol{b}$ usando (2.32). Novamente vamos calcular, elemento por elemento, os lados direitos locais $\boldsymbol{b}_{K_i}$, $i = 1, 2, 3$, em (2.30). Neste exemplo a função do lado direito é $f(x) = -1$ para todo $x \in (0, 1)$. No elemento $K_1 = (x_1, x_2)$ temos

$$\mathcal{F}_{K_1}(\phi_1) = \int_{K_1} -1\phi_1(x)dx = -\int_{x_1}^{x_2} \frac{x - x_1}{x_2 - x_1}dx = -\frac{1}{6}$$
$$\mathcal{F}_{K_1}(\phi_2) = \int_{K_1} -1\phi_2(x)dx = -\int_{x_1}^{x_2} \frac{x_2 - x}{x_2 - x_1}dx = -\frac{1}{6}$$

e então $\boldsymbol{b}_{K_1} = -[\frac{1}{6}, \frac{1}{6}]^T$. Analogamente $\boldsymbol{b}_{K_2} = \boldsymbol{b}_{K_3} = -[\frac{1}{6}, \frac{1}{6}]^T$. Temos finalmente que

$$\boldsymbol{b} = \sum_{i=1}^{3} R_i^T \boldsymbol{b}_{K_i} = \begin{bmatrix} -\frac{1}{6} \\ -\frac{1}{6} - \frac{1}{6} \\ -\frac{1}{6} - \frac{1}{6} \\ -\frac{1}{6} \end{bmatrix} = - \begin{bmatrix} 1/6 \\ 1/3 \\ 1/3 \\ 1/6 \end{bmatrix}.$$

Solução do sistema linear

Temos que resolver o sistema linear

$$A\boldsymbol{u}_h = 3 \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} u_h(x_0) \\ u_h(x_1) \\ u_h(x_2) \\ u_h(x_3) \end{bmatrix} = \begin{bmatrix} -1/6 \\ -1/3 \\ -1/3 \\ -1/6 \end{bmatrix}.$$

Sabemos que $u_h(x_0) = -1$ e $u_h(x_1) = 1$. Podemos substituir no lado esquerdo,

$$\begin{bmatrix} u_h(x_0) \\ u_h(x_1) \\ u_h(x_2) \\ u_h(x_3) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ u_h(x_1) \\ u_h(x_2) \\ 0 \end{bmatrix}$$

e colocando o termo conhecido $A(-1,0,0,1)^T$ no lado direito, obtemos,

$$3\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}\begin{bmatrix} 0 \\ u_h(x_1) \\ u_h(x_2) \\ 0 \end{bmatrix} = \begin{bmatrix} -1/6 \\ -1/3 \\ -1/3 \\ -1/6 \end{bmatrix} - \begin{bmatrix} 3 \\ -3 \\ -3 \\ 3 \end{bmatrix}$$

ou

$$3\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} u_h(x_1) \\ u_h(x_2) \end{bmatrix} = \begin{bmatrix} 8/3 \\ 8/3 \end{bmatrix}$$

que resulta na solução $\begin{bmatrix} u_h(x_1) \\ u_h(x_2) \end{bmatrix} = \begin{bmatrix} 8/9 \\ 8/9 \end{bmatrix}$. A aproximação de elementos finitos é então

$$\boldsymbol{u}_h = \begin{bmatrix} u_h(x_0) \\ u_h(x_1) \\ u_h(x_2) \\ u_h(x_3) \end{bmatrix} = \begin{bmatrix} 1 \\ 8/9 \\ 8/9 \\ 1 \end{bmatrix} \quad \text{e} \quad u_h = \phi_1 + \frac{8}{9}\phi_2 + \frac{8}{9}\phi_3 + \phi_4.$$

A solução exata de (2.33) pode ser calculada facilmente. O objetivo

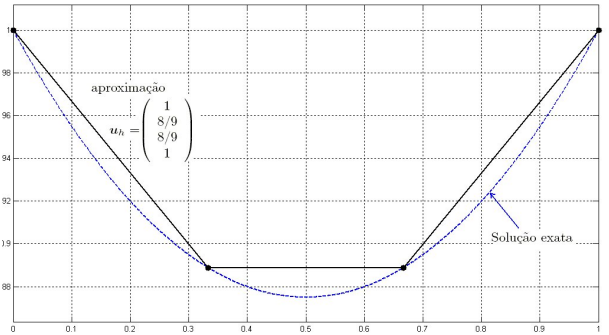


Figura 2.3: Solução exata de (2.33), linha pontilhada, e aproximação de elementos finitos com $h = 1/3$, linha sólida.

deste exemplo é mostrar o potencial do método dos elementos finitos. Na Figura 2.3 comparamos a solução exata u e a aproximação de elementos finitos obtida acima $u_h = \phi_1 + \frac{8}{9}\phi_2 + \frac{8}{9}\phi_3 + \phi_4$. Lembre que

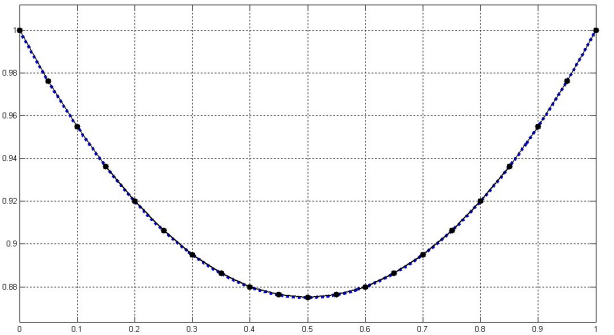


Figura 2.4: Solução exata de (2.33), linha pontilhada, e aproximação de elementos finitos com $h = 1/10$, linha sólida.

neste exemplo $h = 1/3$ e resolvemos um sistema linear de tamanho 2×2.

Para finalizar esta seção resolvemos a equação (2.33) com uma triangulação mais fina que a usada no exemplo, usamos uma triangulação uniforme com $h = 1/10$. Na figura 2.4 vemos a aproximação de elementos finitos para esta malha mais fina. O sistema linear resolvido para obter a aproximação de elementos finitos com $h = 1/10$ é de tamanho 9×9.

2.3.2 O sistema linear obtido

Nas aplicações praticas do método dos elementos finitos o tamanho da matriz do sistema linear em (2.23) é muito grande, especialmente em duas e três dimensões. É necessário então, escolher o método adequado para aproximar a solução deste sistema linear. No caso da equação elíptica básica (2.10) e o espaço de elementos finitos de funções linear por partes, a matriz resultante A tem as seguintes propriedades que podem ser deduzidas do Lema 5 a das propriedades das funções base (2.25),

- O tamanho da matriz é gigantesco.

- A matriz A é esparsa: isto é, uma pequena porcentagem das entradas da matriz é diferente de zero. Este é o principal motivo da escolha de funções base de elementos finitos com *suporte* pequeno.

- A matriz de Neumann é semi-definida positiva.

- A matriz de Dirichlet é definida positiva.

- Para a matriz da forma bilinear $\mathcal{A}$ definida em (2.13) com o coeficiente κ satisfazendo (2.11) temos que existem constantes positivas C e c, que dependem unicamente da triangulação $\mathcal{T}^h$ e do domínio D, tais que

$$\lambda_{\max} \leq C\kappa_{\max}\frac{1}{h} \quad \text{e} \quad \lambda_{\min} \geq c\kappa_{\min}h,$$

onde $\lambda_{\max}$ e $\lambda_{\min}$ são o maior e menor autovalor da matriz A. Podemos então estimar o número de condição (espectral) da matriz A,

$$\text{Cond}(A) := \frac{\lambda_{\max}}{\lambda_{\min}} \leq \frac{C}{c}\frac{\kappa_{\max}}{\kappa_{\min}}\frac{1}{h^2}. \tag{2.34}$$

2.4 Erro de aproximação

Dados $h > 0$ e uma triangulação do intervalo (a, b) pode-se então usar $V = \mathbb{P}^1(\mathcal{T}^h)$ em (2.17). A solução obtida em (2.17) é a aproximação de elementos finitos $\mathbb{P}^1(\mathcal{T}^h)$ da solução de (2.16). Denotamos por u^h esta solução. O erro de aproximação $u - u^h$ poder ser calculado usando resultados da análise, por exemplo temos o seguinte lema; veja [21].

Lema 11 (Estimativa de erro 'a priori')**.** *Sejam u e u^h as soluções da formulação fraca (2.16) e da formulação de Galerkin (2.17) com $V^h = \mathbb{P}^1(\mathcal{T}^h)$, respectivamente. Temos que existe uma constante C que é independente de $h > 0$ e de u, tal que*

$$|u - u^h|_{H^1(a,b)} \leq Ch|u|_{H^2(a,b)} \quad e \quad \|u - u^h\|_{L^2(a,b)} \leq Ch^2|u|_{H^2(a,b)}.$$

2.5 Experimentos numéricos

Defina

$$\kappa_1(x,\mu) = \begin{cases} 1, & x \in (0,\frac{1}{5}) \cup (\frac{2}{5},\frac{3}{5}) \cup (\frac{4}{5},1) \\ \\ \mu, & x \in (\frac{1}{5},\frac{2}{5}) \cup (\frac{3}{5},\frac{4}{5}) \end{cases} \tag{2.35}$$

e

$$k_2(x,p) = 1 + \sin(2\pi p x). \tag{2.36}$$

Considere a equação

$$\begin{array}{l} \textit{Achar } u : [0,1] \to \mathbb{R} \textit{ tal que:} \\ \begin{cases} -(\kappa u'(x))' = -1, & 0 < x < 1 \\ u(0) = 0, u(1) = 1. \end{cases} \end{array} \tag{2.37}$$

onde

$$\kappa(x) = \kappa_1(x,\mu) + 100\kappa_2(x,p)$$

para valores dados de μ e p. Note que entre maior o μ maior o contraste do meio (contraste$(\kappa) = \kappa_{\max}/\kappa_{\min}$) e que p introduz uma variação considerável do coeficiente na escala fina.

Apresentamos alguma das soluções de elementos finitos com diferentes valores de μ, p e h.

Primeiro consideramos o caso $\mu = 1000$ e $n = 1$ que corresponde ao caso de um meio poroso com contraste alto e sem variações nas escalas finas. Na Figura 2.5 observamos aproximações de elementos finitos para diferentes valores de h. Note que para os primeiros valores de h, obtemos aproximações não muito boas quando comparadas com as aproximações para valores menores de h. Compare com o exemplo da equação de Laplace no final da Seção 2.3.1. Concluímos que, neste exemplo, o valor de h requerido para obter aproximações satisfatórias depende do contraste do meio. Observamos também que o sistema linear resolvido é muito mal condicionado, veja (2.34), o que dificulta ainda mais a obtenção de soluções acuradas de forma eficiente. Esta situação é similar ou pior em duas e três dimensões. Observamos novamente a necessidade de usar métodos acurados e eficientes para calcular as soluções de equações diferenciais parciais

elípticas em meios porosos com contraste alto.

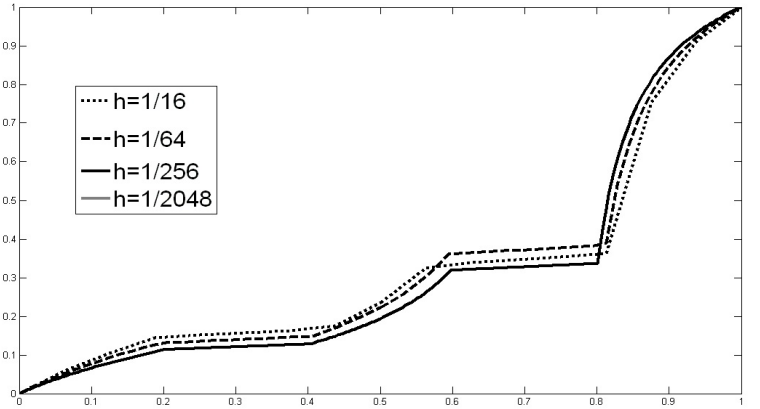


Figura 2.5: Aproximações de elementos finitos para diferentes valores de h. Aqui usamos o coeficiente $\kappa(x) = \kappa_1(x, 1000) + 100\kappa_2(x, 1)$. Notamos que a curva correspondente ao $h = 1/2048$ coincide com a curva correspondente ao $h = 1/256$ na escala da figura.

Agora consideramos o caso $\mu = 100$ e $p = 30$ que corresponde ao caso com contraste alto e características finas. Nas Figuras 2.6 e 2.7 apresentamos a solução de elementos finitos para vários valores do h. Na Figura 2.6 temos que com uma malha do tamanho $h = 1/64$ obtemos uma aproximação razoável do comportamento da solução mais a aproximação do comportamento da solução na escala fina pode ser melhorada. Na Figura 2.7 vemos que com uma malha mais fina obtemos melhor aproximação do comportamento da solução na escala fina. Compare com o exemplo da equação de Laplace no final da Seção 2.3.1. Concluímos que, neste exemplo, para obter uma boa aproximação do comportamento da solução nas escalas de variação do coeficiente κ devemos usar uma malha suficientemente fina. Isto requer a solução de sistemas lineares gigantescos. Entre menor a escala de variação que queremos representar, maior o sistema linear obtido. Lembrei que o tamanho do sistema junto com a condição da matriz tem implicações direitas no tempo e custo computacional. A

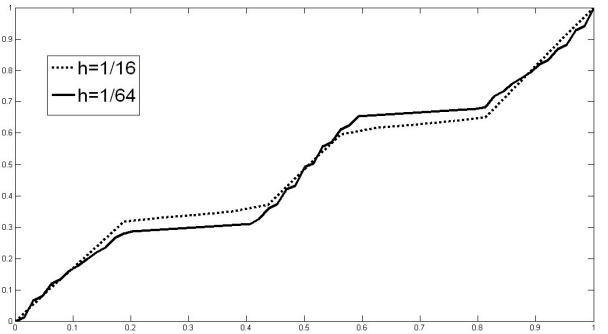


Figura 2.6: Aproximações de elementos finitos para $h = 1/16$ e $h = 1/64$ com $\kappa(x) = \kappa_1(x, 1000) + 100\kappa_2(x, 30)$.

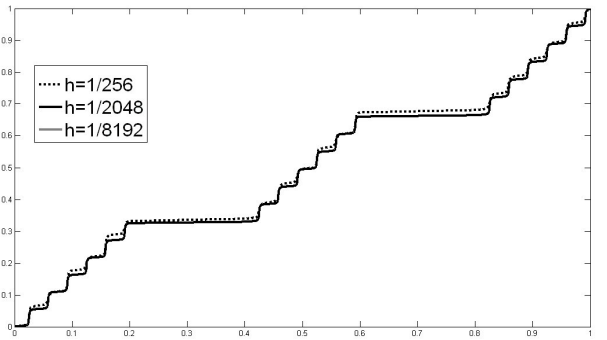


Figura 2.7: Aproximações de elementos finitos para $h = 1/256, 1/2048$ e $1/8192$ e $\kappa(x) = \kappa_1(x, 1000) + 100\kappa_2(x, 30)$. A curva correspondente ao $h = 1/8192$ coincide com a curva correspondente ao $h = 1/2048$ na escala da figura.

situação fica pior em dimensões maiores. Observamos novamente a necessidade de usar métodos acurados e eficientes para calcular as soluções de equações diferenciais parciais elípticas em meios porosos com contraste alto e múltiplas escalas.

Capítulo 3

O método dos elementos finitos em duas dimensões

O objetivo deste capítulo é estudar o método dos elementos finitos usando funções lineares por partes em duas dimensões. O leitor é referido à Seção 2.1 para uma discussão geral. Para um estudo mais detalhado veja [7, 8, 21, 3, 10, 17, 16] e as referências ali citadas.

3.1 Introdução

Neste capítulo trabalhamos num subconjunto aberto, conexo e poligonal $D \subset \mathbb{R}^2$ que é o domínio físico onde a equação diferencial é formulada. Denotamos $x = (x_1, x_2) \in \mathbb{R}^2$. Em geral pode-se considerar subconjuntos *Lipschitz* de $\mathbb{R}^2$ e não somente subconjuntos poligonais. Denotamos por ∂D a fronteira do domínio D. Dada uma função $v : D \to \mathbb{R}$ denotamos por $\frac{\partial v}{\partial x_1} = \partial_1 v$ e $\frac{\partial v}{\partial x_2} = \partial_2 v$ as suas derivadas parciais. Notação similar é usada para as derivadas parciais de ordem dois, $\frac{\partial^2 v}{\partial x_1^2} = \partial_{11}^2 v$, $\frac{\partial^2 v}{\partial_{x_2} \partial x_1} = \partial_{21}^2 v$ $\frac{\partial^2 v}{\partial x_2^2} = \partial_{22}^2 v$. Denotamos

por ∇v o vetor *linha*

$$\nabla v = \left[\frac{\partial v}{\partial x_1}, \frac{\partial v}{\partial x_2} \right].$$

Note que se $w : D \to \mathbb{R}$, temos os seguintes produtos

$$\nabla v \cdot \nabla w = (\nabla v)^T \nabla w = \frac{\partial v}{\partial x_1}\frac{\partial w}{\partial x_1} + \frac{\partial v}{\partial x_2}\frac{\partial w}{\partial x_2} = \sum_{i=1}^{2} \partial_i v \partial_i w$$

$$|\nabla v|^2 = \nabla v \cdot \nabla v = \left| \frac{\partial v}{\partial x_1} \right|^2 + \left| \frac{\partial v}{\partial x_2} \right|^2 = \sum_{i=1}^{2} |\partial_i v|^2$$

e se

$$\kappa(x) = \begin{bmatrix} \kappa_{11}(x) & \kappa_{12}(x) \\ \kappa_{21}(x) & \kappa_{22}(x) \end{bmatrix} \tag{3.1}$$

é uma matriz 2×2 para cada $x \in D$, então

$$\nabla v \kappa \nabla w = \nabla v \kappa \cdot \nabla w = \nabla v \kappa (\nabla w)^T = \sum_{i=1}^{2}\sum_{i=1}^{2} \kappa_{ij} \partial_i v \partial_j w.$$

Dadas duas funções $v, w : D \subset \mathbb{R}^2 \to \mathbb{R}$, definimos o divergente do vetor $(v, w) \in \mathbb{R}^2$ por

$$\operatorname{div}(v, w) = \partial_1 v + \partial_2 w.$$

Para toda função u, o Laplaciano Δu é definido por

$$\Delta u = \operatorname{div}(\nabla u) = \frac{\partial^2}{\partial x_1^2} u + \frac{\partial^2}{\partial x_2^2} u$$

e o operador diferencial elíptico $\operatorname{div}(\kappa \nabla(\cdot))$ é

$$\operatorname{div}(\kappa \nabla u) = \sum_{i=1}^{2}\sum_{j=1}^{2} \partial_i (\kappa_{ij} \partial_j u).$$

Quando $\kappa(x) = \tilde{\kappa}(x) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ usaremos a notação

$$\operatorname{div}(\tilde{\kappa} \nabla u) = \operatorname{div}(\kappa \nabla u) = \sum_{i=1}^{2} \partial_i (\tilde{\kappa} \partial_i u) = \partial_1 (\tilde{\kappa} \partial_1 u) + \partial_2 (\tilde{\kappa} \partial_2 u).$$

Vamos usar as seguintes formulas de integração por partes (primeira identidade de Green)

$$\int_D (\Delta u) v dx = \int_{\partial D} v(\nabla u \cdot \boldsymbol{\eta}) dS - \int_D \nabla u \cdot \nabla v dx, \qquad (3.2)$$

onde para cada $x \in \partial D$, $\boldsymbol{\eta}(x)$ denota o vetor normal com sentido para o exterior de D em x. Com κ em (3.1) a mesma fórmula é

$$\int_D \operatorname{div}(\kappa \nabla u) v dx = \int_{\partial D} v(\kappa \nabla u \cdot \boldsymbol{\eta}) dS - \int_D \nabla u \cdot \kappa \nabla v dx \qquad (3.3)$$

que é o análogo da fórmula de integração por partes em uma dimensão, $\int_a^b (\kappa u')' v = (\kappa u' v)|_a^b - \int_a^b \kappa u' v'$.

3.2 Espaços de funções

Nesta seção definimos o espaço de funções adequado para construir a formulação fraca das equações diferenciais elípticas em duas dimensões. Veja [2, 21, 9, 23].

Denote por $C_0^\infty(D)$ as funções teste infinitamente diferenciáveis e com suporte compacto contido em D. Denotamos por $L^2(D)$ o espaço das funções de quadrado integrável segundo a medida de Lebesgue. Definimos também

$$H^1(D) := \left\{ v \in L^2(D) \mid \quad \frac{\partial v}{\partial x_1}, \frac{\partial v}{\partial x_2} \in L^2(D) \right\}. \qquad (3.4)$$

O espaço $H^1(D)$ é um espaço de Hilbert com o produto interno

$$\begin{aligned} (v, w)_{H^1(D)} &= \int_D v(x) w(x) dx + \int_D \nabla v(x) \cdot \nabla w(x) dx \\ &= (v, w)_{L^2(D)} + (\partial_1 v, \partial_1 w)_{L^2(D)} + (\partial_2 v, \partial_2 w)_{L^2(D)} \end{aligned}$$

e norma

$$\begin{aligned} \|v\|^2_{H^1(D)} &= \int_D |v(x)|^2 dx + \int_D |\nabla v(x)|^2 dx \\ &= \|v\|^2_{L^2(D)} + \|\partial_1 v\|^2_{L^2(D)} + \|\partial_2 v\|^2_{L^2(D)}. \end{aligned}$$

Também usaremos o espaço de funções $H_0^1(D) \subset H^1(D)$ definido por

$$H_0^1(D) = \{v \in H^1(D) \mid \ v = 0 \text{ em } \partial D\} \tag{3.5}$$

onde $v = 0$ em ∂D quer dizer que $\int_{\partial D} v^2 = 0$, isto é, $v = 0$ no sentido $L^2(\partial D)$.

Finalmente definimos

$$H^2(D) := \left\{ v \in L^2(D) \mid \ \frac{\partial v}{\partial x_i}, \frac{\partial^2 v}{\partial x_i \partial x_j} \in L^2(D), \ \ i,j = 1,2 \right\} \tag{3.6}$$

com norma

$$\|v\|^2_{H^2(D)} \ = \ \|v\|^2_{L^2(D)} + \sum_{i=1}^{2} \|\partial_i v\|^2_{L^2(D)} + \sum_{i=1}^{2}\sum_{j=1}^{2} \|\partial_{ij} v\|^2_{L^2(D)}.$$

Note que $H^2(D) \subset H^1(D) \subset L^2(D)$.

3.3 Formulação fraca

Agora vamos estudar o que seria o análogo em duas dimensões a formulação fraca em uma dimensão da Seção 2.2. Para construir a formulação fraca de um problema elíptico posto em $D \subset \mathbb{R}^2$ temos que

1. Supor que existe uma solução u de nossa equação diferencial, multiplicar os dois lados da equação por uma *função teste* $v \in C_0^\infty(D)$ e tomar integrais nos dois lados da igualdade.

2. Usar a fórmula de integração por partes em $\mathbb{R}^2$ (primeira identidade de Green) e a condição de contorno para ficar com expressões que requeiram tomar somente derivadas da mais baixa ordem possível.

3. Trocar $v \in C_0^\infty(a,b)$ por $v \in V = H_0^1(D)$.

Usaremos os exemplos introduzidos no Capítulo 1 para ilustrar o procedimento acima.

3.3.1 Exemplo: a equação de Laplace

Considere a seguinte equação diferencial parcial elíptica de Laplace,

$$\begin{array}{l} \textit{Achar } u : D \subset \mathbb{R}^2 \to \mathbb{R} \textit{ tal que:} \\ \left\{ \begin{array}{rll} -\Delta u(x) = & f(x), & \text{para } x = (x_1, x_2) \in D \\ u(x) = & g(x), & \text{para } x = (x_1, x_2) \in \partial D. \end{array} \right. \end{array} \tag{3.7}$$

Para obter a formulação fraca de (3.7) multiplicamos os dois lados da primeira igualdade por $v \in C_0^\infty(D)$ fixa mais arbitrária, depois integramos os dois lados da equação e obtemos

$$\begin{array}{l} \textit{Achar } u : D \subset \mathbb{R}^2 \to \mathbb{R} \textit{ tal que:} \\ \left\{ \begin{array}{rll} -\displaystyle\int_D \Delta u(x) v(x) dx = & \displaystyle\int_D f(x) v(x) dx, & \forall v \in C_0^\infty(D) \\ u(x) = & g(x), & \text{para } x \in \partial D. \end{array} \right. \end{array} \tag{3.8}$$

Usamos a fórmula de integração por partes (3.2) e o fato $v(x) = 0$ para todo $x \in \partial D$ e obtemos

$$\begin{aligned} -\int_D \Delta u(x) v(x) dx &= \int_D \nabla u(x) \cdot \nabla v(x) dx - \int_{\partial D} v(x) (\nabla u(x) \cdot \boldsymbol{\eta}(x)) dS \\ &= \int_D \nabla u(x) \cdot \nabla v(x) dx - 0 = \int_D \nabla u(x) \cdot \nabla v(x) dx \end{aligned}$$

A equação (3.8) pode ser escrita então como

$$\begin{array}{l} \textit{Achar } u : D \subset \mathbb{R}^2 \to \mathbb{R} \textit{ tal que:} \\ \left\{ \begin{array}{rll} \displaystyle\int_D \nabla u(x) \cdot \nabla v(x) dx = & \displaystyle\int_D f(x) v(x) dx, & \forall v \in C_0^\infty(D) \\ u(x) = & g(x), & \text{para } x \in \partial D. \end{array} \right. \end{array} \tag{3.9}$$

Defina

$$\mathcal{A}(u, v) = \int_D \nabla u(x) \cdot \nabla v(x) dx \quad \text{para toda } u, v \in H^1(D) \tag{3.10}$$

e

$$\mathcal{F}(v) = \int_D f(x) v(x) dx \quad \text{para toda } v \in H^1(D). \tag{3.11}$$

Note que $\mathcal{A} : U \times V \to \mathbb{R}$ é uma forma bilinear e $\mathcal{F} : V \to \mathbb{R}$ é um funcional linear. Seguindo as mesmas ideias da Seção 2.2 concluímos que a formulação fraca de (3.7) é

$$\begin{array}{l} \textit{Achar } u \in U \textit{ tal que:} \\ \left\{ \begin{array}{rcll} \mathcal{A}(u,v) = & \mathcal{F}(v) & \forall v \in H^1_0(D) \\ u = & g, & \text{em } \partial D. \end{array} \right. \end{array} \tag{3.12}$$

3.3.2 Exemplo: equação elíptica básica em meios heterogêneos

Considere a seguinte equação diferencial parcial elíptica,

$$\begin{array}{l} \textit{Achar } u : D \subset \mathbb{R}^2 \to \mathbb{R} \textit{ tal que:} \\ \left\{ \begin{array}{rcll} -\text{div}(\kappa(x)\nabla u(x)) = & f(x), & \text{para } x \in D \\ u(x) = & g(x), & \text{para } x \in \partial D \end{array} \right. \end{array} \tag{3.13}$$

onde o tensor de permeabilidade (ou condutividade) κ é definido em (3.1). Assumimos que existem $\kappa_{\min}$ e $\kappa_{\max}$ tais que para todo $x \in D$ temos

$$0 < \kappa_{\min} \leq \mu_{\min}(x) \leq \mu_{\max}(x) \leq \kappa_{\max}, \tag{3.14}$$

onde $\mu_{\min}(x)$ e $\mu_{\max}(x)$ são o menor e maior autovalor da matriz $\kappa(x)$ em (3.1).

Para obter a formulação fraca de (3.13) multiplicamos os dois lados da primeira igualdade por $v \in C^\infty_0(D)$ fixa mas arbitrária, depois integramos os dois lados da equação, aplicamos a fórmula de integração por partes (3.3) e obtemos

$$\begin{aligned} -\int_a^b \text{div}(\kappa(x)\nabla u(x))v(x)dx &= \int_D \nabla u(x)\kappa \cdot \nabla v(x)dx \\ &\quad - \int_{\partial D} v(x)(\nabla u(x)\kappa(x) \cdot \boldsymbol{\eta}(x))dS \\ &= \int_D \nabla u(x)\kappa(x) \cdot \nabla v(x)dx - 0 \\ &= \int_D \nabla u(x)\kappa(x) \cdot \nabla v(x)dx. \end{aligned}$$

A formulação fraca de (3.13) pode ser escrita como

$$\begin{array}{l} \textit{Achar } u \in H^1(D) \textit{ tal que:} \\ \left\{ \begin{array}{rcll} \mathcal{A}(u,v) = & \mathcal{F}(v) & \forall v \in H_0^1(D) \\ u = & g, & \text{em } x \in \partial D, \end{array} \right. \end{array} \tag{3.15}$$

onde a forma bilinear $\mathcal{A}$ é definida por

$$\mathcal{A}(u,v) = \int_D \nabla u(x)\kappa(x) \cdot \nabla v(x) dx \quad u, v \in H^1(D) \tag{3.16}$$

e o funcional linear $\mathcal{F}$ é o mesmo definido em (3.11).

3.4 Formulação de Galerkin

Como no caso de uma dimensão espacial, podemos considerar a formulação fraca (3.12), (3.15) ou qualquer problema da forma

$$\begin{array}{l} \textit{Achar } u^* \in V \textit{ tal que:} \\ \left\{ \begin{array}{rcll} \mathcal{A}(u^*,v) = & \mathcal{F}(v) & \forall v \in V \end{array} \right. \end{array} \tag{3.17}$$

onde $V = H_0^1(D)$ ou V é um espaço de funções em D. Como antes, trocamos os espaços de dimensão infinita na formulação fraca (3.17) por espaços de dimensão finita, i.e., espaços de elementos finitos.

Considere V^h, $h > 0$, subespaços de dimensão finita $V^h \subset V$. A formulação de Galerkin de (3.17) é então

$$\begin{array}{l} \textit{Achar } u_h^* \in V^h \textit{ tal que:} \\ \left\{ \begin{array}{rcll} \mathcal{A}(u_h^*,v_h) = & \mathcal{F}(v_h) & \text{para toda } v_h \in V^h. \end{array} \right. \end{array} \tag{3.18}$$

Depois de escolher uma base de V^h,

$$\{\phi_1, \phi_2, \ldots, \phi_{N_h}\} \quad \text{onde } N_h \text{ é a dimensão de } V^h, \tag{3.19}$$

obtemos o sistema linear $N_h \times N_h$ (veja a Seção 2.3),

$$A\boldsymbol{u}_h = \boldsymbol{b}$$

onde a representação matricial $A = [a_{ij}]$ da forma bilinear $\mathcal{A}$ tem entradas

$$a_{ij} = \mathcal{A}(\phi_i, \phi_j), \quad i, j = 1, \ldots, N_h, \tag{3.20}$$

e

$$\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{N_h} \end{bmatrix} \in \mathbb{R}^{N_h} \quad \mathrm{b}_j = \mathcal{F}(\phi_j), \quad j = 1, \ldots, N_h, \tag{3.21}$$

e o vetor $\boldsymbol{u}_h = [x_1, \ldots, x_{N_h}]^T \in \mathbb{R}^{N_h}$.

3.4.1 O espaço de elementos finitos de funções lineares por partes em duas dimensões

Seja $D \subset \mathbb{R}^2$ um domínio poligonal. Uma *triangulação* (ou *malha*) $\mathcal{T}^h$ do domínio D é uma partição de D em subconjuntos disjuntos de D chamados *elementos*, isto é, $\mathcal{T}^h = \{K_i\}_{i=1}^{N_h^e}$ com

$$\bigcup_{i=1}^{N_h^e} \overline{K}_i = \overline{\Omega}, \quad K_i \cap K_j = \emptyset \text{ for } i \neq j, \text{ e } h = \max_{1 \leq i \leq N_h^e} \text{diâmetro}(K_i).$$

A triangulação $\mathcal{T}^h$ é chamada *geometricamente conforme* se a interseção dos fechos de dois elementos diferentes ($i \neq j$) $\overline{K}_i \cap \overline{K}_j$ é um lado ou um vértice comum aos dois elementos. Veja Figura 3.1. O triângulo de referência $\widehat{K}$ é o triângulo com vértices $(0, 0)$,

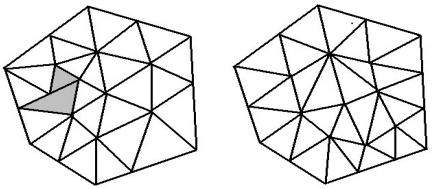


Figura 3.1: Exemplo de triangulação geometricamente não conforme (esquerda) e geometricamente conforme(direita).

$(0,1)$ e $(1,0)$. Neste minicurso consideramos unicamente malhas formadas por triângulos que são a imagem por uma aplicação afim do triângulo de referência, isto é, para cada elemento K da malha existe uma aplicação $F_K : \widehat{K} \to K$ que é da forma

$$F_K(\hat{x}) = \begin{bmatrix} b_{11}^K & b_{12}^K \\ b_{21}^K & b_{22}^K \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} + \begin{bmatrix} c_1^K \\ c_2^K \end{bmatrix}.$$

Para cada elemento $K \in \mathcal{T}^h$ define-se o fator de aspecto $\rho(K)$ por $\rho(K) = \text{diâmetro}(K)/r_K$, onde r_K é o raio do maior círculo contido em K. Veja Figura 3.2.

Figura 3.2: Elemento com fator de aspecto pequeno (esquerda) e grande (direita).

Uma familia de triangulações $\{\mathcal{T}^h\}_{h>0}$ é dita de *aspecto regular* se existe uma constante $C > 0$ independente de h tal que $\rho(K) \leq C$ para todo elemento $K \in \mathcal{T}^h$. A familia de triangulações $\{\mathcal{T}^h\}_{h>0}$ é dita *quase-uniforme* se existe uma constante C independente de h tal que $\text{diam}(K) \geq Ch$ para todo elemento $K \in \mathcal{T}^h$. Assumiremos que todas as triangulações usadas neste minicurso são de aspecto regular e quase-uniformes.

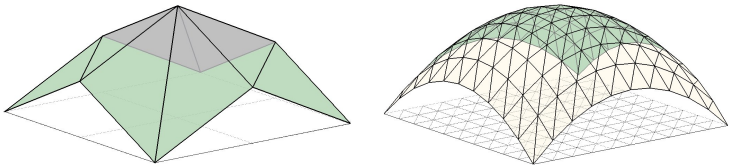
Dada uma triangulação $\mathcal{T}^h$, seja N_h^v o número de vértices da triangulação. Os vértices da triangulação $\{x_i\}$ são divididos em, *fronteira* $\{x_i \in \partial D\}$ e *interiores* $\{x_i \in D\}$. Definimos o espaço de funções contínuas lineares por partes associado a triangulação $\mathcal{T}^h$,

$$\mathbb{P}^1(\mathcal{T}^h) = \left\{ v \in C(D) \quad : \quad \begin{array}{l} v\big|_K \text{ é polinômio em duas variáveis de} \\ \text{grau total 1 para todo elemento } K \text{ da} \\ \text{triangulação } \mathcal{T}^h \end{array} \right\}$$

onde $v\big|_K$ denota a restrição da função v ao elemento K. Na Figura 3.3 a) e b) mostramos duas funções lineares por partes numa triangulação

estruturara com $h = \frac{1}{2}$ e $h = \frac{1}{10}$, respectivamente. Também definimos $\mathbb{P}_0^1(\mathcal{T}^h)$ como o espaço de funções lineares por partes com valor zero nos vértices fronteira da triangulação, isto é,

$$\mathbb{P}_0^1(\mathcal{T}^h) = \left\{ v \in \mathbb{P}^1(\mathcal{T}^h) : \quad v(x) = 0 \text{ para todo } x \in \partial D \right\}.$$



a) b)

Figura 3.3: Exemplo de funções lineares por partes num quadrado: a) com $h = 1/2$, b) com $h = 1/10$.

Lema 12. *O conjunto de funções lineares por partes $\mathbb{P}^1(\mathcal{T}^h)$ é subconjunto do $H^1(D)$. O gradiente de uma função linear por partes é uma função (vetorial) constante por partes em $\mathcal{T}^h$.*

Dados $h > 0$ e uma triangulação do domínio D pode-se então usar $V = \mathbb{P}^1(\mathcal{T}^h)$ em (3.18). A solução obtida em (3.18) é a aproximação de elementos finitos $\mathbb{P}^1(\mathcal{T}^h)$ da solução de (3.17).

Note que se $v^h \in V = \mathbb{P}^1(\mathcal{T}^h)$, então podemos escrever

$$v^h(x) = \sum_{i=1}^{N_h^v} v^h(x_i)\phi_i(x) \tag{3.22}$$

onde N_h^v é o número de vértices da malha e as funções base ou funções chapéu , ϕ_i, $i = 1, \ldots, N_h^v$, estão definidas por

$$\phi_i(x) = \begin{cases} 1, & \text{se } x = x_i, \text{ (1 no vértice } x_i) \\ 0, & \text{se } x = x_j, j \neq i, \text{(0 nos outros vértices)} \\ \text{extensão linear}, & \text{se } x \text{ não é vértice.} \end{cases} \tag{3.23}$$

Também usaremos a notação $\phi_{x_i} = \phi_i$, $i = 1, \ldots, N_h^v$. Veja a Figura 3.4 para um exemplo de uma função base e seu suporte. Concluímos que $\mathbb{P}^1(\mathcal{T}^h)$ é gerado pelas N_h^v funções $\{\phi_i\}_{i=1}^{N_h^v}$. Observe também que $\mathbb{P}_0^1(\mathcal{T}^h)$ é gerado pelas funções $\{\phi_i : x_i \text{ é vértice interior}\}$.

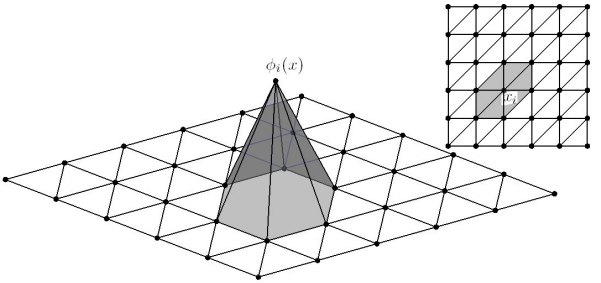


Figura 3.4: Exemplo de uma função base em duas dimensões.

Usando a equação (3.22) o vetor $\boldsymbol{u}^h \in \mathbb{R}^{N_h^v}$ que representa as coordenadas da função de elementos finitos $u^h \in \mathbb{P}^h(\mathcal{T}^h)$ na base das função chapéu é dado pelos valores da função u^h nos vértices da triangulação $\mathcal{T}^h$, isto é,

$$\boldsymbol{u}^h = [u^h(x_1), \ldots, u^h(x_{N_h^v})]^T \in \mathbb{R}^{N_h^v}. \tag{3.24}$$

Analogamente ao caso em uma dimensão, depois de escolher a base, a formulação de Galerkin equivale a resolver um sistema linear. O sistema linear obtido com o espaço de elementos finitos de funções lineares por partes é então

$$A\boldsymbol{u}^h = \boldsymbol{b} \tag{3.25}$$

onde $\boldsymbol{u}^h$ é como em (3.24), $\boldsymbol{b}$ é dado por (3.21) e a matriz A é definida como em (3.20).

Observação 13 (Matrizes de Neumann e de Dirichlet)**.** *Se usamos todos os vértices (i.e., os espaço $\mathbb{P}^1(\mathcal{T}^h)$) a matriz obtida é de dimensão $N_h^v \times N_h^v$ e é chamada de Matriz de Neumann. Se usamos somente os vértices interiores (i.e., o espaço $\mathbb{P}_0^1(\mathcal{T}^h)$) obtemos uma matriz de dimensão menor conhecida como matriz de Dirichlet.*

Agora descrevemos melhor as funções base do espaço $\mathbb{P}^1(\mathcal{T}^h)$.

Lema 14. *Seja $\widehat{K}$ o triângulo de referência com vértices $(0,0)$, $(1,0)$ e $(0,1)$ e $\hat{\phi}_1$, $\hat{\phi}_2$ e $\hat{\phi}_3$ as respectivas funções bases. As três funções bases no $\widehat{K}$ são*

$$\hat{\phi}_1(\hat{x}) = 1 - \hat{x}_1 - \hat{x}_2, \quad \hat{\phi}_2(\hat{x}) = \hat{x}_1, \quad \hat{\phi}_3(\hat{x}) = \hat{x}_2,$$

para todo $\hat{x} = (\hat{x}_1, \hat{x}_2) \in \widehat{K}$.

Lema 15. *Seja K_i um elemento da triangulação $\mathcal{T}^h$ com vértices $u = (u_1, u_2), v = (v_1, v_2)$ e $z = (z_1, z_2)$ (ordenados no sentido anti-horário) e $\phi_1 = \phi_u$, $\phi_2 = \phi_v$ e $\phi_3 = \phi_z$ as respectivas funções base. Seja $\widehat{K}$ o triângulo de referência e $F_{K_i} : \widehat{K} \to K$ a função afim tal que $F_{K_i}(\widehat{K}) = K_i$ e $F_{K_i}(0) = u$. Então*

$$F_{K_i}(\hat{x}) = \begin{bmatrix} v_1 - u_1 & z_1 - u_2 \\ v_2 - u_1 & z_2 - u_2 \end{bmatrix} \hat{x} + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \quad \hat{x} \in \hat{K}$$

e para cada $x \in K_i$, as três funções bases com suporte no elemento K_i são

$$\phi_j(x) = \hat{\phi}_j(\hat{x}), \text{ para todo } \hat{x} = F_{K_i}^{-1}(x), j = 1, 2, 3.$$

Também temos que

$$\nabla_x \phi_j = \nabla_{\hat{x}} \hat{\phi}_j(\hat{x}) B_{K_i}^{-1}$$

onde $B_{K_i} = \begin{bmatrix} v_1 - u_1 & z_1 - u_2 \\ v_2 - u_1 & z_2 - u_2 \end{bmatrix}$.

Observação 16. *De acordo com o Lema 15 dizemos que o método dos elementos finitos lineares por partes em duas dimensões tem três graus de liberdade por cada elemento.*

No triângulo de referência $\hat{K}$ vale a seguinte fórmula de quadratura.

Lema 17 (Quadratura de sete pontos no triângulo)**.** *Vale*

$$\int_{\hat{K}} f(\hat{x}) d\hat{x} \approx \sum_{i=1}^{7} f(\zeta_i) \omega_i$$

Ponto ζ_i	Peso ω_i
$(\frac{1}{3}, \frac{1}{3})$	$\frac{9}{90}$
$(\frac{6+\sqrt{15}}{21}, \frac{6+\sqrt{15}}{21})$	$\frac{155+\sqrt{15}}{2400}$
$(\frac{9-2\sqrt{15}}{21}, \frac{6+\sqrt{15}}{21})$	$\frac{155+\sqrt{15}}{2400}$
$(\frac{6+\sqrt{15}}{21}, \frac{9-2\sqrt{15}}{21})$	$\frac{155+\sqrt{15}}{2400}$
$(\frac{6-\sqrt{15}}{21}, \frac{6-\sqrt{15}}{21})$	$\frac{155-\sqrt{15}}{2400}$
$(\frac{9+2\sqrt{15}}{21}, \frac{6-\sqrt{15}}{21})$	$\frac{155-\sqrt{15}}{2400}$
$(\frac{6-\sqrt{15}}{21}, \frac{9+2\sqrt{15}}{21})$	$\frac{155-\sqrt{15}}{2400}$

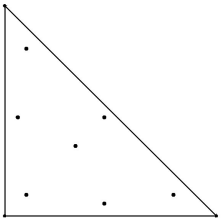


Tabela 3.1: Pontos e pesos da fórmula de quadratura com sete pontos no triângulo de referência (esquerda) e mapa dos pontos de quadratura no triângulo de referência (direita)

onde os pontos de quadratura e pesos ζ_i, ω_i são definidos na Tabela 3.1. A fórmula acima integra exato polinômios de grau total menor o igual que cinco.

Afim de aplicar a formula de quadratura do lema anterior para calcular os termos do lado direito ou da matriz A quando $\kappa(x)$ é um coeficiente complicado usamos a fórmula de mudança de variáveis.

Lema 18. *Seja K_i um elemento da triangulação $\mathcal{T}^h$ com vértices $u = (u_1, u_2), v = (v_1, v_2)$ e $z = (z_1, z_2)$ (ordenados no sentido anti-horário). Para toda função integrável $f : K_i = F_{K_i}(\hat{K}) \to \mathbb{R}$ temos*

$$\int_{K_i} f(x)dx = \int_{\hat{K}} f(F_{K_i}(\hat{x}))|det \begin{bmatrix} v_1 - u_1 & z_1 - u_2 \\ v_2 - u_1 & z_2 - u_2 \end{bmatrix}|d\hat{x}.$$

Análogo ao caso em uma dimensão a matriz A pode ser construída somando as contribuições locais de cada elemento. Note que em cada elemento triangular temos *três graus de liberdade*, isto é, somente três funções bases tem suporte em cada elemento. Consideramos somente o caso da forma bilinear definida em (3.16).

Lema 19. *Seja $\mathcal{A}$ a forma bilinear definida em (3.16) e $\mathcal{T}^h$ uma triangulação de D. Sejam K_i, $i = 1, \ldots, N_h^e$ os elementos da triangulação. Definamos R_i como a matriz $3 \times N_h^v$ de restrição ao elemento K_i, i.e., a matriz R_i^T tem todas as entradas nulas com excessão das posições $(1, i_1)$ (correspondente ao vértice $x_{i_1} \in K_i$), $(2, i_2)$ (do vértice $x_{i_2} \in K_i$) e $(3, i_3)$ (do vértice $x_{i_3} \in K_i$), onde tem o valor um. Seja A_{K_i} a matriz local definida por*

$$A_{K_i} = \begin{bmatrix} \mathcal{A}_{K_i}(\phi_{i_1}, \phi_{i_1}) & \mathcal{A}_{K_i}(\phi_{i_1}, \phi_{i_2}) & \mathcal{A}_{K_i}(\phi_{i_1}, \phi_{i_3}) \\ \mathcal{A}_{K_i}(\phi_{i_2}, \phi_{i_1}) & \mathcal{A}_{K_i}(\phi_{i_2}, \phi_{i_2}) & \mathcal{A}_{K_i}(\phi_{i_2}, \phi_{i_3}) \\ \mathcal{A}_{K_i}(\phi_{i_3}, \phi_{i_1}) & \mathcal{A}_{K_i}(\phi_{i_3}, \phi_{i_2}) & \mathcal{A}_{K_i}(\phi_{i_3}, \phi_{i_3}) \end{bmatrix} \tag{3.26}$$

onde $\mathcal{A}_{K_i}$, a restrição da forma bilinear $\mathcal{A}$ ao elemento K_i, é dada por

$$\mathcal{A}_{K_i}(v, w) = \int_{K_i} \kappa(x) \nabla v(x) \cdot \nabla w(x) dx. \tag{3.27}$$

Finalmente seja $\boldsymbol{b}_{K_i}$ o lado direito local,

$$\boldsymbol{b}_{K_i} = \begin{bmatrix} \mathcal{F}_{K_i}(\phi_{i_1}) \\ \mathcal{F}_{K_i}(\phi_{i_2}) \\ \mathcal{F}_{K_i}(\phi_{i_3}) \end{bmatrix}$$

onde $\mathcal{F}_{K_i}$, a restrição do $\mathcal{F}$ ao elemento K_i, é dada por

$$\mathcal{F}_{K_i}(v) = \int_{K_i} f(x) v(x) dx.$$

Temos então que

$$A = \sum_{i=1}^{N_h^e} R_i^T A_{K_i} R_i \tag{3.28}$$

e

$$\boldsymbol{b} = \sum_{i=1}^{N_h^e} R_i^T \boldsymbol{b}_{K_i}. \tag{3.29}$$

O lema anterior permite calcular a matriz A usando as contribuições locais de cada elemento. Isto representa uma grande vantagem na hora da implementação numérica. Na igualdade (3.28) o papel das matrizes R_i é somente colocar a contribuição local no lugar certo na

matriz *global* A. As matrizes de extensão R_i, $i = 1, \ldots, N_h^e$, podem ser substituídas por funções que transformem os índices locais $[1, 2, 3]$ no elemento K_i, nos índices globais $[i_1, i_2, i_3]$. Analogamente para as matrizes R_i^T, $i = 1, \ldots, N_h^e$. Note também que as matrizes locais são calculadas em cada elemento K_i separadamente.
Para resolver o sistema linear $Ax = b$ com a matriz A em (3.28) usando um método iterativo, somente precisamos de uma rotina que faça a operação de multiplicação matriz vezes vetor, isto é, dado $x \in \mathbb{R}^{N_h^v}$, calcule, Ax. O resultado de aplicar a matriz A ao vetor x pode ser calculado usando (3.28) diretamente. A formulas (3.28) e (3.29) são fundamentais para a construção de muitos algoritmos de decomposição de domínios na aproximação numérica de equações diferenciais parciais elípticas.

Exemplo: a equação de Laplace

Suponha que queremos aproximar a solução da equação de Laplace (3.7) com $D = [0, 1] \times [0, 1]$ e $f(x) = -1$ usando elementos finitos, isto é, queremos resolver

$$\begin{array}{l} \textit{Achar } u : [0,1] \times [0,1] \to \mathbb{R} \textit{ tal que:} \\ \begin{cases} -\Delta u(x) = -1, & x \in [0,1] \times [0,1] \\ \quad u(x) = 1, & x \in \partial([0,1] \times [0,1]). \end{cases} \end{array} \tag{3.30}$$

Formulação fraca

Usamos a formulação fraca construída na Seção 3.3.1. Lembramos a definição da forma bilinear $\mathcal{A}$ em (3.10) e do funcional linear $\mathcal{F}$ definido em (3.11).

Triangulação

Neste exemplo usamos a triangulação da Figura 3.5. Observe que esta triangulação tem 12 vértices, 23 arestas e 12 elementos (triângulos). O primeiro elemento pode ser denotado por $K_1 = [1, 2, 6]$ indicando que seus vértices são x_1, x_2 e x_6, nessa ordem (sentido anti-horário).

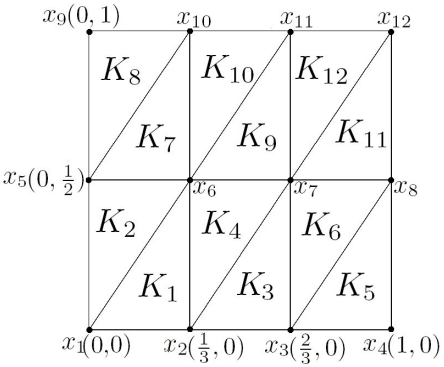


Figura 3.5: Triangulação usada para aproximar a solução de (3.30)

Os doze elementos são (veja Figura 3.5)

$$K_1 = [1,2,6], \quad K_2 = [6,5,1], \quad K_3 = [2,3,7],$$
$$K_4 = [7,6,1], \quad K_5 = [3,4,8], \quad K_6 = [8,7,3],$$
$$K_7 = [5,6,10], \quad K_8 = [10,9,5], \quad K_9 = [6,7,11],$$
$$K_{10} = [11,10,6], \quad K_{11} = [7,8,12], \quad K_{12} = [12,11,7].$$

Montagem da matriz

Depois de definir a triangulação procedemos a construção da matriz A usando a fórmula (3.28). Temos que construir as matrizes locais em (3.26). O primeiro elemento é K_1 com vértices $x_1 = (0,0)$, $x_2 = (1/3,0)$ e $x_6 = (1/3,1/2)$. Vide Figura 3.5. Para este elemento temos que a matriz B_{K_1} do Lema 15 é $B_{K_1} = \begin{bmatrix} 1/3 & 1/3 \\ 0 & 1/2 \end{bmatrix}$ e a sua inversa é dada por $B_{K_1}^{-1} = \begin{bmatrix} 3 & -2 \\ 0 & 2 \end{bmatrix}$. Para as funções em K_1 temos então que

$$\nabla_x \phi_1 = \nabla_{\hat{x}} \hat{\phi_1} B_{K_1}^{-1} = \begin{bmatrix} -1 & -1 \end{bmatrix} B_{K_1}^{-1} = \begin{bmatrix} -3 & 0 \end{bmatrix},$$

e analogamente em K_1 temos $\nabla\phi_2 = \begin{bmatrix} 3 & -2 \end{bmatrix}$ e $\nabla\phi_6 = \begin{bmatrix} 0 & 2 \end{bmatrix}$. As entradas da matriz local A_{K_1} são calculadas como segue,

$$\begin{aligned}
\mathcal{A}_{K_1}(\phi_1,\phi_1) &= \int_{K_1} \nabla\phi_1\cdot\nabla\phi_1 = \int_{K_1} (-3)^2+(0)^2 = 9|K_1| = \frac{9}{12},\\
\mathcal{A}_{K_1}(\phi_1,\phi_2) &= \mathcal{A}_{K_1}(\phi_2,\phi_1) = \int_{K_1} \nabla\phi_1\cdot\nabla\phi_2 = \int_{K_1} -9 = \frac{-9}{12},\\
\mathcal{A}_{K_1}(\phi_1,\phi_6) &= \mathcal{A}_{K_1}(\phi_6,\phi_1) = \int_{K_1} \nabla\phi_1\cdot\nabla\phi_6 = \int_{K_1} 0 = 0,\\
\mathcal{A}_{K_1}(\phi_2,\phi_2) &= \int_{K_1} \nabla\phi_2\cdot\nabla\phi_2 = \int_{K_1} 13 = \frac{13}{12},\\
\mathcal{A}_{K_1}(\phi_2,\phi_6) &= \mathcal{A}_{K_1}(\phi_6,\phi_2) = \int_{K_1} -4 = \frac{-4}{12},\\
\mathcal{A}_{K_1}(\phi_6,\phi_6) &= \int_{K_1} 4 = \frac{4}{12}.
\end{aligned}$$

Portanto a matriz local é $A_{K_1} = \frac{1}{12}\begin{bmatrix} 9 & -9 & 0 \\ -9 & 13 & -4 \\ 0 & -4 & 4 \end{bmatrix}$. Note que R_1 é a matriz 3×12,

$$R_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \dots & 0 \end{bmatrix}$$

e $R_1^T A_{K_1} R_1$ é a matriz 12×12 tal que

$$R_1^T A_{K_1} R_1 = \frac{1}{12}\begin{bmatrix} 9 & -9 & 0 & 0 & 0 & 0 & 0 & \dots \\ -9 & 13 & 0 & 0 & 0 & -4 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & -4 & 0 & 0 & 0 & 4 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \end{bmatrix}.$$

Analogamente podem ser calculadas as matrizes locais e de restrição para os outros elementos K_i, $i = 2, \ldots, 12$. Obtemos

$$A_{K_i} = \frac{1}{12}\begin{bmatrix} 9 & -9 & 0 \\ -9 & 13 & -4 \\ 0 & -4 & 4 \end{bmatrix}, \quad i = 2, \ldots, 12.$$

Com as matrizes locais A_{K_i} e as matrizes R_i, $i = 1, \ldots, 12$, montamos a matriz global

$$A = \sum_{i=1}^{12} R_i^T A_{K_i} R_i.$$

Por exemplo, se queremos calcular a entrada a_{66} da matriz A notamos que x_6 é o terceiro vértice de K_1, o primeiro vértice de K_2, o segundo vértice de K_4,..., donde

$$\begin{aligned} a_{6,6} &= \mathcal{A}(\phi_6, \phi_6) = \sum_{i=1,4,9,10,7,2} \mathcal{A}_{K_i}(\phi_6, \phi_6) \\ &= \frac{4}{12} + \frac{13}{12} + \frac{9}{12} + \frac{4}{12} + \frac{13}{12} + \frac{9}{12} = \frac{52}{12}. \end{aligned}$$

onde os índices na soma acima são os elementos da triangulação que tem vértice 6 (x_6).

Observação 20. *No caso geral, as integrais no triângulo de referência no calculo das matrizes locais acima podem ser calculadas usando uma fórmula de quadratura. Veja Lema 17. O mesmo vale para os calculo do lado direito local.*

Montagem do lado direito

Agora montamos o lado direito $\boldsymbol{b}$ usando (3.29). Como antes vamos elemento por elemento para calcular as contribuições locais. Neste exemplo a função do lado direito é $g(x) = -1$ para todo $x \in D$. No elemento K_1, usando a fórmula de mudança de variáveis, temos

$$\begin{aligned} \mathcal{F}_{K_1}(\phi_1) &= \int_{K_1} -1\phi_1(x)d\boldsymbol{x} = -\int_{\hat{K}} \hat{\phi_1}(\hat{x})|\det B_{K_1}|d\hat{\boldsymbol{x}} = -\frac{1}{36} \\ \mathcal{F}_{K_1}(\phi_2) &= \int_{K_1} -1\phi_2(x)d\boldsymbol{x} = -\int_{\hat{K}} \hat{\phi_2}(\hat{x})|\det B_{K_1}|d\hat{\boldsymbol{x}} = -\frac{1}{36} \\ \mathcal{F}_{K_1}(\phi_3) &= \int_{K_1} -1\phi_3(x)d\boldsymbol{x} = -\int_{\hat{K}} \hat{\phi_3}(\hat{x})|\det B_{K_1}|d\hat{\boldsymbol{x}} = -\frac{1}{36} \end{aligned}$$

e então $\boldsymbol{b}_{K_1} = -(\frac{1}{36}, \frac{1}{36}, \frac{1}{36})^T$. Analogamente $\boldsymbol{b}_{K_i} = -(\frac{1}{36}, \frac{1}{36}, \frac{1}{36})^T$, $i = 2, \ldots, 12$. Temos finalmente que

$$\boldsymbol{b} = \sum_{i=1}^{12} R_i^T \boldsymbol{b}_{K_i}.$$

Por exemplo a sexta coordenada de $\boldsymbol{b}$ é

$$f_6 = \sum_{i=1,4,9,10,7,2} \mathcal{F}_{K_i}(\phi_6) = -\sum_{i=1,4,9,10,7,2} \frac{1}{36} = -\frac{1}{6}.$$

onde os índices na soma acima são os elementos da triangulação que tem vértice 6 (x_6).

Solução do sistema linear

Temos que resolver os sistema linear

$$A\boldsymbol{u}_h = \boldsymbol{b}.$$

Sabemos que $u_h(x_i) = 1$ para $i = 1, 2, 3, 4, 5, 8, 9, 10, 11$. Podemos substituir

$$\boldsymbol{u}_h = \boldsymbol{u}_h^D + \begin{bmatrix} \vdots \\ 0 \\ u_h(x_6) \\ u_h(x_7) \\ 0 \\ \vdots \end{bmatrix}$$

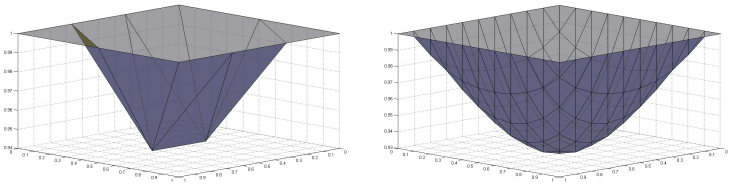
e colocando o termo conhecido $A\boldsymbol{u}_h^D$ no lado direito, obtemos

$$\begin{aligned} \frac{1}{12}\begin{bmatrix} 52 & -18 \\ -18 & 52 \end{bmatrix}\begin{bmatrix} u_h(x_6) \\ u_h(x_7) \end{bmatrix} &= -\begin{bmatrix} 1/6 \\ 1/6 \end{bmatrix} + \begin{bmatrix} 16/6 \\ 16/6 \end{bmatrix} \\ &= \begin{bmatrix} 8/3 \\ 8/3 \end{bmatrix} \end{aligned}$$

que fornece a solução $\begin{bmatrix} u_h(x_6) \\ u_h(x_7) \end{bmatrix} = \begin{bmatrix} 16/17 \\ 16/17 \end{bmatrix}$. Pedimos ao leitor que tente construir uma solução analítica para o problema (3.30). O

objetivo deste exemplo é mostrar o potencial do método dos elementos finitos. Na Figura 3.6 a) mostramos a aproximação de elementos finitos obtida acima. Lembre que neste exemplo o tamanho do sistema linear resolvido é 2×2.

Para finalizar esta seção calculamos a solução de elementos finitos da equação (3.30) numa malha mais fina com $h = 1/10$. Na figura 3.6 b) vemos a aproximação por elementos finitos para esta malha mais fina. O sistema linear resolvido para obter a aproximação de elementos finitos com $h = 1/10$ é de tamanho 81×81.



a) b)

Figura 3.6: Aproximações de elementos finitos da solução da equação (3.30): a) com $h = 1/2$, 12 vértices e 12 elementos; b) com $h = 1/10$, 121 vértices e 200 elementos.

3.4.2 O sistema linear obtido

Como no caso de uma dimensão espacial, é necessário escolher o método adequado para aproximar a solução do sistema linear equivalente à formulação de Galerkin. No caso da equação elíptica básica (3.13) e o espaço de elementos finitos de funções lineares por partes, a matriz resultante A tem as seguintes propriedades: O tamanho da matriz é gigantesco. A matriz A é esparsa, menos esparsa que em uma dimensão, mas ainda a porcentagem de entradas não nulas é muito grande. A matriz de Dirichlet é definida positiva. Para a matriz da forma bilinear $\mathcal{A}$ definida em (3.16) com o coeficiente κ satisfazendo (3.14) temos que existem constantes positivas C e c, que dependem unicamente da triangulação $\mathcal{T}^h$ e do domínio D, tais que

$$\lambda_{\max} \leq C\kappa_{\max} \quad \text{e} \quad \lambda_{\min} \geq c\kappa_{\min}h^2.$$

Onde $\lambda_{\max}$ e $\lambda_{\min}$ são o maior e o menor autovalor da matriz A. Podemos assim estimar o número de condição (espectral) da matriz A por

$$\mathrm{Cond}(A) := \frac{\lambda_{\max}}{\lambda_{\min}} \leq \frac{C}{c}\frac{\kappa_{\max}}{\kappa_{\min}}\frac{1}{h^2}. \tag{3.31}$$

3.4.3 Erro de aproximação

Dados $h > 0$ e uma triangulação do domínio D pode-se então usar $V = \mathbb{P}^1(\mathcal{T}^h)$ em (3.18). A solução obtida em (3.18) é a aproximação de elementos finitos $\mathbb{P}^1(\mathcal{T}^h)$ da solução de (3.17). Denotaremos por u^h esta solução. Temos

Lema 21 (Estimativa de erro 'a priori')**.** *Sejam u e u^h as soluções da formulação fraca (3.15) e da sua formulação de Galerkin com $V^h = \mathbb{P}^1(\mathcal{T}^h)$, respectivamente. Temos que existe uma constante C que é independente de $h > 0$ e de u, tal que*

$$|u - u^h|_{H^1(D)} \leq Ch|u|_{H^2(D)} \quad e \quad \|u - u^h\|_{L^2(D)} \leq Ch^2|u|_{H^2(D)}.$$

Capítulo 4

O método do gradiente conjugado

O material apresentado neste capítulo corresponde ao capítulo do gradiente conjugado num curso de análise numérico no IMPA. Parte do material aqui apresentado coincide com [29]. Veja também [18, 28, 4].

4.1 O método do gradiente conjugado

Neste capítulo A denota uma matriz auto-adjunta e definida posi-tiva de dimensão $n \times n$. A norma gerada pela matriz A é dada por

$$\|x\|_A = (x, Ax) = (A^{1/2}x, A^{1/2}x) = \|A^{1/2}x\|_2$$

onde $\|\cdot\|_2$ é a norma Euclidiana do $\mathbb{R}^n$. Dois vetores $x, y \in \mathbb{R}^n$ são ditos conjugados com relação a A, ou A-ortogonais se eles são ortogonais no produto interno gerado pela matriz A, isto é,

$$(x, Ay) = x^T Ay = 0.$$

O conjunto $\{d_i\}_{i=1}^n$ é dito conjugado se d_i é conjugado a d_j para todo $i \neq j$. Neste caso temos $d_i^T A d_j = 0$ para todo $i \neq j$.

Suponha que $\{d_i\}_{i=1}^n$ são direções conjugadas em $\mathbb{R}^n$, temos que estas direções são linearmente independentes, logo $\text{span}\{d_i\}_{i=1}^n = \mathbb{R}^n$. Seja x_* tal que

$$Ax_* = b.$$

Temos que existem $\alpha_1, \ldots, \alpha_n$ tais que das direções conjugadas,

$$x_* = \sum_{i=1}^{n} \alpha_i d_i. \tag{4.1}$$

Podemos então calcular x_* se conseguimos especificar direções conjugadas $\{d_i\}_{i=1}^n$ e os respectivos coeficientes $\{\alpha_j\}_{i=1}^n$. Se aplicamos a matriz A na equação (4.1), obtemos que $\sum_{i=1}^n \alpha_i A d_i = b$ e tomando produto interno com d_j para j fixo, mas arbitrário, obtemos

$$\sum_{i=1}^{n} \alpha_i d_j^T A d_i = d_j^T b$$

donde, usando o fato $d_j^T A d_i = 0$ para $i \neq j$, obtemos

$$\alpha_j = \frac{d_j^T b}{d_j^T A d_j}, \quad j = 1, \ldots, n.$$

Para $j = 1, \ldots, n$, defina a j-ésima aproximação de x_* por

$$x_j = \sum_{i=1}^{j} \alpha_i d_i. \tag{4.2}$$

Como $Ax_* = b$ temos

$$0 = Ax_* - b = Ax_{j-1} - b + \sum_{i=j}^{n} \alpha_i A d_i,$$

e tomando novamente produto com d_j^T, temos

$$\alpha_j = -\frac{d_j^T (Ax_{j-1} - b)}{d_j^T A d_j} = \frac{d_j^T q_{j-1}}{d_j^T A d_j},$$

onde o $j-$ésimo resíduo é

$$q_j = -(Ax_j - b) = b - Ax_j, \quad j = 1, \ldots, n.$$

Observe que, dado x_k, podemos calcular

$$x_{k+1} = x_k + \alpha_{k+1} d_{k+1} \quad \text{onde } \alpha_{k+1} = \frac{d_{k+1}^T q_k}{d_{k+1}^T A d_{k+1}}. \tag{4.3}$$

Temos o seguinte lema.

Lema 22. *Suponha que $x_0 = 0$. Se $i \leq k$ então $q_k^T d_i = 0$.*

Demonstração. De (4.1) obtemos

$$q_k = b - Ax_k = A(x_* - x_k) = A\left[\sum_{j=k+1}^{n} \alpha_j d_j\right] = \sum_{j=k+1}^{n} \alpha_j A d_j$$

onde, para $i \leq k$, $d_i^T q_k = \sum_{j=k+1}^{n} \alpha_j d_i^T A d_j = 0$. ∎

Corolário 23. *Se temos as k primeiras direções conjugadas $\{d_i\}_{i=1}^k$, e $q_k = b - Ax_k \neq 0$ então $\{d_1, d_2, ..., d_k, q_k\}$ é linearmente independente.*

Suponha que temos $k+1$ direções A-ortogonais $d_1, d_2, ..., d_k$. Para obter outra direção conjugada a partir de q_k, aplicamos Gram-Schmidt no produto interno A, isto é, geramos d_{k+1} com

$$d_{k+1} = q_k - \sum_{j=1}^{k} \bar{a}_{jk} d_j, \tag{4.4}$$

e $\text{span}\{d_1, \ldots, d_k, q_k\} = \text{span}\{d_1, \ldots, d_k, d_{k+1}\}$. Provaremos na frente que somente o último coeficiente $\bar{a}_{kk}$ em (4.4) acima é diferente de zero. Tomando produto interno com Ad_i, $i = 1, ..., k$, obtemos

$$\begin{aligned} 0 &= d_i^T A d_{k+1} = d_i^T A q_k - \sum_{j=1}^{k} \bar{a}_{jk} d_i^T A d_j \\ &= d_i^T A q_k - \bar{a}_{ik} d_i^T A d_i. \end{aligned}$$

Donde para cada coeficiente em (4.4) temos $\bar{a}_{ik} = \dfrac{d_i^T A q_k}{d_i^T A d_i}$.

Teorema 24. *Se $d_1 = b$, e $d_i \neq 0$, $i = 1, \ldots, k$, temos então $\bar{a}_{1k} = \bar{a}_{2k} = ... = \bar{a}_{(k-1)k} = 0$.*

Para provar o teorema anterior precisamos de um lema simples.

Lema 25. *Suponha que $x_0 = 0$, $q_1 = b$, d_1 é um múltiplo escalar de b e $d_i \neq 0$, $i = 1, \ldots, k$. Defina*

$$V_k := span\{d_1, ..., d_k\} = span\{d_i\}_{i=1}^k.$$

Temos que

1. $V_k = span\{b, Ab, ..., A^{k-1}b\}$ *(subespaço de Krylov!)*
2. $V_k = span\{q_1, \ldots, q_k\}$
3. $AV_k \subseteq V_{k+1}$
4. $q_k^T A V_{k-1} = \{0\}$, $q_k^T V_{k-1} = \{0\}$ *e* $q_k^T q_i = 0$ *para* $i \leq k-1$.

Demonstração. Para provar *1.* usamos indução. Sabemos que $V_1 = \text{span}\{d_1\} = \text{span}\{b\}$. Suponha que o lema vale para o inteiro k. Para $x_k \in V_k$ temos que $q_k = b - Ax_k \in \text{span}\{AV_k, b\} = \text{span}\{b, Ab, \ldots, A^{k-1}b, A^k b\}$. Portanto, de (4.4) e a hipótese de indução vemos que

$$d_{k+1} \in \text{span}\{b, Ab, \ldots, A^{k-1}b, A^k b\}.$$

Isto da $\text{span}\{d_i\}_{i=1}^{k+1} \subseteq \text{span}\{A^i b\}_{i=0}^k$ e como os vetores $\{d_i\}_{i=1}^{k+1}$ são linearmente independentes eles devem gerar um espaço de dimensão $k+1$. Isto prova o primeiro enunciado. A prova do *2.* segue do processo de ortogonalização em (4.4). A prova de *3.* seque de *1.* e a prova de *4.* segue do *2.* e o Lema 22. ∎

Agora provamos o Teorema 24.

Demonstração. Se $i \leq k-1$ usando (4.4) temos

$$\begin{aligned} d_i^T A d_{k+1} &= d_i^T A q_k - \sum_{j=1}^{k} \bar{a}_{jk} d_i^T A d_j \\ &= d_i^T A q_k - \bar{a}_{ik} d_i^T A d_j. \end{aligned}$$

Note que $d_i^T A d_{k+1} = 0$, por *3.* do Lema 25 temos que $d_i^T A q_k = 0$, e $d_i^T A d_j \neq 0$, o que implica que $a_{ik} = 0$. ■

Temos então a seguintes formulas para a direção conjugada $k+1$ e o seu coeficiente α_{k+1} em (4.1) e (4.3),

$$d_{k+1} = q_k - a_{kk} d_k, \quad \bar{a}_{kk} = \frac{d_k^T A q_k}{d_k^T A d_k}, \quad \alpha_{k+1} = \frac{d_{k+1}^T q_k}{d_{k+1}^T A d_{k+1}}.$$

Vamos simplificar um pouco mais estas formulas. Note que da primeira fórmula acima e o Lema 22 temos

$$\alpha_{k+1} = \frac{(q_k - a_{kk} d_k) q_k}{d_{k+1}^T A d_{k+1}} = \frac{q_k^T q_k}{d_{k+1}^T A d_{k+1}},$$

e na iteração k teríamos $\alpha_k d_k^T A d_k = q_{k-1}^T q_{k-1}$. Do fato que podemos escrever $q_k = q_{k-1} - \alpha_k A d_k$ e *4.* no Lema 25 temos que

$$q_k^T q_k = q_k^T (q_{k-1} - \alpha_k A d_k) = -\alpha_k q_k^T A d_k.$$

Juntando estas duas últimas igualdades temos

$$\beta_k := -\bar{a}_{kk} = -\frac{d_k^T A q_k}{d_k^T A d_k} = -\frac{-\frac{1}{\alpha_k} q_k^T q_k}{\frac{1}{\alpha_k} q_{k-1}^T q_{k-1}} = \frac{q_k^T q_k}{q_{k-1}^T q_{k-1}}.$$

Usando estas formulas finais descrevemos o algoritmo na Tabela 4.1.

Agora estudamos a convergência do método do gradiente conjugado. Queremos saber a velocidade de convergência do método. Embora o gradiente conjugado tome somente n iterações para achar a solução do sistema linear, n poder ser muito grande e pode ser que o nosso tempo computacional não seja suficiente para chegar ate a última iteração.

1. Inicializar $q_0 = b - Ax_0$
2. Iterar $k = 1, 2, \ldots,$ ate a convergência

$$\begin{aligned}
\beta_k &= \frac{(q_{k-1}, q_{k-1})}{(q_{k-2}, q_{k-2})} \quad [\beta_1 = 0] \\
d_k &= q_{k-1} + \beta_k d_{k-1} \quad [d_1 = q_0] \\
\alpha_k &= \frac{(q_{k-1}, q_{k-1})}{(d_k, Ad_k)} \\
x_k &= x_{k-1} + \alpha_k d_k \\
q_k &= q_{k-1} - \alpha_k A d_k
\end{aligned}$$

Tabela 4.1: Algoritmo do gradiente conjugado

Note que se $x_k := \sum_{i=1}^{k} \alpha_i d_i$, com $\{d_i\}_{i=1}^{n}$ direções conjugadas e $x_* = \sum_{i=1}^{n} \alpha_i d_i$, então x_k é projeção ortogonal (no produto interno gerado pela matriz A) do vetor x_* no espaço $V_k = \text{span}\{d_i\}_{i=1}^{k}$. Das propriedades gerais das projeções ortogonais, temos

$$(d_i, A(x_* - x_k)) \quad = \quad 0 \quad i = 1, ..., k,$$

e

$$\min_{y \in V_k} ||x_* - y||_A \quad = \quad ||x_* - x_k||_A. \tag{4.5}$$

Denote por $\mathbb{P}_{k-1}$ o conjunto de polinômios de grau menor o igual que $k-1$. Dado $y \in V_k$, usando o Lema 25 pode-se exprimir

$$y = \sum_{i=0}^{k-1} \gamma_i A^{i-1} b.$$

Se P é o polinômio de grau $k-1$ definido por $P(x) = \sum_{i=0}^{k-1} \gamma_i x^i$ podemos escrever $y = P(A)b$. Concluímos que

$$\begin{aligned}
V_k &= \{P(A)b \quad : \quad P \in \mathbb{P}_{k-1}\} \\
&= \{P(A)Ax_* \quad : \quad P \in \mathbb{P}_{k-1}\}.
\end{aligned}$$

Com esta igualdade, (4.5) e o fato $\|y\|_A = \|A^{1/2}y\|_2$ para todo $y \in \mathbb{R}^n$ temos

$$\begin{aligned} ||x_* - x_k||_A &= \min_{y \in V_k} ||x_* - y||_A = \min_{P \in \mathbb{P}_{k-1}} ||x_* - P(A)Ax_*||_A \\ &= \min_{P \in \mathbb{P}_{k-1}} ||A^{1/2}\left[I - P(A)A\right]x_*||_2 \end{aligned} \tag{4.6}$$

Considere a decomposição espectral da matriz auto-adjunta e positiva definida $A = Q^*\Lambda Q$, onde Q é uma matriz ortonormal (tem colunas ortonormais) e $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ é a matriz diagonal com os autovetores de A ordenados do maior ao menor, isto é, $\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n = \lambda_{\min}$. Temos que a decomposição espectral $I - P(A)A = Q^*\left[I - P(\Lambda)\Lambda\right]Q$ e usando propriedades das matrizes ortonormais temos

$$\begin{aligned} ||A^{1/2}\left[I - P(A)A\right]x_*||_2 &= ||\left[I - P(A)A\right]A^{1/2}x_*||_2 \\ &\leq ||I - P(A)A||_2||A^{1/2}x_*||_2 \\ &\leq ||Q^*\left[I - P(\Lambda)\Lambda\right]Q||_2||x_*||_A \\ &= ||I - P(\Lambda)\Lambda||_2||x_*||_A \end{aligned}$$

onde temos usado a desigualdade $||Bz||_2 \leq \|B\|_2\|z\|_2$ valida para toda matriz B. Lembrando que para B auto-adjunta e definida positiva $\|B\|_2$ é o máximo autovalor de B, temos,

$$\begin{aligned} ||A^{1/2}\left[I - P(A)A\right]x_*||_2 &= ||I - P(\Lambda)\Lambda||_2||x_*||_A \\ &\leq \max_{1 \leq i \leq n} |1 - P(\lambda_i)\lambda_i|||x_*||_A \\ &\leq \max_{\lambda \in [\lambda_n, \lambda_1]} |1 - P(\lambda)\lambda|||x_*||_A, \end{aligned} \tag{4.7}$$

onde na última linha o máximo é tomado no intervalo $[\lambda_n, \lambda_1]$ e não somente nos n autovalores $\lambda_n, \ldots, \lambda_1$. Finalmente pondo (4.7) em (4.6) obtemos

$$\begin{aligned} ||x_* - x_k||_A &\leq ||x_*||_A \min_{P \in \mathbb{P}_{k-1}} \max_{\lambda \in [\lambda_n, \lambda_1]} |1 - \lambda P(\lambda)| \\ &= ||x_*||_A \min_{P \in \mathbb{P}_k, P(0)=1} \max_{\lambda \in [\lambda_n, \lambda_1]} |P(\lambda)| \end{aligned} \tag{4.8}$$

onde para obter a última igualdade notamos que se P é de grau $k-1$ então o polinômio definido por $1 - \lambda P(\lambda)$ é de grau k e tem o valor

1 quando $\lambda = 0$.

Para entender melhor o problema de minimização (4.8) estudamos polinômios ortogonais de Chebyshev(Tchebyshev). Uma das muitas definições dos Polinômio de Chebyshev é a seguinte. Define-se o polinômio de Chebyshev de grau k por [1]

$$T_k(x) = \begin{cases} \cos(k\cos^{-1}(x)), & \text{se } |x| \leq 1, \\ \\ \cosh(k\cosh^{-1}(x)), & \text{se } |x| > 1. \end{cases} \tag{4.9}$$

Note que $T_k(x) = \cos(k\theta)$ onde $\cos(\theta) = x$ com $\theta \in [-\pi, 0]$ e portanto T_k define um polinômio de grau k na variável x já que $cos(k\theta)$ é um polinômio em $\cos(\theta)$. Temos

$$T_k(1) = 1, \quad |T_k(x)| \leq 1, \quad \text{para todo } x \in [-1, 1].$$

$$\int_{-1}^{1} T_i(x)T_j(x)(1-x^2)^{-\frac{1}{2}}dx = \delta_{ij},$$

onde δ_{ij} é o delta de Kronecker. Os polinômios $T_n(x)$ podem também ser obtidos a partir da recorrência:

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x) \tag{4.10}$$

que pode ser deduzida facilmente de (4.9).

Na equação (4.8) podemos tomar $P = \overline{T}_k$ definido a partir do polinômio de Chebyshev de grau k, T_k, por

$$\overline{T}_k(\lambda) = \frac{T_k(g(\lambda))}{T_k(g(0))} \quad \text{com} \quad g(\lambda) = \frac{\lambda_1 + \lambda_n - 2\lambda}{\lambda_1 - \lambda_n}.$$

[1] Para ver que as duas partes da definição dão os mesmos polinômios podemos verificar a fórmula de recorrência (4.10) para cada uma. Uma outra forma de ver este fato é que passando a variável complexa temos

$$\cosh(k\cosh^{-1}(\zeta)) = \cos(-ik\cosh^{-1}(\zeta)) = \cos(k\cos^{-1}(\zeta))$$

para todo ζ no domínio das duas funções no plano complexo.

Observe que g leva o intervalo $[\lambda_n, \lambda_1]$ no intervalo $[-1, 1]$ com $g(\lambda_n) = 1$, $g(\lambda_1) = -1$. Os zeros do polinômio T_k ficam localizados no intervalo $[-1, 1]$. Temos então que no caso $\lambda_1 \neq \lambda_n$

$$g(0) = \frac{\lambda_1 + \lambda_n}{\lambda_1 - \lambda_n} > 1, \quad \text{donde} \quad T_k(g(0)) \neq 0,$$

vemos que $\overline{T}_k(0) = 1$. Os zeros do polinômio $\overline{T}_k$ ficam localizados no intervalo $[\lambda_n, \lambda_1]$. De (4.8) vemos que

$$||x_* - x_k||_A \quad \leq \quad \max_{[\lambda_n, \lambda_1]} |\overline{T}_k(\lambda)| ||x_*||_A \leq \frac{1}{T_k(g(0))} ||x_*||_A$$

com

$$g(0) \quad = \quad \frac{\lambda_1 + \lambda_n}{\lambda_1 - \lambda_n} = \frac{\text{Cond}(A) + 1}{\text{Cond}(A) - 1},$$

e onde o *número de condição*[2] de A é definido por

$$\text{Cond}(A) = \frac{\lambda_1}{\lambda_n} = \frac{\lambda_{\max}}{\lambda_{\min}}. \tag{4.11}$$

Para continuar com o argumento usaremos o seguinte resultado.

Lema 26. *Para todo x com $|x| \geq 1$ existe z tal que*

$$x = \frac{z + z^{-1}}{2}, \quad e \quad T_k(x) = \frac{z^k + z^{-k}}{2}.$$

Note que $\frac{\sqrt{\text{Cond}(A)}+1}{\sqrt{\text{Cond}(A)}-1} \geq 1$, e que podemos escrever

$$\frac{\text{Cond}(A) + 1}{\text{Cond}(A) - 1} = \frac{1}{2} \left[\frac{\sqrt{\text{Cond}(A)} - 1}{\sqrt{\text{Cond}(A)} + 1} + \frac{1}{\frac{\sqrt{\text{Cond}(A)}-1}{\sqrt{\text{Cond}(A)}+1}} \right]$$

[2] ou número de condição espectral

Aplicando o Lema 26 com $z = \frac{\sqrt{\text{Cond}(A)}-1}{\sqrt{\text{Cond}(A)}+1}$ temos que

$$\begin{aligned} T_k(g(0)) &= T_k\left[\frac{\text{Cond}(A)+1}{\text{Cond}(A)-1}\right] \\ &= \frac{1}{2}\left[\left[\frac{\sqrt{\text{Cond}(A)}-1}{\sqrt{\text{Cond}(A)}+1}\right]^k + \left[\frac{\sqrt{\text{Cond}(A)}+1}{\sqrt{\text{Cond}(A)}-1}\right]^k\right] \end{aligned}$$

logo,

$$\begin{aligned} \frac{1}{T_k(g(0))} &= \frac{2}{\left[\frac{\sqrt{\text{Cond}(A)}-1}{\sqrt{\text{Cond}(A)}+1}\right]^k + \left[\frac{\sqrt{\text{Cond}(A)}+1}{\sqrt{\text{Cond}(A)}-1}\right]^k} \\ &\leq \frac{2}{\left[\frac{\sqrt{\text{Cond}(A)}+1}{\sqrt{\text{Cond}(A)}-1}\right]^k} = 2\left[\frac{\sqrt{\text{Cond}(A)}+1}{\sqrt{\text{Cond}(A)}-1}\right]^k. \end{aligned}$$

Fica provado então o seguinte Teorema.

Teorema 27. *Seja x_* tal que $Ax_* = b$ e x_k o $k-$ésimo iterado do gradiente conjugado. Então*

$$||x_* - x_k||_A \leq 2\left[\frac{\sqrt{Cond(A)}-1}{\sqrt{Cond(A)}+1}\right]^k ||x_*||_A.$$

Corolário 28. *Para o caso em que $x_0 \neq 0$, aplicamos o argumento anterior a*

$$A\delta x = b - Ax_0, \quad (\delta x)_0 = 0$$

e obtemos

$$||x_* - x_k||_A \leq 2\left[\frac{\sqrt{Cond(A)}-1}{\sqrt{Cond(A)}+1}\right]^k ||x_* - x_0||_A$$

para todo k.

Para fechar completamente a prova do teorema anterior temos que provar o Lema 26.

Demonstração. Se $x \geq 1$, pode-se exprimir

$$x \quad = \quad \cosh(\zeta) = \frac{e^{\zeta} + e^{-\zeta}}{2} = \frac{z + z^{-1}}{2}, \qquad \text{com } z = e^{\zeta},$$

donde, usando a definição do T_k em (4.9).

$$T_k(x) \quad = \quad \cosh(k \cosh^{-1} \zeta) = \frac{e^{k\zeta} + e^{-k\zeta}}{2} = \frac{z^k + z^{-k}}{2}.$$

■

Como resultado do Teorema 27 vemos que o erro do gradiente conjugado na norma da energia depende do número de condição da matriz $\text{Cond}(A) = \lambda_{\max}/\lambda_{\max}$.

Em geral, dada uma matriz invertível B, o número de condição na norma $\|\cdot\|_2$ é definido por

$$\text{Cond}(B) = \|B\|_2 \|B^{-1}\|_2.$$

Pode-se provar que quando A é auto-adjunta e definida positiva temos que $\text{Cond}(A) = \lambda_1/\lambda_n$.

4.2 Contagem de número de iterações

No Teorema 27 provamos que o erro inicial e reduzido pelo fator $2\left[\frac{\sqrt{\text{Cond}(A)}-1}{\sqrt{\text{Cond}(A)}+1}\right]^k$, isto é,

$$||x_* - x_k||_A \leq 2\left[\frac{\sqrt{\text{Cond}(A)} - 1}{\sqrt{\text{Cond}(A)} + 1}\right]^k ||x_* - x_0||_A.$$

Se queremos calcular suficientes iterações para reduzir o erro inicial por um fator ϵ, é suficiente tomar k tal que,

$$2\left[\frac{\sqrt{\text{Cond}(A)} - 1}{\sqrt{\text{Cond}(A)} + 1}\right]^k \leq \epsilon$$

donde, utilizando o fato $\ln(\frac{1+x}{1-x}) \geq 2x$ com $x = 1/\sqrt{\text{Cond}(A)}$ concluímos que k é da ordem

$$k = \frac{\ln(\frac{2}{\epsilon})}{\ln\left[\frac{1+\frac{1}{\sqrt{\text{Cond}(A)}}}{1-\frac{1}{\sqrt{\text{Cond}(A)}}}\right]} \leq \frac{1}{2}\sqrt{\text{Cond}(A)}\ln\left[\frac{2}{\epsilon}\right]. \tag{4.12}$$

Note que a desigualdade é assintoticamente uma igualdade quando $\text{Cond}(A) \to \infty$. Observamos então que o número de iterações necessárias para obter uma aproximação da solução com uma tolerância ϵ depende fortemente da condição da matriz $\text{Cond}(A)$.

Por exemplo, para a matriz da forma bilinear $\mathcal{A}$ definida em (2.13) com o coeficiente κ satisfazendo (2.11), temos a estimativa (2.34) para o número de condição, isto da que existe uma constante C independente de h tal que

$$k \leq C\frac{1}{2}h\ln\left[\frac{2}{\epsilon}\right].$$

A constante C pode depender do contraste do coeficiente $\kappa_{\max}/\kappa_{\min}$. Vemos que o número de iterações necessárias para obter a solução de elementos finitos (com uma tolerância ϵ) cresce de forma linear com o parâmetro da triangulação. Lembramos que o parâmetro h controla o tamanho do erro de elementos finitos. Geralmente, para problemas práticos isto representa demasiadas iterações (isto é, demasiado tempo computacional).

4.3 Experimentos numéricos

Nesta seção curta mostramos alguns resultados numéricos obtidos usando o método do gradiente conjugado na solução do sistema linear associado a discretização de elementos finitos da equação elíptica básica em (2.10) e (3.13).

Uma dimensão

Considere primeiro a equação de Laplace (2.3). Em particular considere a equação $u'' = -1$ em $(0,1)$ com $u(0) = u(1) = 0$. A formulação fraca desta equação foi construída na Seção 2.2.2 usando a

forma bilinear $\mathcal{A}$ em (2.6) e do funcional linear $\mathcal{F}$ em (2.7). Usamos uma triangulação *estruturada*, isto é, com vértices igualmente espaçados.

$n \downarrow$	Iterações	(est. cond)
16	8	103.09
32	16	414.35
64	32	1659.38
128	64	6639.52
256	128	26560.07
512	256	106242.29
1024	512	424971.18
2048	1024	1699886.72

Tabela 4.2: Número de iterações do gradiente conjugado para o problema considerado nesta seção. Aqui $h = 1/n$ e usamos uma tolerância de 10^{-6}. Veja a fórmula (4.12).

Seguimos os mesmos passos do exemplo no final da Seção 2.3.1 mas na hora de resolver os sistemas lineares usamos o método do gradiente conjugado. Obtemos os resultados da Tabela 4.2. Vemos que quando o parâmetro $h = 1/n$ é dividido por dois, o número de condição é multiplicado por um fator de 4 e o número de iterações duplica-se. Veja a fórmula (4.12). Concluímos que o número de iteração cresce de forma linear com respeito ao parâmetro h. Vemos também que se precisamos usar $h << 1$, o número de iterações necessárias, para calcular a solução de elementos finitos com a tolerância desejada, é muito grande. A situação fica pior em dimensões dois e três.

Consideramos agora o exemplo da Seção 2.5,

$$\begin{array}{l} \textit{Achar } u : [0,1] \to \mathbb{R} \textit{ tal que:} \\ \begin{cases} -(\kappa(x)u'(x))' = -1, & 0 < x < 1 \\ u(0) = 0, u(1) = 1. \end{cases} \end{array} \tag{4.13}$$

onde

$$\kappa(x) = \kappa_1(x, \mu) + 100\kappa_2(x, p)$$

com κ_1 e κ_2 definidos em (2.35) e (2.36). Com $\mu = 1000$ e $p = 30$ obtemos os resultados da Tabela 4.3. Concluímos que o número de iterações necessárias aumenta com o contraste do meio. A situação fica pior em dimensões maiores.

$n \downarrow$	Iterações	(est. cond)
16	15	696.97
32	35	3166.30
64	86	16806.85
128	228	126804.46
256	588	869532.27
512	1439	5132737.47
1024	3356	24495669.25
2048	>5000	47948970.24

Tabela 4.3: Número de iterações do gradiente conjugado para o problema (4.13) com coeficiente $\kappa(x) = \kappa_1(x, \mu) + 100\kappa_2(x, p)$. Aqui $h = 1/n$ e usamos uma tolerância de 10^{-6}.

Duas dimensões

Suponha que queremos calcular a solução de elementos finitos da equação de Laplace (3.7) com $D = [0, 1] \times [0, 1]$ usando elementos finitos, isto é, queremos resolver

$$\begin{array}{l} \textit{Achar } u : [0,1] \times [0,1] \to \mathbb{R} \textit{ tal que:} \\ \begin{cases} -\Delta u(x) = -1, & x \in [0,1] \times [0,1] \\ u(x) = 1, & x \in \partial([0,1] \times [0,1]). \end{cases} \end{array}$$

Usamos a formulação fraca construída na Seção 3.3.1 com $\mathcal{A}$ em (3.10) e $\mathcal{F}$ em (3.11). Usamos uma triangulação estruturada. Dividimos o quadrado $[0, 1] \times [0, 1]$ em $n \times n$ quadrados e cada quadrado é dividido em dois triângulos. Veja Figura 4.1. Seguimos os mesmos passos do exemplo no final da Seção 3.4.1 mas na hora de resolver os sistemas lineares usamos o método do gradiente conjugado. Obtemos os resultados da Tabela 4.4. Observamos novamente que o número de iteração cresce linearmente com h. O número de condição cresce quadraticamente com h. Vemos que para obter a mesma ordem h

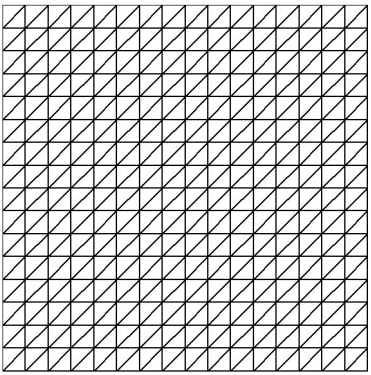


Figura 4.1: Triangulação estruturada com $h = \sqrt{2}/n$ e $n = 16$ usada para aproximar a solução da equação de Laplace em duas dimensões

que no caso de uma dimensão precisamos de muitas mas iterações. A situação piora em dimensões maiores.

$n \downarrow$	N_h^v	Iterações	(est. cond)
8	81	9	25.27
16	289	25	103.09
32	1089	51	414.35
64	4225	100	1659.38
128	16641	197	6639.52

Tabela 4.4: Número de iteração do gradiente conjugado para o problema considerado nesta seção. Aqui $h = \sqrt{2}/n$ e usamos uma tolerância de 10^{-6}. Veja a fórmula (4.12).

Voltamos a atenção para a equação

$$\begin{array}{l} \textit{Achar } u : D \subset \mathbb{R}^2 \to \mathbb{R} \textit{ tal que:} \\ \begin{cases} -\mathrm{div}(\kappa(x)\nabla u(x)) = & f(x) \quad x \in D \\ u(x) = & g(x) \quad x \in \partial D \end{cases} \end{array}$$

onde $\kappa(x) = 1 + 1000(1 + \sin(6\pi x_1)\sin(4\pi x_2))$. O contrate deste coeficiente e ao redor de 2000. Obtemos os resultados na Tabela 4.5.

$n \downarrow$	N_h^v	Iterações	(est. cond)
8	81	17	27.71
16	289	42	155.64
32	1089	95	785.21
64	4225	205	3498.16
128	16641	430	14700.20

Tabela 4.5: Número de iteração do gradiente conjugado para o problema com coeficiente $\kappa(x) = 1+1000(1+\sin(6\pi x_1)\sin(4\pi x_2))$. Aqui $h = \sqrt{2}/n$ e usamos uma tolerância de 10^{-6}. Veja a fórmula (4.12).

Agora consideramos o problema elíptico geral acima mas com o coeficiente

$$\kappa(x) = \Big(\kappa_1(x_1,\mu) + 100\kappa_2(x_1,p)\Big)\Big(\kappa_1(x_2,\mu) + 100\kappa_2(x_2,p)\Big) \quad (4.14)$$

onde κ_1 e κ_2 definidos em (2.35) e (2.36). Com $\mu = 1000$ e $p = 30$ obtemos os resultados da Tabela 4.6.

$n \downarrow$	N_h^v	Iterações	(est. cond)
8	81	27	162.99
16	289	110	1238.65
32	1089	289	5859.17
64	4225	814	31908.59
128	16641	>2000	240630.26

Tabela 4.6: Número de iteração do gradiente conjugado para o problema elíptico com coeficiente (4.14). Aqui $h = \sqrt{2}/n$ e usamos uma tolerância de 10^{-6}. Veja a fórmula (4.12).

Das Tabelas 4.6 e 4.5 concluímos que o número de iterações ate a convergência do método do gradiente conjugado aumenta consideravelmente com a presencia de contraste alto e múltiplas escalas.

4.4 O método do gradiente conjugado precondicionado

O método do gradiente conjugado pode ser modificado para obter o método do gradiente conjugado precondicionado. Com o método do gradiente conjugado precondicionado podemos obter as aproximações de elementos finitos das equações diferenciais mencionadas na Seção 4.3 em muitas menos iterações e portanto menos tempo computacional. O sucesso no uso do método do gradiente conjugado precondicionado depende da escolha de um bom precondicionador.

Para resolver um sistema linear

$$Ax = b$$

onde a matriz A é simétrica e definida positiva mas também muito grande, $n >> 1$, e muito mal condicionada, $\text{Cond}(A) >> 1$, aplicar diretamente o método do gradiente conjugado resultaria em demasiadas iterações para alcançar a tolerância requerida. A idéia é então usar um precondicionador M^{-1} onde M é simétrica e definida positiva, isto é, resolver o sistema linear

$$M^{-1}Ax = M^{-1}b$$

no lugar de resolver o sistema linear original $Ax = b$. Note que os dois sistemas lineares tem a mesma solução. Dado que, em geral, $M^{-1}A$ não é simétrica, para poder aplicar o método do gradiente conjugado usamos a substituição $z = M^{1/2}x$ e obtemos que z satisfaz

$$M^{-1/2}AM^{-1/2}z = M^{-1/2}b. \tag{4.15}$$

A matriz $M^{-1/2}AM^{-1/2}$ é simétrica e definida positiva. Podemos estimar a condição dela usando a rata entre o maior e o menor autovalor. O produto interno gerado por $M^{-1/2}AM^{-1/2}$ é dado por

$$(z, M^{-1/2}AM^{-1/2}w) = (M^{-1/2}z, AM^{-1/2}w) = (x, M^{-1}Ay) \tag{4.16}$$

onde $x = M^{1/2}z$ e $y = M^{-1/2}w$. Note também que para todo vetor $x, y \in \mathbb{R}^n$ temos

$$(M^{-1/2}x, M^{-1/2}y) = (x, M^{-1}y). \tag{4.17}$$

Se resolvemos os sistema linear (4.15) usando o método do gradiente conjugado na Tabela 4.1, obtemos o método do gradiente conjugado precondicionado. Fazendo algumas manipulações usando (4.16), (4.17) e lembrando que queremos calcular x, obtemos o algoritmo na Tabela 4.7.

1. Inicializar $q_0 = b - Ax_0$
2. Iterar $k = 1, 2, \ldots,$ ate a convergência

$$\begin{aligned} z_{k-1} &= M^{-1}q_{k-1} \text{(Precondicionador)} \\ \beta_k &= \frac{(z_{k-1}, q_{k-1})}{(z_{k-2}, q_{k-2})} \quad [\beta_1 = 0] \\ d_k &= z_{k-1} + \beta_k d_{k-1} \quad [d_1 = z_0] \\ \alpha_k &= \frac{(z_{k-1}, q_{k-1})}{(d_k, Ad_k)} \\ x_k &= x_{k-1} + \alpha_k d_k \\ q_k &= q_{k-1} - \alpha_k A d_k \end{aligned}$$

Tabela 4.7: Algoritmo do gradiente conjugado para resolver $Ax = b$ com precondicionador M.

O número de iterações dependera então da condição da matriz $M^{-1/2}AM^{-1/2}$,

$$\text{Cond}(M^{-1/2}AM^{-1/2}) = \frac{\lambda_{\max}}{\lambda_{\min}}$$

onde $\lambda_{\max}$ e $\lambda_{\min}$ são o maior e menor autovalor da matriz simétrica definida positiva $M^{-1/2}AM^{-1/2}$. Temos o seguinte lema que ensina como calcular o número de condição desta matriz.

Lema 29. *Suponha que A e M são matrizes simetrias e definidas positivas. Os seguintes problemas de autovalores tem os mesmos autovalores*

1. $M^{-1/2}AM^{-1/2}x = \lambda x$
2. $M^{-1}Ax = \lambda x$
3. $Ax = \lambda Mx$ *(problema de autovalores generalizados)*

Também vale $Cond(M^{-1/2}AM^{-1/2}) = \frac{\lambda_{\max}}{\lambda_{\min}}$ *onde* $\lambda_{\max} = \max_{1\leq i\leq n} \lambda_i$ *e* $\lambda_{\min} = \min_{1\leq i\leq n} \lambda_i$.

Note que podemos usar os autovalores de qualquer um dos problemas 1), 2) ou 3) no lema anterior. Em particular temos que

$$\text{Cond}(M^{-1}A) = \text{Cond}(M^{-1/2}AM^{-1/2}) = \frac{\lambda_{\max}}{\lambda_{\min}}. \tag{4.18}$$

Onde o primeiro número de condição deve ser interpretado como o número de condição espectral, isto é, definido como a rata entre o maior e o menor autovalor do problema de autovalor 2) ou 3) no Lema 29. O operador M (ou M^{-1}) é conhecido como precondicionador. Para construir um bom precondicionador precisamos levar em conta que,

1. o cálculo $M^{-1}q$ não deve ser muito custoso em termos de memória e tempo de computação.
2. o numero de condição $\text{Cond}(M^{-1}A)$ em (4.18) deve ser menor que o numero de condição $\text{Cond}(A)$. Podemos pensar que M^{-1} é uma aproximação da inversa da matriz A, isto é, $M^{-1} \approx A^{-1}$.

É desejável poder usar computação paralela de forma eficiente no cálculo $M^{-1}q$. Em geral, dado um vetor q, o custo de calcular Aq pode ser menor que o custo de calcular $M^{-1}q$. Note que na Tabela 4.7 temos somente um cálculo da forma $M^{-1}q$ em cada iteração. No caso de sistemas linear de elementos finitos é também desejável que o numero de condição $\text{Cond}(M^{-1}A)$ dependa *pouco* do tamanho da matriz, isto é, da dimensão do espaço de elementos finitos.

Na prática, depois de construir o precondicionador M^{-1}, temos que estimar o número de condição $\text{Cond}(M^{-1}A)$ em (4.18). Um jeito relativamente simples de estimar o número de condição da matriz precondicionada é usando o seguinte lema que pode ser deduzido facilmente do problema 2) do Lema 29.

Lema 30. *Se existem constantes $c, C > 0$ tais que*

$$x^T Ax \leq Cx^T Mx \quad e \quad y^T Ay \geq cy^T My \quad \text{para todo } x, y \in \mathbb{R}^n,$$

então, $Cond(M^{-1}A) = Cond(M^{-1/2}AM^{-1/2}) \leq C/c$.

O seguinte lema também é útil quando precisamos estimar o número de condição do operador precondicionado.

Lema 31. *Sejam A e M^{-1} simétricas positivas definidas e sejam c e C duas constantes positivas. São equivalentes,*

$$c(Ax, x) \leq (AM^{-1}Ax, x) \leq C(Ax, x), \quad \text{para todo } x \in \mathbb{R}^n, \tag{4.19}$$

$$c\|x\|_A \leq \|M^{-1}Ax\|_A \leq C\|x\|_A, \quad \text{para todo } x \in \mathbb{R}^n, \tag{4.20}$$

$$\frac{1}{C}(Ax, x) \leq (M^{-1}x, x) \leq \frac{1}{c}(Ax, x), \quad \text{para todo } x \in \mathbb{R}^n, \tag{4.21}$$

$$\frac{1}{C}(M^{-1}x, x) \leq (Ax, x) \leq \frac{1}{c}(M^{-1}x, x), \quad \text{para todo } x \in \mathbb{R}^n. \tag{4.22}$$

Para estimar o número de condição do operador precondicionado podemos provar qualquer uma das desigualdades acima. Nos Capítulos 5 e 6 usaremos (4.20).

Capítulo 5

Métodos com superposição em dimensão um

Neste capítulo construiremos precondicionadores de decomposição de domínios para sistemas lineares que resultam na aproximação numérica de equações diferenciais elípticas em dimensão um. Consideramos especificamente o sistema linear obtido usando elementos finitos do Capítulo 2. O material apresentado está baseado nas notas de um minicurso e um curso regular no IMPA dirigidos pelo Professor Marcus Sarkis em métodos de decomposição de domínios. Para um estudo mais detalhado veja [31, 24, 30, 27] e as referências ali citadas.

5.1 Decomposição com e sem sobreposição

Seja $\mathcal{T}^h = \{K\}$ uma triangulação do intervalo (a, b). Consideramos a formulação fraca na Seção 2.2.3. Introduzimos uma partição do domínio $D = (a, b)$ da equação diferencial em subdomínios (subintervalos) disjuntos $\{D_i = (a_i, b_i)\}_{i=1}^{N_S}$ com

$$a = a_1 < b_1 = a_2 < b_2 = a_2 < \cdots < b_{N_S} = b$$

onde N_S é o número de subdomínios. Esta decomposição é dita sem sobreposição. Assumimos que cada subdomínio D_i, $i = 1, \ldots, N_s$ é a união de elementos da triangulação $\mathcal{T}$. Com esta decomposição construímos uma nova cobertura $\{D_i'\}_{i=1}^{N_S}$ do intervalo $D = (a, b)$ com sobreposição δ definindo

$$D_i' = \{x \in D \ : \ |x-y| < \delta, \text{ para algum } y \in D_i\} = (a_i-\delta, b_i+\delta) \cap D.$$

Assumimos também que cada subdomínio D_i', $i = 1, \ldots, N_S$ é a união de elementos da triangulação. Vide Figura 5.1. Usaremos a notação

- N_h^e número de elementos da triangulação
- N_h^v número de vértices da triangulação
- $N_h^{(i),e}$ número de elementos do subdomínio D_i
- $N_h^{(i),I}$ número de vértices interiores no subdomínio D_i'.

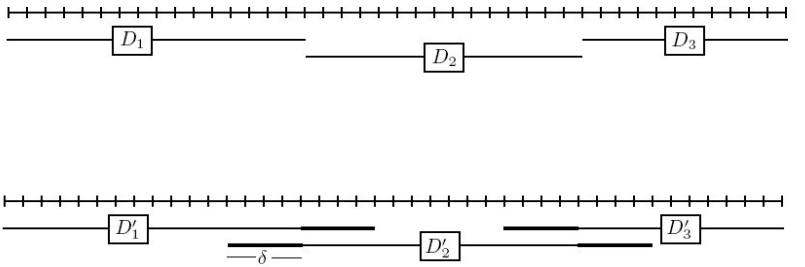


Figura 5.1: Exemplo de uma decomposição sem sobreposição (acima) e a decomposição com sobreposição construída aumentando cada subdomínio com $\delta = 4h$ (embaixo).

Usando a decomposição com sobreposição $\{D_i'\}_{i=1}^{N_S}$ do domínio D, vamos construir um precondicionador de decomposição de domínios. Os ingredientes principais de um precondicionador de decomposição de domínios são:

1. os espaços locais

2. operadores de restrição e extensão

3. em cada espaço local teremos que definir uma aproximação da forma bilinear da formulação fraca, ou seja, uma matriz local que aproxime a sub-matriz da matriz global.

4. um (ou ate vários) espaço(s) grosso(s).

5.2 Espaços locais, operadores de restrição e extensão

No Capítulo 2 estudamos o espaço de elementos finitos de funções lineares por partes. O espaço de elementos finitos usado para aproximar (2.10) é o espaço

$$V := \mathbb{P}_0^1(\mathcal{T}^h).$$

Neste capítulo V é o espaço *global* de nosso método de decomposição de domínios. Dada uma decomposição com sobreposição δ, $\{D_i'\}$ definimos os espaços locais por

$$V^{(i)} = V^{(i)}(D_i') = \mathbb{P}_0^1(D_i') = \text{span}\{\phi_i; x_i \in D_i'\}, \tag{5.1}$$

onde as funções base chapéu $\{\phi_i\}$, foram definidas em (2.25). O espaço local $V^{(i)}$ é somente a restrição do espaço global V aos vértices interiores ao subdomínio D_i', $i = 1, \ldots, N_S$. Vide Figura 5.2.

Definimos a matriz de restrição $R^{(i)}$ de dimensão $N_h^{(i),I} \times N_h^v$ com entradas nulas em todas as posições com excessão das posições (ℓ, j) onde o índice j corresponde ao vértice $x_j \in D_i'$. As matrizes $R^{(i)}$, $i = 1, \ldots, N_S$, são análogas as matrizes de restrição do Lema 10. Note que a matriz $R^{(i)T}$ representa o operador linear extensão por zero para fora do D_i', , $i = 1, \ldots, N_S$,.

Para cada D_i', $i = 1, \ldots, N_S$, definimos $\mathcal{A}^{(i)}$, a restrição da forma bilinear $\mathcal{A}$ ao subdomínio $V^{(i)}$ por,

$$\mathcal{A}^{(i)}(u_i, v_i) = \int_{D_i'} \kappa(x) u_i'(x) v_i'(x) dx, \quad u_i, v_i \in V^{(i)}. \tag{5.2}$$

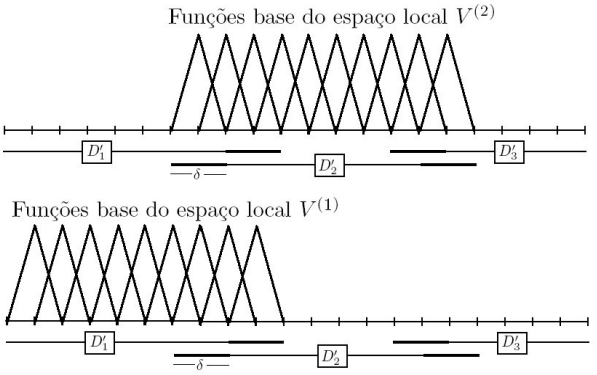


Figura 5.2: Funções base dos espaços locais $V^{(1)}$ (embaixo) e $V^{(2)}$ (acima) para a decomposição $\{D'_1, D'_2, D'_3\}$.

Seja $A^{(i)}$ a matriz $N_h^{(i),I} \times N_h^{(i),I}$ que representa a forma bilinear local $\mathcal{A}^{(i)}$. A matriz $A^{(i)}$, $i = 1, \ldots, N_S$, é a matriz de Dirichlet da mesma equação diferencial que estamos considerando *mas* restrita ao subdomínio D'_i. Note que $A^{(i)}$ é uma matriz invertível. Observe também que para $i = 1, \ldots, N_S$, e $u_i, v_i \in V^{(i)}$ temos

$$\begin{aligned} \boldsymbol{u}_i^T A^{(i)} \boldsymbol{v}_i &= \mathcal{A}^{(i)}(u_i, v_i) \\ &= \mathcal{A}(R^{(i)T} u_i, R^{(i)T} v_i) = (R^{(i)T} \boldsymbol{u}_i)^T A (R^{(i)T} \boldsymbol{v}_i) \end{aligned}$$

onde $R^{(i)T} w_i$ é a função de elementos finitos com representação vetorial $R^{(i)T} \boldsymbol{w}_i \in \mathbb{R}^{N_h^v}$ e A é matriz global, isto é, a representação matricial da forma bilinear global $\mathcal{A}$. A função $R^{(i)T} w_i$ é nula nos vértices fora do D'_i, $i = 1, \ldots, N_S$. Concluímos que

$$A^{(i)} = R^{(i)} A R^{(i)T}, \quad i = 1, \ldots, N_S. \tag{5.3}$$

A matriz local $A^{(i)}$ é o bloco diagonal da matriz A correspondente aos índices associados aos vértices interiores ao subdomínio D'_i, $i = 1, \ldots, N_S$. Veja a ilustração da Figura 5.3.

$$A = \begin{bmatrix} a_{12} & a_{13} & a_{14} & a_{15} & a_{16} & \dots \\ a_{22} & a_{23} & a_{24} & a_{25} & a_{26} & \dots \\ a_{32} & a_{33} & a_{34} & a_{35} & a_{36} & \dots \\ a_{42} & a_{43} & a_{44} & a_{45} & a_{46} & \dots \\ a_{52} & a_{53} & a_{54} & a_{55} & a_{56} & \dots \\ a_{62} & a_{63} & a_{64} & a_{65} & a_{66} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \qquad \begin{aligned} V^{(2)} &= \operatorname{span}\{\phi_3, \phi_4, \phi_5\} \\ A^{(2)} &= \begin{bmatrix} a_{33} & a_{34} & a_{35} \\ a_{43} & a_{44} & a_{45} \\ a_{53} & a_{54} & a_{55} \end{bmatrix} \end{aligned}$$

Figura 5.3: Exemplo de matriz local $A^{(2)}$ que corresponde ao espaço local $V^{(2)} = \operatorname{span}\{\phi_3, \phi_4, \phi_5\}$

5.3 Precondicionador aditivo de um nível

Com a notação introduzida na Seção 5.2 definimos o precondicionador aditivo de um nível M_1^{-1} por

$$M_1^{-1} = \sum_{i=1}^{N_S} R^{(i)T} \left[A^{(i)}\right]^{-1} R^{(i)}. \tag{5.4}$$

Lembre que para usar o precondicionador acima no método do gradiente conjugado precondicionado precisamos discutir como obter $M_1^{-1}q$ dado um vetor q do tamanho apropriado. Observamos que

1. $M_1^{-1}q = \sum_{i=1}^{N_S} R^{(i)T} \left[A^{(i)}\right]^{-1} R^{(i)}q = \sum_{i=1}^{N_S} R^{(i)T}u_i$ onde definimos $u_i = \left[A^{(i)}\right]^{-1} R^{(i)}q$.

2. Cada parcela da soma na definição de M_1^{-1} em (5.4) pode ser calculada independente das outras. Pode-se então utilizar computação paralela para implementar o precondicionador M_1^{-1}.

3. Para calcular a i-ésima parcela $u_i = \left[A^{(i)}\right]^{-1} R^{(i)}q$ no lugar de aplicar a inversa da matriz local $A^{(i)}$ podemos resolver o sistema linear
$$A^{(i)}u_i = R^{(i)}q$$
que como sabemos equivale a solução da mesma equação diferencial parcial mas no subdomínio D_i' com condição de contorno

de Dirichlet. Note que a dimensão deste sistema linear é pequena quando comparada com a dimensão do sistema linear global.

4. Nas aplicações práticas pode-se substituir u_i (solução exata do problema local) por alguma aproximação da mesma. Isto é, não precisamos resolver os sistemas lineares locais com precisão total.

5. Depois de calcular as soluções dos *problemas locais*, montamos $M_1^{-1}q$ usando as matrizes de extensão: $M_1^{-1}q = \sum_{i=1}^{N_S} R^{(i)T} u_i$.

A idéia é então usar o método do gradiente conjugado precondicionado da Tabela 4.7 na página 70 com o precondicionador M_1^{-1} acima. Em cada iteração do método na Tabela 4.7 devemos

1. aplicar a matriz global A. Isto pode ser feito usando a fórmula (2.31).

2. aplicar $M_1^{-1}q$ levando em conta as recomendações acima.

O número de iterações até a convergência depende da condição do operador precondicionado $M_1^{-1}A$. O seguinte teorema fornece uma estimativa para o número de condição desta matriz. A prova deste resultado não será apresentada aqui. Na Seção 5.7 será apresentada a idéia da prova para o caso do precondicionador de dois níveis que é um pouco mais complicada que a prova do resultado enunciado nesta seção.

Teorema 32. *Considere o precondicionador M_1^{-1} definido em (5.4). A matriz A é a matriz da forma bilinear definida em (2.13) com o coeficiente κ satisfazendo (2.11). Existe uma constante C independente de h tal que*

$$Cond(M_1^{-1}A) \leq C\left(1 + \frac{1}{\delta H}\right)$$

onde $H = \max_{1\leq i\leq N_s} diâmetro(D_i)$ e δ é o parâmetro da decomposição com superposição $\{D_i'\}$. A constante C pode depender do contraste do meio.

5.4 Experimentos numéricos

Nesta seção consideramos a equação de Laplace $u'' = -1$ com condição de Dirichlet no intervalo $(0, 1)$. Dividimos o intervalo $(0, 1)$ em N subdomínios, i.e., $H = 1/N$. Cada subdomínio é dividido em n elementos, i.e., $h = 1/(nN)$. Calcularemos a solução usando o método do gradiente conjugado precondicionado com o precondicionador aditivo de um nível. Mostramos os resultados na Tabela 5.1. Compare com os resultados da Tabela 4.2 na página 65. Por exemplo, para $h = 1/1024$ temos 512 iterações do gradiente conjugado sem precondicionador e um número de condição de 424971.18. Para o mesmo valor de h na Tabela 5.1 obtemos 34 iterações e um número de condição de 3323.64 se tomamos $N = 32$ e $n = 32$. Neste caso, em cada uma das 34 iterações do método do gradiente conjugado precondicionado temos que resolver 32 problemas locais de tamanho 33×33. Podemos também obter a solução em 18 iterações se tomamos $N = 16$ e $n = 64$ ou em 64 iterações se $N = 64$ e $n = 16$.

$n\backslash N$	4	8	16	32	64
4	3(4.53)	6(26.53)	10(104.21)	18(415.41)	34(1660.41)
8	4(11.47)	9(53.19)	17(208.74)	31(831.22)	54(3321.28)
16	5(28.21)	9(105.92)	17(417.13)	33(1662.15)	64(6642.28)
32	5(55.57)	9(211.10)	17(833.59)	34(3323.64)	67(13283.92)
64	5(110.23)	9(421.34)	18(1666.34)	34(6646.46)	68(26567.01)

Tabela 5.1: Número de iterações do gradiente conjugado precondicionado (estimativa do número de condição) para o problema considerado nesta seção. Aqui $h = 1/(nN)$ e $H = 1/N$ e o fixamos $\delta = 2h$.

Na Tabela 5.1 o tamanho da sobreposição é fixado em $2h$ para todos os casos. Note que a sobreposição fica menor com h. Pelo Lema 32 o número de condição do operador precondicionado $M_1^{-1}A$ é limitado pelo número

$$1 + \frac{1}{\delta H} = 1 + \frac{1}{2}nN^2.$$

Os resultados da Tabela 5.1 coincidem com esta afirmação. O número de condição multiplica por dois (aproximadamente) nas colunas e por

um fator de quatro nas linhas. Na Figura 5.4 ilustramos as cinco iterações do caso $N = 4$ e $n = 32$ na Tabela 5.1. Vemos a efetividade das iterações e a dependência no tamanho da sobreposição δ. Repetimos o mesmo experimento mas agora usamos uma so-

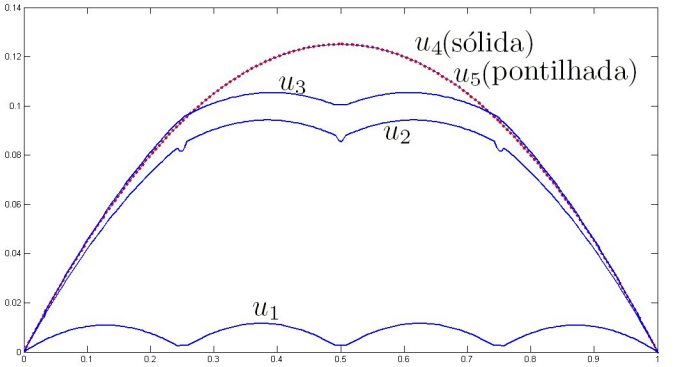


Figura 5.4: Convergência rápida do método do gradiente conjugado precondicionado com o precondicionador de um nível. Aqui u_i, é a $i-$ésima iteração do método, $i = 1, \ldots, 5$. Observe que as iteração u_4 e u_5 estão muito próximas e u_5 é a solução com tolerância de 10^{-6}. Aqui temos $N = 4$ subdomínios ($H = 1/4$) e $n = 32$ elementos em cada subdomínio ($h = 1/(4 \times 32)$). Veja Tabela 5.1.

breposição generosa $\delta = nh = 1/N$. Veja os resultados na Tabela 5.2. Os resultados coincidem com a teoria também neste caso já que $1 + \frac{1}{\delta H} = 1 + N^2$.

5.5 Precondicionador de dois níveis

Nesta seção introduziremos o método aditivo de dois níveis. Para este fim introduzimos uma triangulação grossa $\mathcal{T}^H$ adicional. Supomos que $H > h$. A nova triangulação pode, em geral, ser independente da triangulação mais fina, da decomposição $\{D_i\}_{i=1}^{N_S}$ e da decomposição $\{D_i'\}_{i=1}^{N_S}$. Para simplificar a apresentação assumiremos que a triangulação grossa é dada por $\mathcal{T}^H = \{D_i\}_{i=1}^{N_S}$, isto é, a malha

$n\backslash N = 1/\delta$	4	8	16	32	64
4	3(3.00)	5(10.18)	9(39.32)	14(156.02)	24(622.90)
8	3(3.00)	5(10.18)	9(39.32)	14(156.02)	24(622.90)
16	3(3.00)	5(10.18)	9(39.32)	15(156.02)	24(622.90)
32	3(3.00)	5(10.18)	9(39.32)	15(156.02)	25(622.90)
64	3(3.00)	5(10.18)	9(39.32)	15(156.02)	25(622.90)

Tabela 5.2: Número de iteração e estimativa do número de condição do gradiente conjugado precondicionado para o problema desta seção. Aqui $h = 1/(nN)$ e $H = 1/N$ e o fixamos $\delta = H$

grossa coincide com a decomposição original da qual construímos o precondicionador de um nível.

Dada uma malha grossa consideramos um espaço de elementos finitos grossos baseado nos vértices de $\mathcal{T}^H$, isto é, um espaço de elementos finitos da forma

$$V^H = \text{span}\{\Phi_j \in V^h \ : \ i = 1, \ldots, N_H^v\}$$

onde as funções bases do espaço grosso serão definidas logo e N_H^v é o número de vértices da malha grossa. Como $\Phi_j \in V^h$ pode-se escrever

$$\Phi_j = \sum_{i=1}^{N_h^v} \Phi_j(x_i)\phi_i \quad \text{ou} \quad \boldsymbol{\Phi}_j = [\Phi_j(x_1), \ldots, \Phi_j(x_{N_h^v})]^T$$

onde ϕ_i, $i = 1, \ldots, N_h^v$, são as funções base da malha fina $\mathcal{T}^h$ definidas em (2.25). Denotemos por $R^{(0)}$ a transposta da matriz

$$R^{(0)T} = [\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_{N_H^v}]$$

e definamos a matriz grossa $A^{(0)}$ como a matriz global A na base grossa,

$$A^{(0)} = R^{(0)} A R^{(0)T}. \tag{5.5}$$

Note que se $u_0, v_0 \in V^{(0)}$ temos

$$\boldsymbol{u}_0 A^{(0)} \boldsymbol{v}_0 = (R^{(0)T}\boldsymbol{u}_0)^T A (R^{(0)T}\boldsymbol{v}_0) = \mathcal{A}(R^{(0)T}u_0, R^{(0)T}v_0)$$

onde $R^{(0)T}w_0 \in V$ é a função de elementos finitos com representação $R^{(0)}\boldsymbol{w}_0 \in \mathbb{R}^{N_h^v}$. Concluímos que $A^{(0)}$ é a representação matricial da

forma bilinear $\mathcal{A}^{(0)}$ definida como a restrição da forma bilinear $\mathcal{A}$ ao subespaço $V^{(0)} \subset V$,

$$\mathcal{A}^{(0)}(u_0, v_0) = \mathcal{A}(R^{(0)T}u_0, R^{(0)T}v_0). \tag{5.6}$$

Definimos o precondicionador aditivo de dois níveis M_2^{-1}

$$M_2^{-1} = R^{(0)T}\left[A^{(0)}\right]^{-1}R^{(0)} + \sum_{i=1}^{N_S} R^{(i)T}\left[A^{(i)}\right]^{-1}R^{(i)} \tag{5.7}$$

$$= R^{(0)T}\left[A^{(0)}\right]^{-1}R^{(0)} + M_1^{-1}, \tag{5.8}$$

onde M^{-1} é o precondicionador aditivo de um nível definido em (5.4). Para aplicar M_2^{-1} temos que aplicar M_1^{-1} como antes e aplicar o termo $R^{(0)}\left[A^{(0)}\right]^{-1}$. Observe que,

1. $M_2^{-1}q = \sum_{i=0}^{N_S} R^{(i)T}\left[A^{(i)}\right]^{-1}R^{(i)}q = \sum_{i=0}^{N_S} R^{(i)T}u_i$ onde definimos $u_i = \left[A^{(i)}\right]^{-1}R^{(i)}q$, $i = 0, 1, \ldots, N_S$.

2. Como no caso de M_1, cada parcela da soma na definição de M_2^{-1} pode ser calculada em paralelo.

3. Para calcular a 0-ésima parcela $u_0 = \left[A^{(0)}\right]^{-1}R^{(0)}q$ no lugar de aplicar a inversa da matriz grossa $A^{(0)}$ resolvemos o sistema linear
$$A^{(0)}u_0 = R^{(0)}q$$
que como sabemos equivale a solução da mesma equação diferencial no domínio D no espaço de elementos finitos $V^{(0)}$ baseado na triangulação grossa $\mathcal{T}^H$. Note que a dimensão deste sistema linear é pequena quando comparada com a dimensão do sistema linear global na malha fina $\mathcal{T}^h$. Em particular a dimensão do problema grosso é da ordem do número de vértices na malha grossa $\mathcal{T}^H$.

4. Depois de calcular as soluções dos problemas locais u_i, $i = 1, \ldots, N_S$, e do problema grosso u_0, montamos $M_2^{-1}q$ usando as matrizes de extensão: $M_2^{-1}q = \sum_{i=0}^{N_S} R^{(i)T}u_i$.

5.5.1 Espaços grossos

Nesta seção vamos a descrever as funções bases $\{\Phi_j\}$ que definem o espaço grosso $V^{(0)}$. Em geral as funções base grossa devem ser calculadas antecipadamente. Pode-se por exemplo usar funções bases cuja construção requer um alto custo computacional já que este cálculo é feito somente uma vez.

As funções grossas devem ser escolhidas de tal forma que gerem funções com comportamento similar à solução do problema considerado. Existem muitas escolhas possíveis para as funções bases grossas. Vamos descrever unicamente duas escolhas para as funções base grossas.

Funções base lineares por partes na malha grossa

Podemos escolher Φ_j como sendo as funções base chapéu na malha grossa $\mathcal{T}^H = \{D_i\}_{i=1}^{N_S}$. A vantagem desta escolha é o baixo custo computacional requerido para construir as funções bases. Existem uma função por vértice da malha grossa e duas funções com suporte não nulo em cada elemento grosso D_i. Sejam $y_0, \ldots, y_{N_S+1}$ os vértices da triangulação grossa $\mathcal{T}^H$, temos

$$\Phi_j(x) = \begin{cases} 1, & \text{se } x = y_j, \text{ (1 no vértice } y_j) \\ 0, & \text{se } x = y_k, k \neq j, \text{(0 nos outros vértices)} \\ \text{extensão linear}, & \text{se } x \text{ não é vértice de } \mathcal{T}^H. \end{cases} \tag{5.9}$$

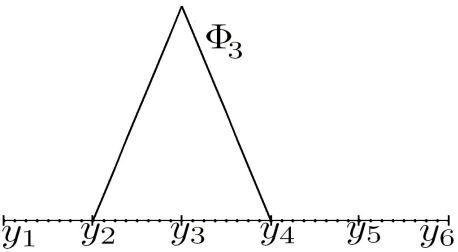


Figura 5.5: Função base grossa linear por partes na malha grossa.

Funções de elementos finitos multi-escala

No lugar de usar funções lineares por partes podemos usar funções que localmente resolvem a equação diferencial, isto é, que a função base seja uma solução da (ou de uma) equação diferencial em cada elemento da malha grossa. No lugar de usar uma extensão linear dentro de cada elemento da malha grossa Φ_j, podemos usar uma *extensão harmônica* Φ_j^{MS} definida por

$$\Phi_j^{MS}(x) = \begin{cases} 1, & \text{se } x = y_j, \\ 0, & \text{se } x = y_k, k \neq j, \\ \text{extensão harmônica}, & \text{se } x \text{ não é vértice de } \mathcal{T}^H, \end{cases} \tag{5.10}$$

para $j = 1, \ldots, N_S + 1$. Aqui extensão harmônica no interior do elemento quer dizer que Φ_j satisfaz a equação

$$\begin{aligned} &\mathcal{A}(\Phi_j^{MS}, v) = 0 \quad \forall v \in \mathbb{P}_0^1(\mathcal{T}^h|_{D_j}) \\ &\Phi_j(y_{j-1}) = 0 \text{ e } \Phi_j(y_j) = 1, \end{aligned}$$

onde $\mathcal{T}^h|_{D_j}$ denota a restrição da malha fina ao elemento da malha grossa $D_j = (y_{j-1}, y_j)$. No caso da forma bilinear $\mathcal{A}$ definida em (2.13), Φ_j é a aproximação de elementos finitos da equação,

$$\begin{aligned} &-(\kappa(x)\Phi_j^{MS}(x))' = 0 \quad x \in D_j = (y_{j-1}, y_j) \\ &\Phi_j^{MS}(y_{j-1}) = 0 \text{ e } \Phi_j(y_j)^{MS} = 1. \end{aligned}$$

Quando $\kappa(x) = 1$ para todo $x \in D_j$ temos que Φ_j^{MS} é uma extensão linear dentro do elemento grosso D_j e portanto $\Phi_j^{MS} = \Phi_j$ para todo j.

5.5.2 Número de condição

O seguinte teorema fornece uma estimativa para o número de condição do operador precondicionado $M_2^{-1}A$. A idéia da prova deste resultado é apresentada na Seção 5.7.

Teorema 33. *Considere o precondicionador M_2^{-1} definido em (5.7) com o espaço grosso de funções lineares ou funções de elementos finitos multi-escala. Se A é a matriz da forma bilinear definida em (2.13)*

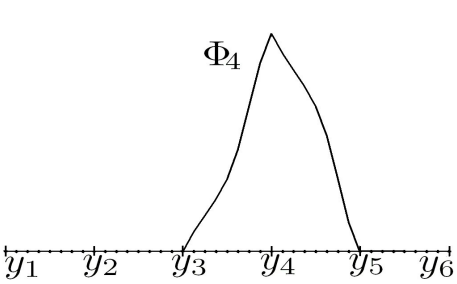


Figura 5.6: Função base grossa de elementos finitos multi-escala com coeficiente $\kappa(x) = 2 + \sin(10\pi x)$.

com o coeficiente κ satisfazendo (2.11), então existe uma constante C independente de h e H tal que

$$Cond(M_2^{-1}A) \leq C\left(1 + \frac{H}{\delta}\right)$$

onde $H = \max_{1\leq i\leq N_s} diâmetro(D_i)$ e δ é o parâmetro da decomposição com superposição $\{D_i'\}$. A constante C pode depender do contraste do coeficiente κ.

5.6 Experimentos numéricos

Vamos repetir os experimentos numéricos da Seção 5.4 mas com o precondicionador de dois níveis M_2^{-1} definido em (5.7). Mostramos os resultados na Tabela 5.3. Compare com a Tabela 4.2 (página 65) que mostra os resultados se usamos o gradiente conjugado sem precondicionador, e com a Tabela 5.1 (página 79) que mostra os resultados usando o precondicionador de um nível. Na Tabela 5.3 o tamanho da sobreposição é fixado em $2h$ para todos os casos e o espaço grosso é gerado pelas funções lineares por partes na malha grossa de elementos retangulares com $H = 1/N$. Note que a sobreposição fica menor com $h = 1/(nN)$. Pelo Teorema 33 o número de condição do operador precondicionado $M_2^{-1}A$ depende do número

$$1 + \frac{H}{\delta} = 1 + \frac{n}{2}.$$

Os resultados da Tabela 5.3 são melhores que os resultados esperados.

$n\backslash N$	4	8	16	32	64
4	5(2.93)	9(2.99)	9(2.99)	11(2.96)	10(2.95)
8	5(2.60)	8(2.65)	8(2.65)	9(2.62)	8(2.62)
16	5(2.35)	7(2.39)	7(2.39)	7(2.36)	7(2.36)
32	5(2.19)	6(2.21)	6(2.21)	6(2.19)	6(2.19)
64	5(2.10)	5(2.10)	5(2.10)	5(2.10)	5(2.10)

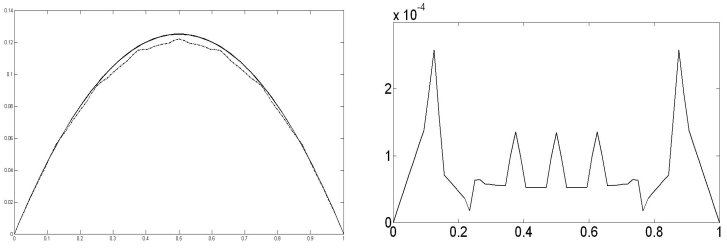
Tabela 5.3: Número de iterações do gradiente conjugado precondicionado (em parênteses estimativa do número de condição) para o problema considerado nesta seção. Consideramos $h = 1/(nN)$, $H = 1/N$ e $\delta = 2h$. Usamos o espaço grosso de funções lineares por partes na malha grossa.

Repetimos o mesmo experimento mas agora usamos uma sobreposição generosa $\delta = nh = 1/N$. Veja os resultados na Tabela 5.4. Os resultados coincidem com a teoria também neste caso pois $1 + \frac{H}{\delta} = 2$. Compare com a Tabela 5.2. Na Figura 5.7 ilustramos a

$n\backslash N = 1/\delta$	4	8	16	32	64
4	2(1.50)	5(2.98)	9(3.68)	12(3.88)	13(3.94)
8	2(1.50)	5(2.98)	9(3.68)	12(3.88)	14(3.93)
16	2(1.50)	5(2.98)	9(3.68)	12(3.88)	14(3.93)
32	2(1.50)	5(2.98)	9(3.68)	12(3.88)	14(3.93)
64	2(1.50)	5(2.98)	9(3.68)	12(3.88)	15(3.93)

Tabela 5.4: Número de iteração e em parênteses estimativa do número de condição do gradiente conjugado precondicionado para o problema desta seção. Consideramos $h = 1/(nN)$, $H = 1/N$ e $\delta = H$. Usamos o espaço grosso de funções lineares por partes na malha grossa.

convergência rápida do gradiente conjugado precondicionado com o precondicionador de dois níveis.



a) b)

Figura 5.7: Convergência rápida do método do gradiente conjugado precondicionado com o precondicionador de dois níveis. Aqui temos $N = 8$ subdomínios ($H = 1/8$) e $n = 8$ elementos em cada subdomínio ($h = 1/(8)(8)$). Veja Tabela 5.3. a) Aqui u_i, é a $i-$ésima iteração do método, $i = 1, \ldots, 8$. Observe que as iterações $u_2, \ldots, u_8$ estão muito próximas e u_8 é a solução com tolerância de 10^{-6}. b) Valor absoluto da diferença entre as iterações u_2 e u_8 que é da ordem 10^{-4}.

5.7 Introdução à análise: como estimar o número de condição?

Nesta seção descrevemos como estimar o número de condição do operador precondicionado $M_2^{-1}A$. Da definição de M_2^{-1} em (5.7) temos

$$M_2^{-1}A = \sum_{i=0}^{N_S} R^{(i)T} \left[A^{(i)}\right]^{-1} R^{(i)} A = \sum_{i=0}^{N_S} T^{(i)}$$

onde para $i = 0, 1, \ldots, N_S$, temos definido

$$T^{(i)} = R^{(i)T} \widetilde{T}^{(i)} \tag{5.11}$$

com

$$\widetilde{T}^{(i)} = \left[A^{(i)}\right]^{-1} R^{(i)} A. \tag{5.12}$$

Dado $u \in V = \mathbb{P}_0^1(\mathcal{T}^h)$, temos que $\widetilde{T}^{(i)} u \in V^{(i)}$ é a solução do sistema linear

$$A^{(i)} \left[\widetilde{T}^{(i)} u\right] = R^{(i)} A u.$$

Se colocamos este sistema linear na sua formulação de Galerkin equivalente obtemos a formulação

$$\mathcal{A}^{(i)}(\widetilde{T}^{(i)}u, v_i) = \mathcal{A}(u, R^{(i)T}v_i) \quad \forall v_i \in V^{(i)} \tag{5.13}$$

que também pode ser usada para definir $\widetilde{T}^{(i)}$. De (5.13) concluímos que $\widetilde{T}^{(i)}u$ tem a forma de uma 'projeção' da função u no espaço local $V^{(i)}$ segundo o produto interno gerado pela forma bilinear $\mathcal{A}^{(i)}$. Observe também que, pela simetria de $\mathcal{A}$ obtemos

$$\begin{aligned}\mathcal{A}(T^{(i)}u, v) &= \mathcal{A}(R^{(i)T}\widetilde{T}^{(i)}u, v) = \mathcal{A}(v, R^{(i)T}\widetilde{T}^{(i)}u) \\ &= \mathcal{A}^{(i)}(\widetilde{T}^{(i)}v, \widetilde{T}^{(i)}u) \quad \text{por (5.13)} \\ &= \mathcal{A}^{(i)}(\widetilde{T}^{(i)}u, \widetilde{T}^{(i)}v).\end{aligned}$$

Ou seja

$$\mathcal{A}(T^{(i)}u, v) = \mathcal{A}^{(i)}(\widetilde{T}^{(i)}u, \widetilde{T}^{(i)}v). \tag{5.14}$$

Usando os mesmos argumentos obtemos $\mathcal{A}(u, T^{(i)}v) = \mathcal{A}^{(i)}(\widetilde{T}^{(i)}u, \widetilde{T}^{(i)}v)$ e portanto

$$\mathcal{A}(u, T^{(i)}v) = \mathcal{A}(T^{(i)}u, v).$$

Concluímos que o operador $T^{(i)}$, $i = 0, \ldots, N$, é simétrico com respeito ao produto interno gerado pela forma bilinear $\mathcal{A}$ e portanto T é também simétrico no produto interno gerado por $\mathcal{A}$.

O seguinte resultado determina um limite inferior para o menor autovalor do operador precondicionado

$$T := M_2^{-1}A = \sum_{i=0}^{N_S} T^{(i)} \tag{5.15}$$

com $T^{(i)}$ definido em (5.11) e $\widetilde{T}^{(i)}$ definido em (5.12) ou (5.13).

Lema 34. *Suponha que existe $C_0 > 0$ tal que para todo $v \in V$, existe a decomposição $v = \sum_{i=0}^{N_S} R^{(i)T}v_i$, com $v_i \in V^{(i)}$, $i = 0, \ldots, N_S$, e*

$$\sum_{i=0}^{N_S} \mathcal{A}^{(i)}(v_i, v_i) \leq C_0^2 \mathcal{A}(v, v). \tag{5.16}$$

Então

$$\lambda_{\min}(T) \geq C_0^{-2}.$$

Observação 35. *O Lema 34 implica que $T = M^{-1}A$ é invertível.*

Observação 36. *A desigualdade (5.16) em forma matricial é: para toda $\boldsymbol{v} \in \mathbb{R}^{N_h^v}$, existe a decomposição $\boldsymbol{v} = \sum_{i=0}^{N_S} R^{(i)T}\boldsymbol{v}_i$, com $\boldsymbol{v}_i \in \mathbb{R}^{N_h^{(i),v}}$, $i = 1, \ldots, N_S$, e*

$$\sum_{i=0}^{N_S} \boldsymbol{v}_i^T A^{(i)} \boldsymbol{v}_i \leq C_0^2 \boldsymbol{v} A \boldsymbol{v}. \tag{5.17}$$

Demonstração. Usando a desigualdade de Cauchy Schwarz

$$\mathcal{A}^{(i)}(z,w) \leq \mathcal{A}^{(i)}(z,z)^{1/2}\mathcal{A}^{(i)}(w,w)^{1/2},$$

como $v = \sum_{i=0}^{N} R^{(i)T} v_i$ e usando (5.13), obtemos que

$$\begin{aligned} \mathcal{A}(v,v) &= \sum_{i=0}^{N_S} \mathcal{A}(v, R^{(i)T}v_i) = \mathcal{A}^{(i)}(\widetilde{T}^{(i)}v, v_i) \\ &\leq \sum_{i=0}^{N_S} \mathcal{A}^{(i)}(\widetilde{T}_i v, \widetilde{T}_i v)^{1/2}\mathcal{A}^{(i)}(v_i,v_i)^{1/2}. \end{aligned}$$

Agora usamos a desigualdade $x^T y \leq ||x||_2||y||_2$, $x, y \in \mathbb{R}^{N_S+1}$ para obter

$$\begin{aligned} \mathcal{A}(v,v) &\leq \left[\sum_{i=0}^{N_S} \mathcal{A}^{(i)}(\widetilde{T}^{(i)}v, \widetilde{T}^{(i)}v)\right]^{1/2} \left[\sum_{i=0}^{N_S} \mathcal{A}^{(i)}(v_i,v_i)\right]^{1/2} \\ &\leq \left[\sum_{i=0}^{N_S} \mathcal{A}(T^{(i)}v, v)\right]^{1/2} \left[\sum_{i=0}^{N_S} \mathcal{A}^{(i)}(v_i,v_i)\right]^{1/2} \quad \text{por (5.14)} \\ &\leq \mathcal{A}(\sum_{i=0}^{N} T^{(i)}v, v)^{1/2} \left[\sum_{i=0}^{N_S} \mathcal{A}^{(i)}(v_i,v_i)\right]^{1/2} \\ &\leq \mathcal{A}(Tv,v)^{1/2} \left[\sum_{i=0}^{N_S} \mathcal{A}^{(i)}(v_i,v_i)\right]^{1/2} \quad \text{por (5.15)} \\ &\leq \mathcal{A}(Tv,v)^{1/2} C_0 \mathcal{A}(v,v)^{1/2} \end{aligned}$$

onde na última desigualdade usamos a hipótese (5.16). Isto implica que $\mathcal{A}(v,v) \leq C_0^2 \mathcal{A}(v, Tv)$ o que da (veja (4.20))

$$\lambda_{\min}(T) = \min_{v \neq 0} \frac{\mathcal{A}(v,Tv)}{\mathcal{A}(v,v)} \geq C_0^{-2}.$$

■

O seguinte lema determina um limite superior para o maior autovalor.

Lema 37 (Lema do limite superior)**.** *Suponha que existe uma constante $\omega > 0$, tal que*

$$\mathcal{A}(R^{(i)T} v_i, R^{(i)T} v_i) \leq \omega \mathcal{A}^{(i)}(v_i, v_i) \quad \forall v_i \in V^{(i)}, \quad 0 \leq i \leq N_S.$$

Suponha também que existem constantes $\mathcal{E}_{ij}$, $1 \leq i,j \leq N_S$, tais que $\forall v_i \in V^{(i)}$, $v_j \in V^{(j)}$

$$\mathcal{A}(R^{(i)T} v_i, R^{(j)T} v_j) \leq \mathcal{E}_{ij} \mathcal{A}(R^{(i)T} v_i, R^{(i)T} v_i)^{1/2} \mathcal{A}(R^{(j)T} v_j, R^{(j)T} v_j)^{1/2}.$$

Então

$$\lambda_{\max}(T) \leq (\rho(\mathcal{E}) + 1)\omega$$

onde $\mathcal{E} = \{\mathcal{E}_{ij}\} \in R^{N_S \times N_S}$ e $\rho(\mathcal{E}) = \lambda_{\max}(\mathcal{E})$.

Demonstração. Usando a primeira hipótese do lema e (5.14) temos que

$$\begin{aligned} \mathcal{A}(T^{(i)}v, T^{(i)}v) &= \mathcal{A}(R^{(i)T}\tilde{T}^{(i)}v, R^{(i)T}\tilde{T}^{(i)}v) \\ &\leq \omega \mathcal{A}^{(i)}(\widetilde{T}^{(i)}v, \widetilde{T}^{(i)}v) = \omega \mathcal{A}(v, T^{(i)}v). \quad (5.18) \end{aligned}$$

Pela primeira hipótese do lema, para $i = 0$,

$$\mathcal{A}(v, T^{(0)}v) \leq \mathcal{A}(v,v)^{1/2} \mathcal{A}(T^{(0)}v, T^{(0)}v) \leq \omega \mathcal{A}(v,v). \quad (5.19)$$

Note que $x^T \mathcal{E} x \leq \rho(\mathcal{E})||x||^2$ para todo $x \in \mathbb{R}^{N_S}$. De (5.18) e da

segunda hipótese do lema

$$\begin{aligned}
&\mathcal{A}\left(\sum_{i=1}^{N_S} T^{(i)}v, \sum_{j=1}^{N_S} v_{j=1}T^{(j)}v\right) = \sum_{i,j=1}^{N_S} \mathcal{A}(T^{(i)}v, T^{(j)}v) \\
\leq &\sum_{i,j=1}^{N_S} \mathcal{E}_{i,j}\mathcal{A}(T^{(i)}v, T^{(i)}v)^{1/2}\mathcal{A}(T^{(j)}v, T^{(j)}v)^{1/2} \\
\leq &\rho(\mathcal{E})\sum_{i=1}^{N_S} \mathcal{A}(T^{(i)}v, T^{(i)}v) \\
\leq &\rho(\mathcal{E})\omega\mathcal{A}\left(v, \sum_{i=1}^{N_S} T_i v\right)
\end{aligned}$$

onde na última desigualdade usamos (5.18). Temos assim

$$\mathcal{A}\left(v, \sum_{i=1}^{N_S} T^{(i)}v\right) \leq \rho(\mathcal{E})\omega\mathcal{A}(v, v)$$

e usando (5.19) obtemos

$$\mathcal{A}(v, Tv) = \mathcal{A}(v, T_0 v) + \mathcal{A}(v, \sum_{i=1}^{N_S} T^{(i)}v) \leq (\rho(\mathcal{E}) + 1)\omega a(v, v),$$

o que finaliza a prova. ∎

Podemos estimar $C_0, \omega, \rho(\mathcal{E})$ e

$$\text{Cond}(M_2^{-1}A) = \kappa(T) \leq \frac{\lambda_{\max}}{\lambda_{\min}} \leq (\rho(\mathcal{E}) + 1)\omega C_0^{-2}.$$

Vamos fazer o resumo destes resultados no seguinte lema.

Lema 38. *Suponhamos,*

1. **Decomposição estável:**. *Existe* $C_0^2 > 0$ *tal que para toda* $v \in V$*, existe a decomposição* $v = \sum_{i=0}^{N_S} R^{(i)T} v_i$*, com* $v_i \in V^{(i)}$*,* $i = 0, \ldots, N_S$*, e*

$$\sum_{i=0}^{N_S} a(v_i, v_i) \leq C_0^2 a(v, v). \tag{5.20}$$

2. **Estabilidade local:** *Existe* $\omega > 0$, *tal que*

$$\mathcal{A}(R^{(i)T}v_i, R^{(i)T}v_i) \leq \omega \mathcal{A}^{(i)}(v_i, v_i) \quad \forall v_i \in V^{(i)}, \quad 0 \leq i \leq N_s.$$

3. **Desigualdades fortes de Cauchy :** *Existem* $\mathcal{E}_{ij}$, $1 \leq i, j \leq N_S$, *tais que para toda* $v_i \in V^{(i)}$, $v_j \in V^{(j)}$

$$\mathcal{A}(R^{(i)T}v_i, R^{(j)}v_j) \leq \mathcal{E}_{ij}\mathcal{A}(R^{(i)T}v_i, R^{(i)T}v_i)^{1/2}\mathcal{A}(R^{(j)T}v_j, R^{(j)T}v_j)^{1/2}.$$

Então,

$$Cond(M_2^{-1}A) = Cond(T) \leq (\rho(\mathcal{E}) + 1)\omega C_0^{-2}. \tag{5.21}$$

Agora aplicamos esta teoria ao método aditivo de dois níveis com sobreposição. Para isto, temos que verificar as três hipóteses do Lema 38. Para simplificar a apresentação consideramos somente o caso da forma bilinear $\mathcal{A}$ definida em (2.6) associada a equação diferencial de Laplace na Seção 2.2.2 com condição de contorno de Dirichlet. Provaremos que a condição do operador precondicionado usando o precondicionador aditivo de dois níveis pode ser estimada por $\text{Cond}(M_2^{-1}A) \leq C\left[1 + \frac{H}{\delta}\right]$ onde C é independente dos parâmetros h e H.
Uma análise similar pode ser aplicada para a forma bilinear (2.13) quando o coeficiente κ é limitado e outras condição de contorno. A constante C pode depender do contraste do coeficiente neste caso.

Verificamos cada uma das três hipóteses do Lema 38. A decomposição estável é a mais difícil de verificar e será apresentada depois.

Estabilidade local: De (6.2) concluímos que

$$\mathcal{A}(R^{(i)T}v_i, R^{(i)T}v_i) = \boldsymbol{v}_i R^{(i)T} A R^{(i)} \boldsymbol{v}_i = \boldsymbol{v}_i A^{(i)} \boldsymbol{v}_i = \mathcal{A}^{(i)}(v_i, v_i)$$

donde $\omega = 1$. Observamos que ω pode ser difícil de obter quando no lugar da matriz local exata $A^{(i)}$ usamos uma aproximação $\tilde{A}^{(i)} \approx A^{(i)}$.

Desigualdades fortes de Cauchy: Da desigualdade usual de Cauchy-Schwarz concluímos que $\mathcal{E}_{ij} \leq 1$ em todos os casos. Como

para $|i-j|>2$ temos $D_i' \cap D_j' = \emptyset$, concluímos então que $\mathcal{E}_{ij}=0$ quando $|i-j|>2$ pois suporte$(v_i) \subset D_i$ e suporte$(v_j) \subset D_j$ da

$$\mathcal{A}(R^{(i)T}v_i, R^{(j)T}v_j) = 0.$$

Para a matriz $\mathcal{E}$ podemos tomar a matriz tridiagonal

$$\mathcal{E} = \begin{bmatrix} 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 1 & 0 & \dots & 0 \\ 0 & 1 & 1 & 1 & \dots & 0 \\ \dots & & & & \dots & \\ 0 & 0 & 0 & & 1 & 1 \end{bmatrix}.$$

É fácil ver que o maior autovalor desta matriz tridiagonal é no máximo três, isto é, $\rho(\mathcal{E}) \leq 3$. Este argumento é um caso particular de um argumento mais geral conhecido como *técnica de colorir*, veja [31].

Decomposição estável: Dado $v \in V$, temos que achar $v_i \in V^{(i)}$, $i=0,...,N_S$, tais que

$$\sum_{i=0}^{N_S} \mathcal{A}^{(i)}(v_i, v_i) \leq C_0^2 \mathcal{A}(v,v). \tag{5.22}$$

A forma matricial da desigualdade acima é (5.17). Antes de apresentar esta decomposição introduzimos um pouco de notação e alguns lemas necessários. Dado um vértice y_j da malha grossa, defina

$$W_{y_j} = \text{suporte de } \Phi_j = D_j \cup D_{j+1}.$$

Seja $v \in V = \mathbb{P}^1(\mathcal{T}^h)$. Defina a interpolação de v na malha grossa por

$$I^H v = v_0 = \sum_{j=2}^{N_H^v - 1} \bar{v}_j \Phi_j^H$$

onde $\bar{v}_j$ é a meia na vizinhança de y_j da função v, isto é, $\bar{v}_j = \frac{1}{|W_{y_j}|}\int_{W_{y_j}} v$, $j=2,\dots,N_H^v-1$, onde N_H^v é o número de vértices da malha grossa $\mathcal{T}^H$. Note que $I^H v(a) = I^H v(b) = 0$ e $I^H v \in V$. Temos o seguinte lema que da as propriedades de aproximação e estabilidade da interpolação I^H, veja [31, 7, 8, 24].

Lema 39. *Para a interpolação definida acima temos que para todo* $v \in V$,

$$||v - R^{(0)T} v_0||_{L^2(D)} \leq CH\mathcal{A}(v,v) \tag{5.23}$$

$$\mathcal{A}^{(0)}(v_0, v_0) \leq C\mathcal{A}(v,v). \tag{5.24}$$

Também vamos usar a interpolação linear na triangulação $\mathcal{T}^h$. Seja f uma função contínua, denotamos por $I^h(f)$ a sua interpolação linear nos vértices de $\mathcal{T}^h$. Vemos que $I^h(f)$ é a função de elementos finitos com valores $I^h(f)(x_i) = f(x_i)$, para todo vértice x_i da malha. Temos o seguinte lema.

Lema 40. *Se* f *é contínua em* $D = (a,b)$ *e diferenciável por partes na malha* $\mathcal{T}^h$ *então*

$$\int_D |(I^h(f))'|^2 \leq \int_D |f'|^2.$$

Introduzimos também uma partição da unidade $\{\theta_i\}_{i=1}^{N_S}$ subordinada à decomposição $\{D_i'\}_{i=1}^{N_S}$, com

$$\left.\begin{array}{l} 0 \leq \theta_i \leq 1, \quad \text{suporte}(\theta_i) \subset D_i' \\ |\theta_i'| \leq \frac{C}{\delta}, \quad i = 1, \ldots, N_S, \text{ e} \\ \sum_{i=1}^{N_S} \theta_i(x) = 1 \end{array}\right\} \tag{5.25}$$

Definimos $v_0 = I^H v$. A interpolação na malha grossa v_0 é a componente grossa da decomposição. Seja $z = v - R^{(0)T} v_0 \in V$. Definimos $v_i = I^h(\theta_i z)$, $i = 1, \ldots, N_S$. Como $\{\theta_i\}_{i=1}^{N_S}$ é uma partição da unidade subordinada à decomposição $\{D_i'\}_{i=1}^{N_S}$ temos $v_i \in V^{(i)}$ com $V^{(i)}$ definido em (5.1), $i = 1, \ldots, N_S$, e vale

$$\begin{aligned} \sum_{i=0}^{N_S} R^{(i)T} v_i &= v_0 + \sum_{i=1}^{N_S} I_h(\theta_i(v - v_0)) \\ &= v_0 + I_h\left((\sum_{i=1}^{N_S} \theta_i)(v - v_0)\right) \\ &= v_0 + I^h(v - v_0) = v_0 + (v - v_0) = v. \end{aligned}$$

Desta maneira achamos uma decomposição para v. Temos ainda que provar a estabilidade desta decomposição, isto é, a equação (5.22).

Cada termo em (5.22) pode ser estimado usando o Lema 40 e a regra para a derivada do produto como segue

$$\begin{aligned}\mathcal{A}^{(i)}(v_i,v_i) &= \int_{D_i'} |(I^h(\theta_i z)'|^2 \\ &\leq \int_{D_i'} |(\theta_i z)')|^2 \\ &\leq 2\int_{D_i'} \theta_i^2 |z'|^2 + 2\int_{D_i'} |\theta_i'|^2 |z|^2 \\ &\leq 2\int_{D_i'} |z'|^2 + \frac{2C}{\delta^2}\int_{D_i'\setminus\overline{D_i}} |z|^2.\end{aligned}$$

Somando nos subdomínios obtemos

$$\sum_{i=1}^{N_S} \mathcal{A}^{(i)}(v_i,v_i) \leq 2\sum_{i=1}^{N_S}\int_{D_i'} |z'|^2 + \frac{2C}{\delta^2}\sum_{i=1}^{N_S}\int_{D_i'\setminus\overline{D_i}} |z|^2. \quad (5.26)$$

Para estimar o primeiro termo na última linha acima usamos que $z' = v' - (R^{(0)T}v_0)'$ e obtemos

$$\int_{D_i'} |z'|^2 \leq 2\int_{D_i'} |v'|^2 + 2\int_{D_i'} |(R^{(0)T}v_0)'|^2,$$

donde

$$\sum_{i=1}^{N_S}\int_{D_i'} |z'|^2 \leq 2\int_D |v'|^2 + 2\int_D |R^{(0)}v_0'|^2 = 2\mathcal{A}(v,v) + 2\mathcal{A}^{(0)}(v_0,v_0) \quad (5.27)$$

Na última desigualdade, D_i' sobreposiciona somente D_{i-1}' e D_{i+1}'.

Para estimar o segundo termo em (5.26) lembramos que $v_0 = I^H v$ e com ajuda do Lema 39 obtemos

$$\sum_{i=1}^{N_S}\int_{D_i'\setminus\overline{D_i}} |z|^2 \leq \sum_{i=1}^{N_S}\int_{D_i'} |z|^2 \leq CH^2\mathcal{A}(v,v). \quad (5.28)$$

Substituindo (5.27) e (5.28) em (5.26) e adicionando o termo de ordem zero (5.24) obtemos

$$\sum_{i=0}^{N_S} \mathcal{A}^{(i)}(v_i, v_i) \quad \leq \quad C\left[1 + \frac{H^2}{\delta^2}\right] \mathcal{A}(v, v).$$

Isto finaliza a prova de (5.22) com

$$C_0^2 = C\left[1 + \frac{H^2}{\delta^2}\right].$$

Finalmente, juntando as três hipóteses verificadas temos pelo Lema 38 que existe uma constante C tal que

$$\text{Cond}(M_1^{-1}A) = \kappa(T) \leq (\rho(\mathcal{E}) + 1)\omega C_0^{-2} \leq C\left[1 + \frac{H^2}{\delta^2}\right].$$

Provamos então o seguinte resultado.

Lema 41. *Assumindo as hipóteses acima na decomposição de D temos que*

$$Cond(M_2^{-1}A) \leq C\left[1 + \frac{H^2}{\delta^2}\right].$$

A prova do lema anterior pode ser melhorada para obter o seguinte resultado.

Teorema 42. *Assumindo as hipóteses acima na decomposição de D temos que*

$$Cond(M_2^{-1}A) \leq C\left[1 + \frac{H}{\delta}\right].$$

A idéia para provar o Teorema 41 é modificar a prova do Lema 41. No lugar de usar a estimativa trivial (5.28) usamos a seguinte estimativa que será provada no Lema 43 a seguir. A estimativa é

$$\int_{D_i' \setminus D_i} |z|^2 \leq C\left[\frac{\delta}{H}\int_{D_i'} |z|^2 + \delta H \int_{D_i'} |z'|^2\right] \tag{5.29}$$

que somando em i, usando (5.27) e o Lema 39 para $z = v - R^{(0)T}v$,

$$\begin{aligned}\sum_{i=1}^{N_S} \int_{D_i' \setminus \overline{D_i}} |z|^2 \\ &\leq C\left[\frac{\delta}{H}\int_D |z|^2 + \delta H\left(\mathcal{A}(v,v) + \mathcal{A}^{(0)}(v_0,v_0)\right)\right] \\ &\leq \tilde{C}\delta H\left[\mathcal{A}(v,v) + \mathcal{A}^{(0)}(v_0,v_0)\right]. \qquad (5.30)\end{aligned}$$

Substituindo (5.27) e (5.30) em (5.26) e somando o termo de ordem zero (5.24) obtemos

$$\sum_{i=0}^{N_S} \mathcal{A}^{(i)}(v_i,v_i) \leq C\left[1 + \frac{H}{\delta}\right]\mathcal{A}(v,v).$$

Para fechar a ideia da prova do Teorema 41 apresentamos uma ideia da prova da estimativa usada.

Lema 43. *Para todo $z \in H_0^1(D)$ vale*

$$\int_{D_i' \setminus D_i} |z|^2 \leq C\left[\frac{\delta}{H}\int_{D_i'} |z|^2 + \delta H \int_{D_i'} |z'|^2\right]. \qquad (5.31)$$

Se a derivada z' é integrável (que é o caso) podemos escrever

$$z(0) \quad = \quad z(x) - \int_0^x z'(y)dy \quad x \in D_i'$$

e usando uma desigualdade de Cauchy-Schwarz na última integral,

$$|z(0)|^2 \quad \leq \quad C\left[|z(x)|^2 + H\int_{D_i'} |z'|^2\right],$$

onde utilizamos que o diâmetro de D_i' é da ordem H, $i = 1, \ldots, N_S$. Tomando integrais em D_i' nos dois lados obtemos

$$|z(0)|^2 \quad \leq \quad C\left[\frac{1}{H}\int_{D_i'} |z|^2 + H\int_{D_i'} |z'|^2\right].$$

Similarmente

$$|z(x)|^2 \quad \leq \quad C\left[\frac{1}{H}\int_{D_i'}|z|^2 + H\int_{D_i'}|z'|^2\right]$$

e tomando integrais em $D_i' \setminus D_i$ e lembrando que o diâmetro de $D_i' \setminus D_i$ é δ obtemos o enunciado do Lema 43.

Capítulo 6

Métodos com superposição em dimensão dois

Neste capítulo estudamos precondicionadores de decomposição de domínio em duas dimensões. Consideramos especificamente o sistema linear obtido usando elementos finitos em duas dimensões; veja o Capítulo 3. O leitor interessado pode consultar [31, 24, 30] e as referências ali citadas.

6.1 Decomposição com e sem sobreposição

Trabalhamos num domínio poligonal conexo $D \subset \mathbb{R}^2$ que tem associada uma triangulação $\mathcal{T}^h$ com parâmetro $h > 0$ e elementos $\{K\}_{i=1}^{N_h^e}$. Assumimos que a triangulação $\mathcal{T}^h$ é conforme e quase-uniforme e consideramos a formulação fraca da Seção 3.3.2.

Introduzimos uma partição do domínio D da equação diferencial

em subdomínios poligonais disjuntos $\{D_i\}_{i=1}^{N_S}$ com

$$\bigcup_{i=1}^{N_S} \overline{D}_i = \overline{D}, \quad D_i \cap D_j = \emptyset, \text{ for } i \neq j.$$

Aqui N_S é o número de subdomínios. Esta decomposição é dita sem sobreposição. Vide Figura 6.1. Com a decomposição sem so-

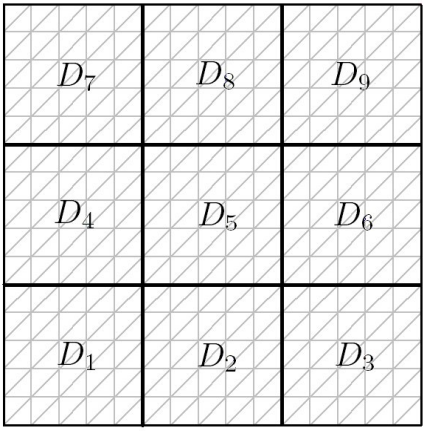


Figura 6.1: Exemplo de uma decomposição sem sobreposição do quadrado $D = (0,1) \times (0,1)$.

breposição construímos uma nova cobertura $\{D_i'\}_{i=1}^{N_S'}$ do domínio D com sobreposição δ definindo

$$D_i' = \{x \in D \quad : \quad d(x,y) < \delta, \text{ para algum } y \in D\}$$

onde $d(\cdot,\cdot)$ denota alguma função distancia em $\mathbb{R}^2$. Assumimos também que cada subdomínio D_i', $i = 1, \ldots, N_S$, é a união de elementos da triangulação, isto é, para cada $i = 1, \ldots, N_S$,

$$\overline{D'}_i = \bigcup_{i:K_i \subset D_i'} \overline{K}_i.$$

Vide Figura 6.1. Usaremos a notação

- N_h^e número de elementos da triangulação
- N_h^v número de vértices da triangulação
- $N_h^{(i),e}$ número de elementos do subdomínio D_i
- $N_h^{(i),I}$ número de vértices interiores no subdomínio D_i'.

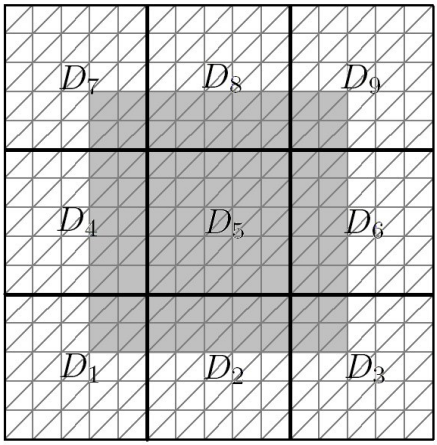


Figura 6.2: Exemplo de uma decomposição com sobreposição do quadrado $D = (0,1) \times (0,1)$. Esta decomposição foi construída a partir da decomposição da Figura 6.1. Somente mostramos o subdomínio D_5' (região cinza). Analogamente são construídos os outros subdomínios. Neste exemplo temos $\delta = 2h$.

Usando a decomposição com sobreposição $\{D_i'\}_{i=1}^{N_S}$ do domínio D, vamos construir um precondicionador de decomposição de domínios. Como no caso de uma dimensão, os ingredientes principais de um precondicionador de decomposição de domínios são: os espaços locais, operadores de restrição e extensão, em cada espaço local teremos que definir uma aproximação da forma bilinear da formulação fraca e um (ou ate vários) espaços grossos.

6.2 Espaços locais, operadores de restrição e extensão

No Capítulo 3 estudamos o espaço de elementos finitos de funções lineares por partes. O espaço de elementos finitos usado para aproximar (3.13) é o espaço $V := \mathbb{P}^1(\mathcal{T}^h)$ onde $\mathcal{T}^h$ é a malha do domínio D. O espaço V é o espaço *global* de nosso método de decomposição de domínios. Dada uma decomposição com sobreposição δ, $\{D'_i\}_{i=1}^{N_S}$, definimos os espaços locais por

$$V^{(i)} = V^{(i)}(D'_i) = \mathbb{P}^1_0(D'_i),$$

isto é, o espaço local é somente a restrição do espaço global V aos vértices interiores do subdomínio D'_i, $i = 1, \ldots, N_S$.

Definimos a matriz de restrição $R^{(i)}$ de dimensão $N_h^{(i),I} \times N_h^v$ com entradas nulas menos nas posições (ℓ, j) onde o índice j corresponde ao vértice $x_j \in D'_i$. Note que as matrizes $R^{(i)}$, $i = 1, \ldots, N_S$, são análogas as matrizes de restrição definidas no Lema 19.

Para cada D'_i, $i = 1, \ldots, N_S$, definimos $\mathcal{A}^{(i)}$, a restrição da forma bilinear $\mathcal{A}$ ao subdomínio $V^{(i)}$ por,

$$\mathcal{A}^{(i)}(u_i, v_i) = \int_{D'_i} \kappa(x) u'_i(x) v'_i(x) dx, \quad u_i, v_i \in V^{(i)}. \tag{6.1}$$

Seja $A^{(i)}$ a matriz $N_h^{(i),I} \times N_h^{(i),I}$ que representa a forma bilinear local $\mathcal{A}^{(i)}$. Observe que

$$\boldsymbol{u}_i^T A^{(i)} \boldsymbol{v}_i = \mathcal{A}^{(i)}(u_i, v_i) = \mathcal{A}(R^{(i)T} u_i, R^{(i)T} v_i) = (R^{(i)T} \boldsymbol{u}_i)^T A (R^{(i)T} \boldsymbol{v}_i)$$

onde $R^{(i)T} w_i$ é a função de elementos finitos com representação vetorial $R^{(i)} \boldsymbol{w}_i \in \mathbb{R}^{N_h^v}$ e A é matriz global, isto é, a representação matricial da forma bilinear global $\mathcal{A}$. Concluímos que

$$A^{(i)} = R^{(i)} A R^{(i)T}. \tag{6.2}$$

A matriz local $A^{(i)}$ é o bloco diagonal da matriz A correspondente aos índices associados aos vértices interiores ao subdomínio D'_i, $i = 1, \ldots, N_S$. Veja a ilustração da Figura 5.3.

6.3 Precondicionador aditivo de um nível

Com a notação introduzida na Seção 6.2 definimos o precondicionador aditivo de um nível M_1^{-1} por

$$M_1^{-1} = \sum_{i=1}^{N_S} R^{(i)T} \left[A^{(i)}\right]^{-1} R^{(i)}. \tag{6.3}$$

Esta é exatamente a mesma fórmula do caso unidimensional (5.7). Esta fórmula é geral e independente da dimensão espacial do problema. Para poder aplicar a fórmula (6.3) somente precisamos da definição das matrizes locais e das matrizes de restrição e extensão.

Lembre que para usar o precondicionador acima no método do gradiente conjugado precondicionado temos que poder calcular $M_1^{-1}q$ dado um vetor q do tamanho apropriado. Temos que levar em conta as mesmas observações que para o caso unidimensional. Podemos escrever

$$M_1^{-1}q = \sum_{i=1}^{N_S} R^{(i)T} \left[A^{(i)}\right]^{-1} R^{(i)}q = \sum_{i=1}^{N_S} R^{(i)T}u_i$$

onde definimos $u_i = \left[A^{(i)}\right]^{-1} R^{(i)}q$, $i = 1, \ldots, N_S$. Cada parcela desta soma pode ser calculada em paralelo. Para calcular a i-ésima parcela $u_i = \left[A^{(i)}\right]^{-1}$ resolvemos o sistema linear local

$$A^{(i)}u_i = R^{(i)}q.$$

Note que a dimensão deste sistema linear é pequena quando comparada com a dimensão do sistema linear global. Nas aplicações praticas este sistema linear não precisa ser resolvido com precisão total, uma aproximação pode ser usada.

O número de iterações do gradiente conjugado precondicionado com precondicionador M_1^{-1} acima depende do número de condição do operador precondicionado $M_1^{-1}A$. Temos a seguinte estimativa para o número de condição de $M_1^{-1}A$.

Teorema 44. *Suponha que a malha $\mathcal{T}^h$ é conforme e quase-uniforme. Considere o precondicionador M_1^{-1} definido em (6.3). A matriz A é a matriz da forma bilinear definida em (3.16) com o coeficiente κ satisfazendo (3.14). Temos que existe uma constante C independente h e de H tal que*

$$Cond(M_1^{-1}A) \leq C\left(1 + \frac{1}{\delta H}\right)$$

onde $H = \max_{1\leq i\leq N_s} diâmetro(D_i)$ e δ é o parâmetro da decomposição com superposição $\{D_i'\}$. A constante C pode depender do contraste do coeficiente κ.

6.4 Experimentos numéricos

Nesta seção consideramos a equação de Laplace $\Delta u = -1$ em $D = (0,1)\times(0,1)$ com condição de Dirichlet. Calculamos a solução usando o método do gradiente conjugado precondicionado com o precondicionador aditivo de um nível M_1^{-1} definido em (6.6). Usamos uma malha uniforme como na Seção 4.3. Dividimos o domínio D em $N\times N$ subdomínios quadrados e usamos uma malha triangular baseada em $n\times n$ quadrados dentro de cada subdomínio. Os resultados aparecem na Tabela 6.1. Compare com os resultados da Tabela 4.4 na página 67 usando o gradiente conjugado sem precondicionador. Por exemplo, para $h = \sqrt{2}/128$ temos 197 iterações no gradiente conjugado sem precondicionador e um número de condição de 6639.53. Na Tabela 6.1 para a mesma malha e o mesmo valor de h podemos calcular a solução em 32 iterações no gradiente conjugado precondicionado se usamos $16 \cdot 16 = 256$ subdomínios ($N = 16$) e dentro de cada subdomínio uma malha triangular baseada em 8×8 quadrados ($n = 8$). O número de condição é 351.47. Neste caso, em cada uma das 32 iterações do gradiente conjugado precondicionado temos que resolver 256 sistemas lineares pequenos de tamanho 81×81 (considerando a sobreposição $\delta = 2h$). Note que a dimensão do sistema linear original é 16641×16641. Podemos também obter a solução em 25 iterações usando $N = 8$ e $n = 16$.

$n \backslash N$	2	4	8	16
2	1 (1.00)	7 (4.27)	9 (13.61)	15 (51.56)
4	5 (4.89)	8 (10.60)	13 (36.61)	23 (141.94)
8	7 (8.52)	11 (24.45)	18 (89.69)	32 (351.47)
16	8 (16.36)	14 (51.82)	25 (194.94)	47 (768.03)

Tabela 6.1: Número de iterações do gradiente conjugado precondicionado e parênteses em estimativa do número de condição para o problema considerado nesta seção. Consideramos $h = 1/(nN)$, $H = 1/N$ e $\delta = 2h$.

Na Tabela 6.1 o tamanho da sobreposição é fixado em $2h$ para todos os casos. Pelo Teorema 32 o número de condição do operador precondicionado $M_1^{-1}A$ depende do número

$$1 + \frac{1}{\delta H} = 1 + \frac{1}{2}nN^2.$$

Os resultados da Tabela 6.1 mostram-se em concordância com esta afirmação. Nas linhas o número de condição duplica-se (aproximadamente) e nas colunas vemos um incremento com um fator aproximadamente quatro.

Repetimos o mesmo experimento com a sobreposição $\delta = nh$. Mostramos os resultados na Tabela 6.2. Nesta caso $1 + \frac{1}{\delta H} = 1 + N^2$. Observamos resultados em concordância com a teoria.

$n\backslash N$	2	4	8	16
2	1 (1.00)	7 (4.27)	9 (13.61)	15 (51.56)
4	1 (1.00)	7 (4.13)	9 (13.05)	15 (49.35)
8	1 (1.00)	7 (4.09)	10 (12.87)	15 (48.66)
16	1 (1.00)	8 (4.07)	10 (12.81)	16 (48.44)

Tabela 6.2: Número de iterações do gradiente conjugado precondicionado e em parênteses estimativa do número de condição) para o problema considerado nesta seção. Aqui $h = 1/(nN)$ e $H = 1/N$ e o fixamos $\delta = nh = H$.

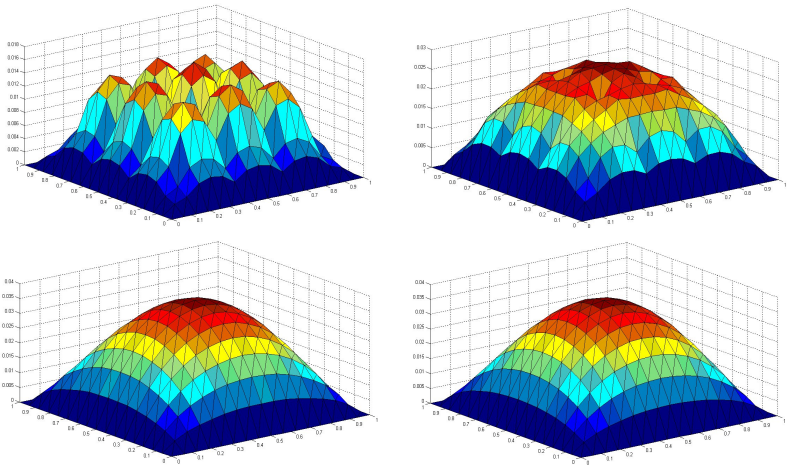


Figura 6.3: Convergência rápida do método do gradiente conjugado precondicionado com o precondicionador de um nível. Aqui u_i, é a $i-$ésima iteração do método, $i = 1, 2, 4, 8$ (de esquerda para direita e de cima para baixo). Neste exemplo $n = N = 4$. Veja a Tabela 6.1. Observe que u_8 é a solução com tolerância de 10^{-6}.

6.5 Precondicionador de dois níveis

No domínio D introduzimos uma triangulação grossa $\mathcal{T}^H$ adicional. Supomos que $H > h$. A nova triangulação pode, em geral, ser independente da triangulação mais fina, da decomposição $\{D_i\}_{i=1}^{N_S}$ e da decomposição $\{D_i'\}_{i=1}^{N_S}$. Assumimos que a triangulação grossa é dada por $\mathcal{T}^H = \{D_i\}_{i=1}^{N_S}$, isto é, a malha grossa coincide com a decomposição original da qual construímos o precondicionador de um nível.

Dada uma triangulação grossa consideramos um espaço de elementos finitos grossos baseado nos vértices de $\mathcal{T}^H$, isto é, um espaço de elementos finitos da forma

$$V^H = \text{span}\{\Phi_j \in V^h \ : \ i = 0, 1, \ldots, N_H^v\}$$

onde N_H^v é a quantidade de vértices da malha grossa $\mathcal{T}^H$ e as funções base grossa serão definidas logo. Como $\Phi_j \in V^h$ pode-se escrever

$$\Phi_j = \sum_{i=1}^{N_h^v} \Phi_j(x_i)\phi_i \text{ ou } \quad \boldsymbol{\Phi}_j = [\Phi_j(x_1), \ldots, \Phi_j(x_{N_h^v})]^T$$

onde ϕ_i, $i = 1, \ldots, N_h^v$, são as funções base da malha fina $\mathcal{T}^h$ definidas em (3.23). Denotemos por $R^{(0)}$ a matriz

$$R^{(0)T} = [\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_{N_H^v}]$$

e definamos a matriz grossa $A^{(0)}$ como a matriz global A na base grossa,

$$A^{(0)} = R^{(0)} A R^{(0)T}. \tag{6.4}$$

Note que se $u_0, v_0 \in V^{(0)}$ temos

$$\boldsymbol{u}_0 A^{(0)} \boldsymbol{v}_0 = (R^{(0)T}\boldsymbol{u}_0)^T A (R^{(0)T}\boldsymbol{v}_0) = \mathcal{A}(R^{(0)T}u_0, R^{(0)T}v_0)$$

onde $R^{(0)T}w_0 \in V$ é a função de elementos finitos com representação $R^{(0)}\boldsymbol{w}_0 \in \mathbb{R}^{N_h^v}$. Concluímos que $A^{(0)}$ é a representação matricial da forma bilinear $\mathcal{A}^{(0)}$ definida como a restrição da forma bilinear $\mathcal{A}$ ao subespaço $V^{(0)} \subset V$,

$$\mathcal{A}^{(0)}(u_0, v_0) = \mathcal{A}(R^{(0)T}u_0, R^{(0)T}v_0). \tag{6.5}$$

Definimos o precondicionador aditivo de dois níveis M_2^{-1}

$$\begin{aligned} M_2^{-1} &= R^{(0)T}\left[A^{(0)}\right]^{-1} R^{(0)} + \sum_{i=1}^{N_S} R^{(i)T}\left[A^{(i)}\right]^{-1} R^{(i)} \quad (6.6) \\ &= R^{(0)T}\left[A^{(0)}\right]^{-1} R^{(0)} + M_1^{-1}, \end{aligned}$$

onde M^{-1} é o precondicionador aditivo de um nível definido em (6.3). Para aplicar M_2^{-1} temos que aplicar M_1^{-1} como antes e aplicar o termo $R^{(0)T}\left[A^{(0)}\right]^{-1} R^{(0)}$. Neste caso podemos exprimir

$$M_2^{-1}q = \sum_{i=0}^{N_S} R^{(i)T}\left[A^{(i)}\right]^{-1} R^{(i)}q = \sum_{i=0}^{N_S} R^{(i)T}u_i$$

onde definimos $u_i = \left[A^{(i)}\right]^{-1} R^{(i)}q$, $i = 0, 1, \ldots, N_S$. Como no caso de M_1, cada parcela da soma na definição de M_2^{-1} pode ser calculada em paralelo. Para calcular a 0-ésima parcela $u_0 = \left[A^{(0)}\right]^{-1} R^{(0)}q$ no lugar de aplicar a inversa da matriz grossa $A^{(0)}$ resolvemos o sistema linear

$$A^{(0)}u_0 = R^{(0)}q$$

que como sabemos equivale a solução da mesma equação diferencial no domínio D no espaço de elementos finitos $V^{(0)}$ baseado na triangulação grossa $\mathcal{T}^H$. A dimensão deste sistema linear é pequena quando comparada com a dimensão do sistema linear global. Em particular a dimensão do problema grosso é da ordem do número de vértices na malha grossa $\mathcal{T}^H$.

6.5.1 Espaços grossos

Nesta seção vamos a descrever as funções bases $\{\Phi_j\}$ que definem o espaço grosso $V^{(0)}$. As funções grossas devem ser escolhidas de tal forma que gerem funções com comportamento similar à solução do problema considerado. Existem muitas escolhas possíveis para as funções bases grossas. Vamos mencionar unicamente duas escolhas para as funções base grossas.

Funções base lineares por partes na malha grossa

Podemos escolher Φ_j como sendo as função base chapéu na malha grossa $\mathcal{T}^H = \{D_i\}_{i=1}^{N_S}$. A vantagem desta escolha é o baixo custo computacional requerido para construir as funções bases. Note que existe uma função por vértice da malha grossa. Sejam $y_0, \ldots, y_{N_H^v}$ os vértices da triangulação grossa $\mathcal{T}^H$. Para simplificar assumimos que os elementos da malha grossa são retângulos ou triângulos. Para os elementos triangulares temos

$$\Phi_j(x) = \begin{cases} 1, & \text{se } x = y_j, \text{ (1 no vértice } y_j) \\ 0, & \text{se } x = y_k, k \neq j, \text{(0 nos outros vértices)} \\ \text{extensão linear}, & \text{se } x \text{ não é vértice de } \mathcal{T}^H. \end{cases} \tag{6.7}$$

e para os elementos retangulares,

$$\Phi_j(x) = \begin{cases} 1, & \text{se } x = y_j, \\ 0, & \text{se } x = y_k, k \neq j, \\ \text{polinômio linear} & \\ \text{de grau um em} & \text{se } x \text{ não é vértice de } \mathcal{T}^H. \\ \text{cada variável}, & \end{cases} \tag{6.8}$$

Lembramos que um polinômio de grau um em cada variável é da forma $p(x, y) = ax + bx + cxy + d$. Esta é a versão numa malha retangular do métodos dos elementos finitos lineares por partes; veja [7]. Na Figura 6.4 temos uma destas funções base grossas.

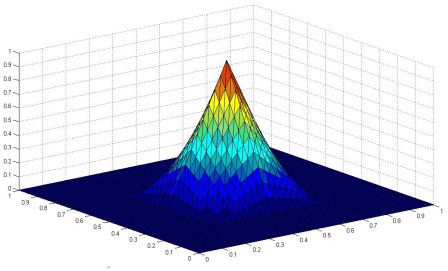


Figura 6.4: Exemplo de uma função base grossa numa partição em subdomínios retangulares.

Funções de elementos finitos multi-escala

No lugar de usar funções lineares por partes podemos usar funções que em cada elemento da malha grossa sejam uma solução da (ou de uma) equação diferencial. No lugar estender linearmente dentro de cada elemento da malha grossa para obter Φ_j, podemos usar uma *extensão harmônica* Φ_j^{MS} definida por

$$\Phi_j^{MS}(x) = \begin{cases} 1, & \text{se } x = y_j, \\ 0, & \text{se } x = y_k, k \neq j, \\ \text{extensão} & \text{se } x \text{ pertence as arestas de} \\ \text{LINEAR}, & \text{algum elemento.} \\ & \\ \text{extensão} & \text{se } x \in K \text{ é ponto interior} \\ \text{harmônica}, & \text{de algum elemento } \mathcal{T}^H \end{cases} . \tag{6.9}$$

Aqui extensão harmônica no interior do elemento quer dizer que Φ_j^{MS} satisfaz a equação

$$\begin{aligned} &\mathcal{A}(\Phi_j^{MS}, v) = 0 \quad \forall v \in \mathbb{P}_0^1(\mathcal{T}^h|_{D_\ell}) \\ &\Phi_j^{MS}(x) = \Phi_j(x) \text{ para } x \in \partial D_\ell \end{aligned}$$

para todos os elementos do suporte de Φ_j e onde $\mathcal{T}^h|_{D_\ell}$ denota a restrição da malha fina ao elemento da malha grossa D_ℓ. No caso da forma bilinear $\mathcal{A}$ definida em (3.16) temo que Φ_j^{MS} é a aproximação de elementos finitos da equação,

$$\begin{aligned} &-\text{div}(\kappa(x)\nabla\Phi_j^{MS}(x)) = 0 \quad x \in D_\ell. \\ &\Phi_j^{MS}(x) = \Phi_j(x) \text{ para } x \in \partial D_\ell. \end{aligned}$$

Quando $\kappa(x) = 1$ para todo $x \in D_\ell$ temos que $\Phi_j^{MS} = \Phi_j$ em D_ℓ. Na Figura 6.5 mostramos um exemplo de função base grossas de elementos finitos multi-escalas. Neste exemplo o coeficiente é apresentado na Figura 6.6.

O seguinte resultado fornece uma estimativa para o número de condição do operador precondicionado $M_2^{-1}A$ com M_2^{-1} definido em (6.6) se usamos qualquer um dois espaços grossos descritos acima.

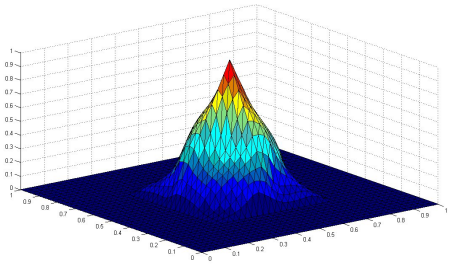


Figura 6.5: Função base grossa de elementos finitos multi-escala numa partição em subdomínios retangulares calculada com o coeficiente na Figura 6.6.

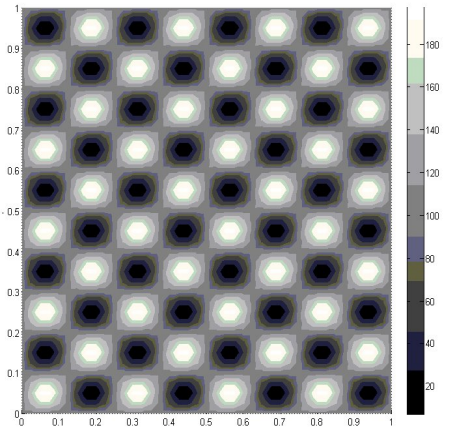


Figura 6.6: Coeficiente oscilatório usado para calcular a função base grossa da Figura 6.5

Teorema 45. *Assuma que a triangulação $\mathcal{T}^h$ é conforme e quasi-uniforme. Considere precondicionador M_2^{-1} definido em (6.6) com o espaço grosso de funções lineares por partes ou o espaço grosso das funções de elementos finitos multi-escala. Seja A a matriz da forma bilinear (3.16) com κ como em (3.1) satisfazendo (3.14). Temos que o número de condição do operador precondicionado é*

$$Cond(M_2^{-1}A) \leq C\frac{\kappa_{\max}}{\kappa_{\min}}\left[1+\frac{H}{\delta}\right],$$

onde a constante C é independente de h.

6.6 Experimentos numéricos

Consideramos a equação de Laplace $\Delta u = -1$ com condição de Dirichlet em $D = (0,1)\times(0,1)$. Calculamos a solução usando o método do gradiente conjugado precondicionado com o precondicionador aditivo de dois níveis. Usamos uma malha uniforme como na Seção 4.3. Dividimos o domínio D em $N \times N$ subdomínios quadrados ($H = 1/N$) e dentro de cada subdomínio consideramos uma malha triangular baseada em $n \times n$ quadrados ($h = \sqrt{2}/(nN)$). A malha grossa coincide com esta decomposição. Usamos o espaço de funções (bi)lineares por partes na malha grossa como espaço grosso, veja 6.9 e [7, 31, 24]. Mostramos somente resultados usando uma sobreposição generosa $\delta = nh$. Veja Tabela 6.3. Compare com a Tabela 6.2 na página 105 (precondicionador de um nível) e com a Tabela 4.4 da página 67 (gradiente conjugado sem precondicionador). Pelo Teorema 45 temos que o número de condição depende do número $1+\frac{H}{\delta} \simeq 1+\frac{n}{2}$. Observamos resultados em concordância com a teoria.

Agora consideramos o problema elíptico geral

$$\begin{array}{l} \textit{Achar } u : D \subset \mathbb{R}^2 \to \mathbb{R} \textit{ tal que:} \\ \left\{ \begin{array}{rcl} -\mathrm{div}(\kappa(x)\nabla u(x)) = & f(x) & x \in D \\ u(x) = & g(x) & x \in \partial D \end{array} \right. \end{array}$$

com o coeficiente

$$\kappa(x) = \Big(\kappa_1(x_1,\mu) + 100\kappa_2(x_1,p)\Big)\Big(\kappa_1(x_2,\mu) + 100\kappa_2(x_2,p)\Big) \quad (6.10)$$

$n\backslash N$	2	4	8	16
2	2(1.2500)	8(2.9063)	10(5.4219)	14(7.6644)
4	7(4.5667)	12(4.9929)	15(5.1662)	15(5.1252)
8	9(4.5957)	14(5.3112)	16(5.4764)	15(5.4472)
16	11 (6.0372)	16(7.3533)	18(7.6699)	18(7.65)

Tabela 6.3: Número de iteração do gradiente conjugado precondicionado (estimativa do número de condição) para o problema considerado nesta seção. Aqui $h = 1/(nN)$ e $H = 1/N$ e o fixamos $\delta = 2h$.

onde κ_1 e κ_2 definidos em (2.35) e (2.36), e usamos $\mu = 1000$ e $p = 30$ que corresponde ao caso de um coeficiente com contraste alto e variações nas escalas finas. Usamos o espaço de funções (bi)lineares por partes na malha grossa quadrada gerada pelos subdomínios. Usamos uma sobreposição generosa $\delta = nh = H$. Veja os resultados na Tabela 6.4. Compare com a Tabela 4.6 onde usamos o método do gradiente conjugado sem precondicionador. Temos uma redução considerável no número de iterações. Por exemplo, na Tabela 4.6 temos 289 iterações quando $n = 32$. Na Tabela 6.4 temos 22 iterações quando usamos 8×8 (N=8) subdomínios e uma malha triangular baseada em 4×4 quadrados dentro de cada subdomínio ($n = 4$). Neste caso, em cada uma das 22 iterações temos que resolver 64 sistemas lineares locais de tamanho 25×25 (que é pequeno quando comparado a dimensão 1089×1089 do sistema linear original). Na Tabela 6.4 também temos 18 iterações quando $n = 8$ e $N = 4$, ou 2 iteração quando $n = 16$ e $N = 2$.

6.7 Introdução à análise: como estimar o número de condição?

Como antes temos o lema abstrato de decomposição de domínios.

Lema 46. *Suponhamos,*

1. **Decomposição estável:**. *Existe* $C_0^2 > 0$ *tal que, para toda* $v \in V$, *existe a decomposição* $v = \sum_{i=0}^{N_S} R^{(i)T} v_i$, *com* $v_i \in V^{(i)}$,

$n\backslash N$	2	4	8	16
2	2 (2.00)	16 (9.01)	20 (7.56)	19 (6.74)
4	2 (2.00)	18 (10.42)	22 (8.82)	25 (9.35)
8	2 (2.00)	18 (9.91)	25 (12.70)	36 (18.31)
16	2 (2.00)	18 (11.00)	30 (23.97)	45 (30.83)

Tabela 6.4: Número de iterações do gradiente conjugado precondicionado e em parênteses estimativa do número de condição para o problema considerado nesta seção. Consideramos $h = 1/(nN)$, $H = 1/N$ e $\delta = 2h$.

$i = 0, \ldots, N_S$, *e*

$$\sum_{i=0}^{N_S} a(v_i, v_i) \leq C_0^2 a(v, v). \tag{6.11}$$

2. **Estabilidade local:** *Existe* $\omega > 0$, *tal que*

$$\mathcal{A}(R^{(i)} v_i, R^{(i)T} v_i) \leq \omega \mathcal{A}^{(i)}(v_i, v_i) \quad \forall v_i \in V^{(i)}, \quad 0 \leq i \leq N_S.$$

3. **Desigualdades forte de Cauchy :** *Existem* $\mathcal{E}_{ij}$, $1 \leq i, j \leq N_S$, *tais que para toda* $v_i \in V^{(i)}$, $v_j \in V^{(j)}$

$$\mathcal{A}(R^{(i)T} v_i, R^{(j)} v_j) \leq \mathcal{E}_{ij} \mathcal{A}(R^{(i)T} v_i, R^{(i)T} v_i)^{1/2} \mathcal{A}(R^{(j)T} v_j, R^{(j)T} v_j).$$

Então

$$Cond(M_1^{-1} A) = \kappa(T) \leq (\rho(\mathcal{E}) + 1) \omega C_0^{-2}. \tag{6.12}$$

Para provar o Teorema 45 e estimar o número de condição do precondicionador aditivo de dois níveis temos que verificar as três hipóteses do lema anterior. A prova é a extensão das ideias da prova em uma dimensão (Seção 5.7) para o caso de duas dimensões e não será apresentada aqui. Veja [31, 24, 30].

Capítulo 7

Comentários finais

7.1 Introdução aos métodos sem sobreposição

Como antes queremos resolver o sistema linear global de elementos finitos

$$A\boldsymbol{u} = \boldsymbol{b}$$

que é elíptico e definido positivo. Considere uma decomposição sem sobreposição $\{D_i\}_{i=1}^{N_S}$. Denotemos por $A^{(i)}$ a matriz local (Neumann) da forma bilinear no subdomínio D_i, $i = 1, \ldots, N_S$, isto é $A^{(i)}$ é a representação matricial da forma bilinear $\mathcal{A}$ restrita ao subespaço

$$\mathbb{P}^1(\mathcal{T}^h|_{D_i})$$

isto é, a restrição do espaço global $V = V^h(D) = P_0^1(\mathcal{T})$ aos vértices *interiores e fronteira* do subdomínio D_i, $i = 1, \ldots, N_S$.

Definimos as interfaces locais por $\Gamma_i = \partial D_i \cap D$, $i = 1, \ldots, N_S$ e a interface (global) por $\Gamma = \bigcup_{i=1}^{N} \Gamma_i$. Veja a Figura 7.1.

Dada uma função de elementos finitos $u \in V_0^h(D)$ classificamos os seus graus de liberdade (isto é, os valores da função nos vértices) em

- $\boldsymbol{u}_\Gamma$, os valores na interface que são os valores que corespondem a vértices em Γ (veja a Figura 7.1) e

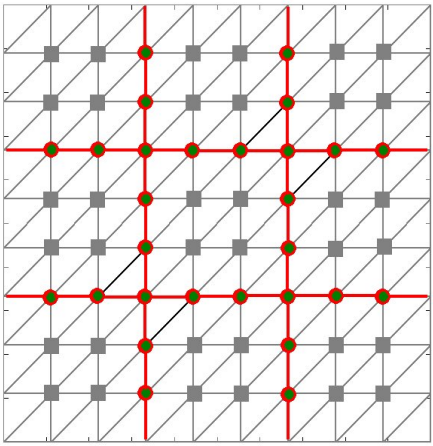


Figura 7.1: Classificação dos graus de liberdade em interface Γ (●) e interiores I (■) .

- $\boldsymbol{u}_I$, os valores interiores, aqueles associados aos vértices fora do Γ e no interior dos subdomínios (veja a Figura 7.1).

Se reordenamos os vértices da triangulação e colocamos primeiro todos os vértices interiores e depois os vértices na interface, obtemos a seguinte estrutura matricial 2×2 em blocos do sistema linear de elementos finitos,

$$A\boldsymbol{u} = \begin{bmatrix} A_{II} & A_{I\Gamma} \\ A_{I\Gamma}^T & A_{\Gamma\Gamma} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_I \\ \boldsymbol{u}_\Gamma \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_I \\ \boldsymbol{b}_\Gamma \end{bmatrix} = \boldsymbol{b}. \tag{7.1}$$

ou

$$\left.\begin{array}{rcl} A_{II}\boldsymbol{u}_I + A_{I\Gamma}\boldsymbol{u}_\Gamma & = & \boldsymbol{b}_I \\ A_{I\Gamma}^T\boldsymbol{u}_I + A_{\Gamma\Gamma}\boldsymbol{u}_\Gamma & = & \boldsymbol{b}_\Gamma \end{array}\right\} \tag{7.2}$$

Podemos fazer o mesmo em cada subdomínio e classificar os graus de liberdade em interiores e na interface, obtemos, o mesmo formato por blocos para as matrizes locais,

$$A^{(i)} = \begin{bmatrix} A_{II}^{(i)} & A_{I\Gamma}^{(i)} \\ A_{I\Gamma}^{(i)T} & A_{\Gamma\Gamma}^{(i)} \end{bmatrix} \quad i = 1, \ldots, N_S. \tag{7.3}$$

7.1.1 O complemento de Schur

Da primeira equação do sistema linear 2×2 em (7.1) ou (7.2) obtemos

$$\boldsymbol{u}_I = A_{II}^{-1}\boldsymbol{b}_I - A_{II}^{-1}A_{I\Gamma}\boldsymbol{u}_\Gamma$$

e substituindo na segunda equação ficamos com o sistema linear

$$S\boldsymbol{u}_\Gamma = \tilde{\boldsymbol{b}} \tag{7.4}$$

onde

$$S = A_{\Gamma\Gamma} - A_{I\Gamma}^T A_{II}^{-1} A_{I\Gamma} \quad \text{e} \quad \tilde{\boldsymbol{b}} = \boldsymbol{b}_\Gamma - A_{I\Gamma}^T A_{II}^{-1}\boldsymbol{b}_I. \tag{7.5}$$

A matriz S é chamada complemento de Schur da matriz A com respeito de Γ. Se reordenamos os vértices interiores de acordo com os subdomínios observamos que A_{II} é diagonal por blocos com respeito aos subdomínios, isto é,

$$A_{II} = \operatorname{diag}(A_{II}^{(i)})_{i=1}^{N_S} = \begin{pmatrix} A_{II}^{(1)} & & & \\ & A_{II}^{(2)} & & \\ & & \ddots & \\ & & & A_{II}^{(N_S)} \end{pmatrix} \tag{7.6}$$

e portanto concluímos que $A_{II}^{-1} = \operatorname{diag}((A_{II}^{(i)})^{-1})_{i=1}^{N_S}$. Note também que

$$A_{I\Gamma} = \begin{pmatrix} A_{I\Gamma}^{(1)} \\ A_{I\Gamma}^{(2)} \\ \vdots \\ A_{I\Gamma}^{(N_S)} \end{pmatrix}. \tag{7.7}$$

Dado um vetor $\boldsymbol{u}_\Gamma$, podemos calcular $S\boldsymbol{u}_\Gamma$ usando (7.5), (7.6) e (7.7) como segue. Usando o formato por blocos (7.3) em cada subdomínio definimos o complemento de Schur local por

$$S^{(i)} = A_{\Gamma\Gamma}^{(i)} - A_{I\Gamma}^{(i)T}(A_{II}^{(i)})^{-1}A_{I\Gamma}^{(i)} \quad i = 1, \ldots, N_S. \tag{7.8}$$

Pode-se verificar facilmente usando (7.5), (7.6) e (7.7) que

$$S = \sum_{i=1}^{N_S} R^{(i)T} S^{(i)} R^{(i)}$$

onde, para $i = 1, \ldots, N_S$, $R^{(i)}$ é a matriz de restrição ao domínio D_i. Concluímos que para calcular $S\boldsymbol{u}_\Gamma$ temos que aplicar os complementos de Schur locais definidos em (7.8), cada um deles envolve resolver um sistema linear da forma

$$A_{II}x_i = R^{(i)}\boldsymbol{u}_\Gamma.$$

Se queremos resolver (7.4) usando o gradiente conjugado, em cada iteração, aplicar S envolve resolver N_S problemas Dirichlet, un em cada subdomínio. A dimensão do sistema linear (7.4) é muito menor que a dimensão do sistema linear original pois envolve somente os graus de liberdade em Γ. Pode-se provar que o número de condição do complemento de Schur é da ordem $\mathrm{Cond}(S) = O(\frac{\kappa_{\max}}{\kappa_{\min}}\frac{1}{hH})$ onde h é o parâmetro da triangulação $\mathcal{T}^h$ e H é o máximo dos diâmetros dos subdomínios $\{D_i\}$, veja [31, 24]. Em principio o sistema linear para o complemento de Schur resulta em menos iterações que para o sistema linear original, mas ainda é um número de iterações muito grande. O uso de um precondicionador é ainda necessário.

7.1.2 Precondicionadores

Pode-se construir vários precondicionadores de decomposição de domínios. Vamos a descrever somente um deles na sua forma matricial. Pode-se como antes, especificar os espaços locais, globais, formas bilineares locais e globais da construção e usar o Lema 38 para estimar o numero de condição do operador precondicionado.

Precondicionador: forma matricial

Um precondicionador aditivo de um nível é

$$B^{-1} = \sum_{i=1}^{N_S} R^{(i)T}(S^{(i)})^\dagger R^{(i)}$$

onde $(S^{(i)})^\dagger$ é a inversa generalizada de $S^{(i)}$. Para aplicar B^{-1} precisa-se calcular $(S^{(i)})^\dagger$ aplicado num vetor do tamanho apropriado. Da definição do complemento de Schur local vemos que para

cada $i = 1, \ldots, N_S$,

$$\begin{bmatrix} A_{II}^{(i)} & A_{I\Gamma}^{(i)} \\ A_{I\Gamma}^{(i)T} & A_{\Gamma\Gamma}^{(i)} \end{bmatrix} \begin{bmatrix} 0 \\ \boldsymbol{u}_\Gamma^{(i)} \end{bmatrix} = \begin{bmatrix} 0 \\ S^{(i)}\boldsymbol{u}_\Gamma^{(i)} \end{bmatrix}, \tag{7.9}$$

e portanto, calcular a ação da inversa do complemento de Schur local $S^{(i)}$, equivale a resolver um problema no subdomínio D_i com a matriz de Neumann $A^{(i)}$ associada ao subdomínio. Lembramos que resolver um problema de Neumann requer uma condição de compatibilidade; veja [7, 31, 22].

O seguinte resultado pode ser provado usando o lema abstrato de decomposição de domínios Lema 38.

Lema 47. *Suponha que consideramos a forma bilinear (3.16) com κ satisfazendo (3.14). Baixo hipóteses adequadas na triangulação $\mathcal{T}^h$ e na decomposição $\{D_i\}_{i=1}^{N_S}$ temos que existe uma constante C tal que*

$$Cond(B^{-1}S) \leq C\frac{1}{H^2}\left(1 + \log(H/h)^2\right).$$

Como antes este limite pode ser melhorado se introduzimos um problema grosso. Por exemplo, podemos usar o problema grosso gerado pelas funções de elementos finitos multi-escala da Seção 6.5.

O precondicionador B^{-1} é o precondicionador de decomposição de domínios mais básico que pode ser construído quando trabalhamos com métodos sem sobreposição. Este precondicionador pode ser melhorado de muitas formas. Para mas detalhes veja [31, 24] e as referências ali citadas.

7.2 Outras equações diferenciais parciais

Nesta seção curta mencionamos algumas outras equações para as quais podemos aplicar o método dos elementos finitos e as técnicas de decomposição de domínios. Certamente esta lista fica muito curta. O objetivo aqui é somente passar a idéia ao leitor da ampla gama de problemas onde as ideias básicas aqui introduzidas podem ser aplicadas. Entre mais complicado o problema considerados, mais detalhes técnicos deveram ser tratados. Em cada equação o espaço de

elementos finitos dever ser escolhido adequadamente. O mesmo acontece com os métodos de decomposição de domínios. Nem todas as equações mencionadas caem na solução de sistemas lineares definidos positivos. Neste caso, no lugar do gradiente conjugado, um outro método iterativo pode ser empregado, entre outros, podemos usar o GMRES ou MINRES; veja [4, 28, 18].

Equação elíptica geral

Seja $D \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$. A equação é

$$\begin{cases} -\mathrm{div}(\kappa(x)\nabla u(x)) + \boldsymbol{B}(x)\nabla u + c(x)u = & f(x) \quad x \in D \\ u(x) = & g(x) \quad x \in \partial D. \end{cases}$$

Esta equação generaliza a equação da pressão em meios porosos e aparece em muitas outras aplicações. Também podemos incluir outras condição de contorno. Veja [31, 24, 27]

Equação de Stokes

Esta é a versão linear e estacionaria das equações de Navier-Stokes. Esta equação é da forma

$$\begin{cases} -\nabla \cdot T(\boldsymbol{u}, p) &= \boldsymbol{b} \quad \text{in } D \\ \nabla \cdot \boldsymbol{u} &= g \quad \text{in } D \\ \boldsymbol{u} &= \boldsymbol{h} \quad \text{on } \partial D \\ \int_D g &= \int_{\partial D} \boldsymbol{h} \cdot \boldsymbol{\eta}. \end{cases} \tag{7.10}$$

Aqui $\boldsymbol{u} = (u_1(x), u_2(x)) : D \to \mathbb{R}^2$, $p : D \to \mathbb{R}$, e definimos $T(\boldsymbol{u}, p) := -pI + 2\mu \boldsymbol{D}\boldsymbol{u}$ onde $\boldsymbol{D}\boldsymbol{u} := \frac{1}{2}(\nabla \boldsymbol{u} + \nabla^T \boldsymbol{u})$ é o tensor de estresse linearizado. Temos que μ é a viscocidade, $\boldsymbol{u}$ é a velocidade e p é a pressão do fluido compressível considerado. A última equação é uma condição de compatibilidade. Veja [10, 16, 11, 15]. Pode-se também incluir as equações de elasticidade quase incompressivel; veja [31].

Lei de Darcy

Como foi mencionado anteriormente, este sistema de equações modela o fluxo de fluidos em meios porosos. O sistema de equações é da

forma,

$$\left\{\begin{array}{rcll} \boldsymbol{u} & = -\frac{\kappa}{\mu}\nabla p + \boldsymbol{b} & \text{in } D & \text{(Darcy's law)} \\ \nabla\cdot\boldsymbol{u} & = g & \text{in } D & \\ \boldsymbol{u}\cdot\boldsymbol{\eta} & = h & \text{on } \partial D & \\ \int_D g & = & \int_{\partial D} h. & \end{array}\right. \tag{7.11}$$

Aqui $\boldsymbol{u}$ é a velocidade do escoamento, p é a pressão, $\boldsymbol{b}$ é uma força externa, g representa a existência de fontes e/ou sumidouros e κ representa a permeabilidade do meio. A última equação é uma condição de compatibilidade. Veja [8, 7, 31, 24, 15, 1, 33].

Equação de advecção difusão

Mencionamos também a equação de advecção difusão. Esta equação é da forma

$$\left\{\begin{array}{rcll} \nu\Delta u + \boldsymbol{B}\cdot u & = f & \text{in } D \\ u & = h & \text{on } \partial D \end{array}\right. \tag{7.12}$$

onde $\nu > 0$ é o coeficiente de difusão e $\boldsymbol{B}$ é uma função vetorial com entradas limitas. Veja [11] e as referências ali citadas.

Equações do tipo Oseen

Mencionamos também sistemas de equações da forma

$$\left\{\begin{array}{rcll} -2\nu\mathrm{div}(\epsilon\Delta\boldsymbol{u}) + \boldsymbol{B}\cdot u + \nabla p & = f & \text{in } D \\ \nabla\cdot\boldsymbol{u} & = g & \text{in } D \\ \boldsymbol{u} & = \boldsymbol{h} & \text{on } \partial D \\ \int_D g & = \int_{\partial D} \boldsymbol{h}\cdot\boldsymbol{\eta} & \end{array}\right. \tag{7.13}$$

Este tipo de equações são obtido das equações de Navier-Stokes depois de uma linearização. Veja [11].

7.3 Bibliografia recomendada

Finalizamos o minicurso recomendando algumas referências.

Método dos elementos finitos

Para uma discussão introdutória recomendamos [21] e as referências ali citadas. Para dicas de implementação veja [3]. Para uma introdução ao método dos elementos finitos com conteúdo matemático no nível de pós-graduação, recomendamos [21, 7, 8, 17, 16, 10] e as referências ali citadas. Para a parte de teoria de equações diferenciais parciais e espaços de funções necessários como pré-requisitos recomendamos [22, 2, 12, 19, 23, 26] entre outros.

Decomposição de domínios

Para uma discussão introdutória recomendamos [30], o primeiro capítulo de [31] e o primeiro capítulo de [24]. Para estudo dos métodos de decomposição de domínios no nível de pós-graduação veja [31, 24, 27] e as referências ali citadas. Atualmente existem muitas aplicações, ainda em desenvolvimento, das tecnicas de decomposição de domínios aos diferentes modelos encontrados na pratica. Convidamos ao leitor a visitar o sítio oficial de decomposição de domínios `www.ddm.org` onde além de um lista de referências atualizadas também pode-se achar as pré-publicações das reuniões e eventos da área. Por últimos citamos algumas teses de doutorado e dissertações de mestrado do IMPA em elementos finitos e/ou decomposição de domínios: [32, 11, 33, 20, 15, 14, 25, 6].

Bibliografia

[1] Eduardo C. Abreu. *Modelagem e Simulação Computational de Escoamentos Trifásicos em Reservatórios de Petróleo Heterogêneos*. PhD thesis, Instituto Politécnico da Universidade de Estado do Rio de Janeiro, 2007.

[2] Robert A. Adams and John J. F. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003.

[3] Jochen Alberty, Carsten Carstensen, and Stefan A. Funken. Remarks around 50 lines of Matlab: short finite element implementation. *Numer. Algorithms*, 20(2-3):117–137, 1999.

[4] Kendall Atkinson and Weimin Han. *Theoretical numerical analysis*, volume 39 of *Texts in Applied Mathematics*. Springer, New York, second edition, 2005. A functional analysis framework.

[5] P Bedrikovetsky. *Mathematical Theory of Oil and Gas Recovery*. Kluwer Academic, London, 1993.

[6] Carlos Borges. Coarse grid correction operator splitting for parabolic partial differential equations. Master's thesis, Instituto Nacional de Matemática Pura e Aplicada, 2007.

[7] Dietrich Braess. *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*. Cambridge University Press, Cambridge, 2001. Second Edition.

[8] Susanne C. Brenner and L. Ridgway Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 1994.

[9] Haïm Brezis. *Analyse fonctionnelle*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1983. Théorie et applications. [Theory and applications].

[10] Franco Brezzi and Michel Fortin. *Mixed and hybrid finite element methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, 1991.

[11] Duilio Conceição. *Balancing Domain Decomposition Preconditioners for Non-symmetric Problems*. PhD thesis, Instituto de Matemática Pura e Aplicada, IMPA, 2006.

[12] Robert Dautray and Jacques-Louis Lions. *Analyse mathématique et calcul numérique pour les sciences et les techniques. Vol. 3-4*. INSTN: Collection Enseignement. [INSTN: Teaching Collection].

[13] Richard E. Ewing. *The mathematics of reservoir simulation*, volume 1 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1983.

[14] Juan Galvis. Finite elements for well-reservoir coupling. Master's thesis, Instituto Nacional de Matemática Pura e Aplicada, April 2004.

[15] Juan Galvis. *Domain Decomposition Aanlysis for Hetrogeneous Darcy's Flow*. PhD thesis, Instituto de Matemática Pura e Aplicada, IMPA, 2008.

[16] Vivette Girault and Pierre-Arnaud Raviart. *Finite element methods for Navier-Stokes equations*, volume 5 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1986. Theory and algorithms.

[17] Vivette Girault, Béatrice Rivière, and Mary F. Wheeler. A discontinuous Galerkin method with nonoverlapping domain decomposition for the Stokes and Navier-Stokes problems. *Math. Comp.*, 74(249):53–84, 2005.

[18] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[19] P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1985.

[20] Etereldes Gonçalves Júnior. *Preconditioners for Elliptic Control Problems*. PhD thesis, Instituto de Matemática Pura e Aplicada, IMPA, 2009.

[21] Claes Johnson. *Numerical solution of partial differential equations by the finite element method*. Cambridge University Press, Cambridge, 1987.

[22] C. Evans Lawrence. *Partial differential equations*. Graduate studies in mathematics. American Mathematical Society, 1990.

[23] J. T. Marti. *Introduction to Sobolev spaces and finite element solution of elliptic boundary value problems*. Computational Mathematics and Applications. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], London, 1986.

[24] Tarek Mathew. *Domain Decomposition Methods for the Numerical Solution of Partial Differential Equations*. Lecture Notes in Computational Science and Engineering. Springer, 2008.

[25] Martha Miranda. The weighted extended B-Splines finite element method. Master's thesis, Instituto Nacional de Matemática Pura e Aplicada, 2005.

[26] Jindřich Nečas. *Les méthodes directes en théorie des équations elliptiques*. Masson et Cie, Éditeurs, Paris, 1967.

[27] Alfio Quarteroni and Alberto Valli. *Domain decomposition methods for partial differential equations*. Numerical Mathematics and Scientific Computation. The Clarendon Press Oxford University Press, New York, 1999. , Oxford Science Publications.

[28] Yousef Saad. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, second edition, 2003.

[29] Jonathan Richard Shewchuk. An introduction to the conjugate gradient method without the agonizing pain. Technical report, August 1994. http://www.cs.cmu.edu/~jrs/jrspapers.html.

[30] Barry F. Smith, Petter E. Bjørstad, and William D. Gropp. *Domain decomposition*. Cambridge University Press, Cambridge, 1996. Parallel multilevel methods for elliptic partial differential equations.

[31] Andrea Toselli and Olof Widlund. *Domain decomposition methods—algorithms and theory*, volume 34 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2005.

[32] Henrique Versieux. *Numerical boundary corrector methods and analysis for a second order elliptic PDE with highly oscillatory periodic coefficients with applications to porous media*. PhD thesis, Instituto de Matemática Pura e Aplicada, IMPA, 2006.

[33] Julia S. Wrobel. *Perda de Injetividade em Reservatórios Estratificados*. PhD thesis, Instituto de Matemática Pura e Aplicada, IMPA, 2005.