

Matemática Numérica Avançada

Pedro H A Konzen

20 de novembro de 2024

Konzen, Pedro Henrique de Almeida

Matemática numérica avançada / Pedro Henrique de Almeida Konzen.
-2024. Porto Alegre.- 2024.

Versão: 20 de novembro de 2024.

"Esta obra é uma edição independente feita pelo próprio autor."

1. Métodos numéricos. 2. Análise numérica. 3. Linguagem Python.

Licença
CC-BY-SA 4.0.

Licença

Este texto é disponibilizado sob a Licença Atribuição-CompartilhaIgual 4.0 Internacional Creative Commons. Para visualizar uma cópia desta licença, visite

http://creativecommons.org/licenses/by-sa/4.0/deed.pt_BR

ou mande uma carta para Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Prefácio

O site [notaspedrok.com.br](https://www.notaspedrok.com.br) é uma plataforma que construí para o compartilhamento de minhas notas de aula. Essas anotações feitas como preparação de aulas é uma prática comum de professoras/es. Muitas vezes feitas a rabiscos em rascunhos com validade tão curta quanto o momento em que são concebidas, outras vezes, com capricho de um diário guardado a sete chaves. Notas de aula também são feitas por estudantes - são anotações, fotos, prints, entre outras formas de registros de partes dessas mesmas aulas. Essa dispersão de material didático sempre me intrigou e foi o que me motivou a iniciar o site.

Com início em 2018, o site contava com apenas três notas incipientes. De lá para cá, conforme fui expandido e revisando os materiais, o site foi ganhando acessos de vários locais do mundo, em especial, de países de língua portuguesa. No momento, conta com 13 notas de aula, além de minicursos e uma coleção de vídeos e áudios.

As notas de **Matemática Numérica Avançada** abordam tópicos sobre sistemas lineares de médio/grande porte, sistemas não-lineares, problemas de otimização, problemas de autovalores e integração auto-adaptativa. Códigos exemplos são apresentados em linguagem [Python](#).

Aproveito para agradecer a todas/os que de forma assídua ou esporádica contribuem com correções, sugestões e críticas! ;)

Pedro H A Konzen

<https://www.notaspedrok.com.br>

Conteúdo

Licença	iii
Prefácio	iv
Conteúdo	vi
1 Sistemas Lineares	1
1.1 Matrizes Esparsas	1
1.1.1 Sistemas Tridiagonais	3
1.1.2 Matrizes Banda	6
1.1.3 Esquemas de Armazenamento	11
1.2 Métodos Iterativos	18
1.2.1 GMRES	18
1.2.2 Método do Gradiente Conjugado	24
1.2.3 Precondicionamento	30
2 Sistemas Não Lineares e Otimização	34
2.1 Sistemas Não-Lineares	34
2.1.1 Método de Newton	35
2.1.2 Método Tipo Newton	38
2.2 Problemas de Minimização	39
2.2.1 Métodos de Declive	40
2.2.2 Método do Gradiente	42
2.2.3 Método de Newton	46
2.2.4 Método do Gradiente Conjugado	51
3 Autovalores e Autovetores	56

3.1	Método da Potência	56
3.1.1	Autovalor dominante	56
3.1.2	Método da Potência Inverso	59
3.2	Iteração QR	60
4	Integração	64
4.1	Integração Autoadaptativa	64
4.2	Integrais múltiplas	67
4.2.1	Regras de Newton-Cotes	67
4.2.2	Regras Compostas de Newton-Cotes	70
	Notas	75
	Referências	78
	Índice de Comandos	80

Capítulo 1

Sistemas Lineares

Em revisão

Neste capítulo, estudamos métodos numéricos para a resolução de sistemas lineares de médio e grande porte. Salvo explicitado ao contrário, assume-se que os sistemas são quadrados e têm solução única.

1.1 Matrizes Esparsas

Em revisão

Uma matriz é dita ser **esparsa** quando ela tem apenas poucos elementos **não nulos**. A ideia é que os elementos não nulos não precisam ser guardados na memória do computador, gerando um grande benefício na redução da demanda de armazenamento de dados. O desafio está no desenvolvimento de estruturas de dados para a alocação eficiente de tais matrizes, i.e. que sejam suficientemente adequadas para os métodos numéricos conhecidos.



Figura 1.1: Em cima: exemplo de uma matriz esparsa estruturada. Em baixo: exemplo de uma matriz esparsa não-estruturada.

Matrizes esparsas podem ser classificadas como **estruturadas** ou **não-estruturadas**. Uma matriz estruturada é aquela em que as entradas não-nulas formam um padrão regular. Por exemplo, estão dispostas em poucas diagonais ou formam blocos (submatrizes densas) ao longo de sua diagonal principal. No caso de não haver um padrão regular das entradas não-nulas, a matriz esparsa é dita ser não-estruturada. Consulte a Figura 1.1 para exemplos.

A **esparsidade** de uma matriz é a porcentagem de elementos nulos que ela

tem, i.e. para uma matriz quadrada $n \times n$ tem-se que a esparsidade é

$$e := \frac{n_{\text{nulos}}}{n^2} \times 100\%. \quad (1.1)$$

Por exemplo, a matriz identidade de tamanho $n = 100$ tem esparsidade

$$e = \frac{100^2 - 100}{100^2} \times 100\% = 99\%. \quad (1.2)$$

1.1.1 Sistemas Tridiagonais

Em revisão

Um **sistema tridiagonal** tem a seguinte forma matricial

$$\begin{bmatrix} a_{1,1} & a_{1,2} & & & 0 \\ a_{2,1} & a_{2,2} & a_{2,3} & & \\ & a_{3,2} & a_{3,3} & a_{3,4} & \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots & a_{n-1,n} \\ 0 & & & & a_{n,n-1} & a_{n,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix}. \quad (1.3)$$

Ou seja, é um sistema cuja a matriz dos coeficientes é tridiagonal.

Uma **matriz tridiagonal** é uma matriz esparsa estruturada. Mais especificamente, é um caso particular de uma **matriz banda**, em que os elementos não nulos estão dispostos apenas em algumas de suas diagonais. Para armazenarmos tal matriz precisamos alocar apenas os seguintes três vetores

$$d_1 = (0, a_{1,2}, \dots, a_{n-1,n}), \quad (1.4)$$

$$d_0 = (a_{1,1}, a_{2,2}, \dots, a_{n,n}), \quad (1.5)$$

$$d_{-1} = (a_{2,1}, \dots, a_{n,n-1}, 0). \quad (1.6)$$

Ou seja, precisamos armazenar $3n$ pontos flutuantes em vez de n^2 , como seria o caso se a matriz dos coeficientes fosse densa. Com isso, podemos alocar a matriz do sistema da seguinte forma

$$\tilde{A} = \begin{bmatrix} * & a_{1,2} & \cdots & a_{n-1,n} \\ a_{1,1} & a_{2,2} & \cdots & a_{n,n} \\ a_{2,1} & \cdots & a_{n,n-1} & * \end{bmatrix}. \quad (1.7)$$

Ou seja, $\tilde{A} = [\tilde{a}_{i,j}]_{i,j=1}^{3,n}$, sendo

$$\tilde{a}_{1+i-j,j} = a_{i,j}. \quad (1.8)$$

Exemplo 1.1.1. Seja o seguinte sistema linear

$$2x_1 - x_2 = 0 \quad (1.9)$$

$$x_{i-1} - 6x_i + 4x_{i+1} = \sin\left(i\frac{\pi}{2(n-1)}\right) \quad (1.10)$$

$$x_{n-1} + x_n = 1 \quad (1.11)$$

O seguinte código [Python](#) faz a alocação de seu vetor dos termos constantes b e de sua matriz de coeficientes no formato compacto de \tilde{A} .

Código 1.1: diagSis.py

```
1 import numpy as np
2 n = 100000
3
4 # alocação
5 # vetor dos termos constantes
6 b = np.empty(n)
7 b[0] = 0.
8 for i in range(1,n-1):
9     b[i] = np.sin(i*np.pi/(2*(n-1)))
10 b[n-1] = 1.
11 print(b)
12 print(f"b size: {b.size*b.itemsize/1024} Kbytes")
13
14 # matriz compacta
15 tA = np.zeros((3,n))
16
17 # indexação
18 def ind(i,j):
19     return 1+i-j,j
20
21 tA[ind(0,0)] = 2.
22 tA[ind(0,1)] = -1.
```

```

23 for i in range(1,n-1):
24     tA[ind(i,i-1)] = 1.
25     tA[ind(i,i)] = -3.
26     tA[ind(i,i+1)] = 4.
27 tA[ind(n-1,n-2)] = 1.
28 tA[ind(n-1,n-1)] = 1.
29 print(tA)
30 print(f"tA size: {tA.size*tA.itemsize/1024**2:1.1f
    } Mbytes")

```

Algoritmo de Thomas (TDMA)

Em revisão

O algoritmo de Thomas⁵ ou TDMA (do inglês, *Tridiagonal Matrix Algorithm*) é uma forma otimizada do método de eliminação gaussiana⁶ aplicada a sistemas tridiagonais. Enquanto este requer $O(n^3)$ operações, esse demanda apenas $O(n)$.

Eliminando os termos abaixo da diagonal em (1.3), obtemos o sistema equivalente

$$\left[\begin{array}{cccc|c} a_{1,1} & a_{1,2} & & 0 & b_1 \\ & \tilde{a}_{2,2} & a_{2,3} & & \tilde{b}_2 \\ & & \tilde{a}_{3,3} & \ddots & \vdots \\ & & & \ddots & a_{n-1,n} \\ 0 & & & & \tilde{a}_{n,n} \end{array} \right] \begin{array}{c} \\ \\ \\ \tilde{b}_{n-1} \\ \tilde{b}_n \end{array} \quad (1.12)$$

Este é obtido pela seguinte iteração

$$w \leftarrow \frac{a_{i+1,i}}{a_{i,i}} \quad (1.13)$$

$$a_{i,i} \leftarrow a_{i,i} - w \cdot a_{i-1,i} \quad (1.14)$$

$$b_i \leftarrow b_i - w \cdot b_{i-1} \quad (1.15)$$

onde, o \sim foi esquecido de propósito, indicando a reutilização da matriz \tilde{A} e do vetor b . A solução do sistema é, então, obtida de baixo para cima, i.e.

$$x_n \leftarrow \frac{b_n}{a_{n,n}} \quad (1.16)$$

$$x_i \leftarrow \frac{b_i - a_{i,i+1}x_{i+1}}{a_{i,i}}, \quad (1.17)$$

com $i = n - 1, n - 2, \dots, 1$.

Código 1.2: tdma.py

```
1 def tdma(ta, b):
2     a = ta.copy()
3     x = b.copy()
4     # eliminação
5     for i in range(1,n):
6         w = a[2,i-1]/a[1,i-1]
7         a[1,i] -= w * a[0,i]
8         x[i] -= w * x[i-1]
9     # resolve
10    x[n-1] = x[n-1]/a[1,n-1]
11    for i in range(n-2,-1,-1):
12        x[i] = (x[i] - a[0,i+1]*x[i+1])/a[1,i]
13    return x
```

1.1.2 Matrizes Banda

Em revisão

Uma **matriz banda** é aquela em que os elementos não nulos estão dispostos em apenas algumas de suas diagonais. Consulte a Figura 1.2.

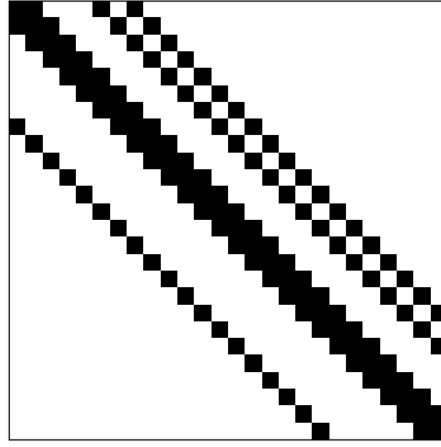


Figura 1.2: Exemplo de uma matriz banda.

Exemplo 1.1.2. Consideramos o seguinte problema de Poisson⁷

$$-\Delta u = f(x, y), \quad (x, y) \in (0, \pi) \times (0, \pi), \quad (1.18)$$

$$u(0, y) = 0, \quad y \in [0, \pi], \quad (1.19)$$

$$u(\pi, y) = 0, \quad y \in [0, \pi], \quad (1.20)$$

$$u(x, 0) = 0, \quad x \in [0, \pi], \quad (1.21)$$

$$u(x, \pi) = 0, \quad x \in [0, \pi], \quad (1.22)$$

onde, $\Delta := \left(\frac{\partial^2}{\partial x^2}, \frac{\partial^2}{\partial y^2} \right)$ é o operador laplaciano⁸. Para fixarmos as ideias, vamos assumir

$$f(x, y) = \sin(x) \sin(y) \quad (1.23)$$

Vamos empregar o **método de diferenças finitas** para computar uma aproximação para a sua solução. Começamos assumindo uma malha uniforme de n^2 nodos

$$x_i = (i - 1)h \quad (1.24)$$

$$y_j = (j - 1)h \quad (1.25)$$

com tamanho de malha $h = \pi/(n - 1)$, $i = 1, 2, \dots, n$ e $j = 1, 2, \dots, n$. Empregando a fórmula de diferenças central, encontramos o seguinte problema

discreto associado

$$u_{i,1} = u_{1,j} = 0 \quad (1.26)$$

$$\begin{aligned} -\frac{1}{h^2}u_{i-1,j} - \frac{1}{h^2}u_{i,j-1} + \frac{4}{h^2}u_{i,j} \\ - \frac{1}{h^2}u_{i+1,j} - \frac{1}{h^2}u_{i,j+1} = f(x_i, y_j) \end{aligned} \quad (1.27)$$

$$u_{i,n} = u_{n,j} = 0 \quad (1.28)$$

Este é um sistema linear $n^2 \times n^2$. Tomando em conta as condições de contorno, ele pode ser reduzido a um sistema $(n-2)^2 \times (n-2)^2$

$$Aw = b \quad (1.29)$$

usando a enumeração das incógnitas

$$(i, j) \rightarrow k = i - 1 + (j - 2)(n - 2), \quad (1.30)$$

i.e.

$$u_{i,j} = w_{k=i-1+(j-2)(n-2)} \quad (1.31)$$

para $i, j = 2, \dots, n - 2$. Consulte a Figura 1.3 para uma representação da enumeração em relação a malha.

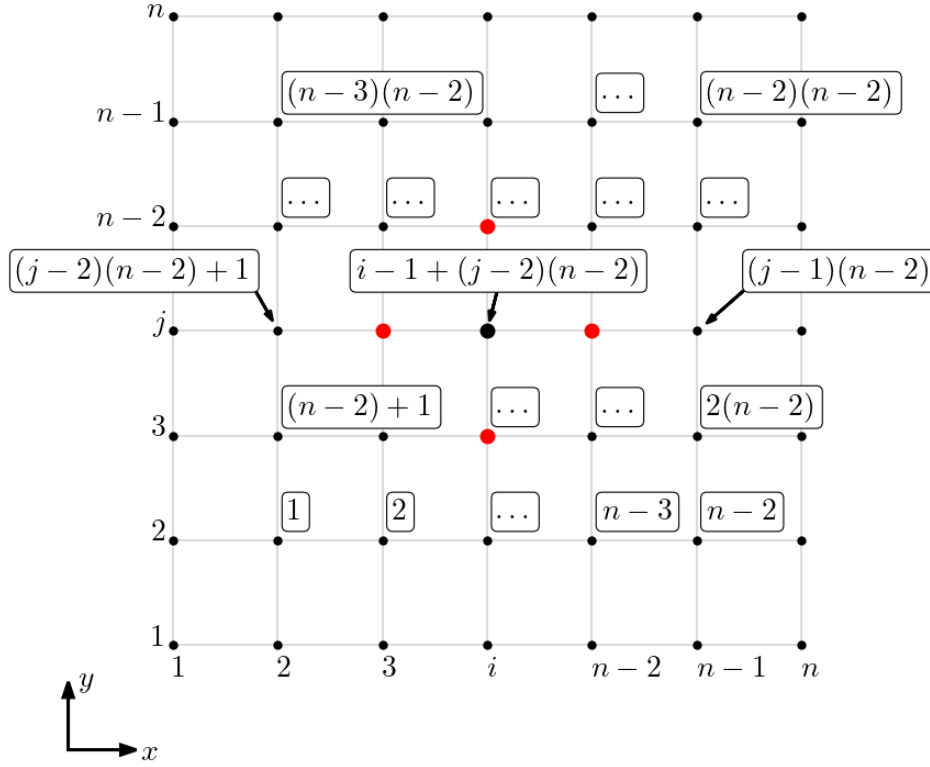


Figura 1.3: Representação da enumeração das incógnitas referente ao problema discutido no Exemplo 1.1.2.

Afim de obtermos uma matriz diagonal dominante, vamos ordenar as equações do sistema discreto como segue

- $j = 2, i = 2$:

$$4w_k - w_{k+1} - w_{k+n-2} = h^2 f_{i,j} \quad (1.32)$$

- $j = 2, i = 3, \dots, n-2$:

$$-w_{k-1} + 4w_k - w_{k+1} - w_{k+n-2} = h^2 f_{i,j} \quad (1.33)$$

- $j = 2, i = n - 1$:

$$-w_{k-1} + 4w_k - w_{k+n-2} = h^2 f_{i,j} \quad (1.34)$$

- $j = 3, \dots, n - 2, i = 2$:

$$-w_{k-(n-2)} + 4w_k - w_{k+1} - w_{k+n-2} = h^2 f_{i,j} \quad (1.35)$$

- $j = 3, \dots, n - 2, i = 3, \dots, n - 2$:

$$-w_{k-1} - w_{k-(n-2)} + 4w_k - w_{k+1} - w_{k+n-2} = h^2 f_{i,j} \quad (1.36)$$

- $j = 3, \dots, n - 2, i = n - 1$:

$$-w_{k-1} - w_{k-(n-2)} + 4w_k - w_{k+n-2} = h^2 f_{i,j} \quad (1.37)$$

- $j = n - 1, i = 2$:

$$-w_{k-(n-2)} + 4w_k - w_{k+1} = h^2 f_{i,j} \quad (1.38)$$

- $j = n - 1, i = 3, \dots, n - 2$:

$$-w_{k-1} - w_{k-(n-2)} + 4w_k - w_{k+1} = h^2 f_{i,j} \quad (1.39)$$

- $j = n - 1, i = n - 1$:

$$-w_{k-1} - w_{k-(n-2)} + 4w_k = h^2 f_{i,j} \quad (1.40)$$

Com isso, temos um sistema com matriz com 5 bandas, consulte a [Figura 1.4](#).

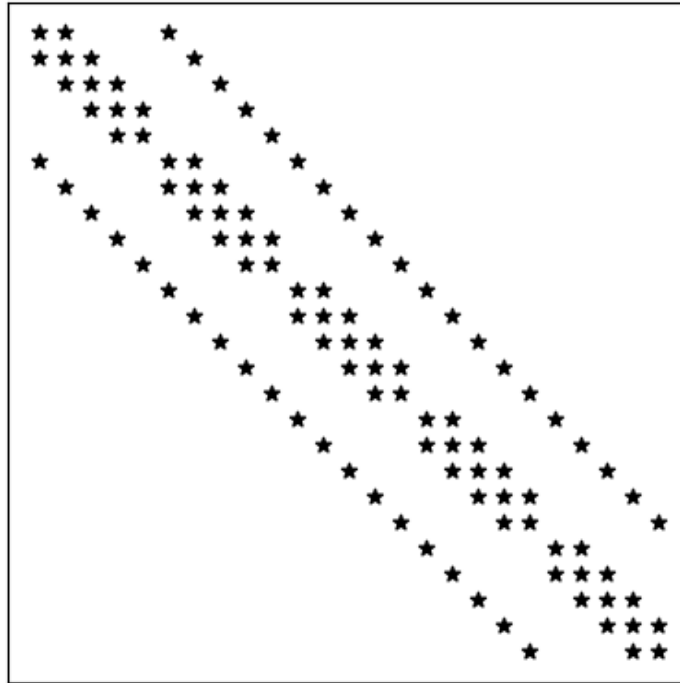


Figura 1.4: Representação da matriz do sistema discreto construído no Exemplo 1.1.2.

1.1.3 Esquemas de Armazenamento

Em revisão

A ideia é armazenar apenas os elementos não-nulos de uma matriz esparsa, de forma a economizar a demanda de armazenamento computacional. Cuidados devem ser tomados para que a estrutura de armazenamento utilizada seja adequada para a computação das operações matriciais mais comuns.

Formato COO

Em revisão

O **formato COO** (do inglês, *COOrdinate format*) é o esquema de armazenamento simples de matrizes esparsas. A estrutura de dados consiste em

três arranjos:

1. um arranjo contendo as entradas não-nulas da matriz;
2. um arranjo contendo seus índices de linha;
3. um arranjo contendo seus índices de coluna.

O método `scipy.sparse.coo_array` permite a alocação de matrizes no formato COO.

Exemplo 1.1.3. O seguinte código armazena a matriz

$$A = \begin{bmatrix} 2. & 0. & 1. & 0. \\ 0. & 3. & 2. & -1. \\ 0 & -1 & -2. & 0. \\ 0 & 0 & 0 & 1. \end{bmatrix} \quad (1.41)$$

no formato COO.

```
1 import numpy as np
2 from scipy.sparse import coo_array
3
4 data = np.array([2., 1., 3., 2., -1., -1., -2., 1.])
5 row = np.array([0, 0, 1, 1, 1, 2, 2, 3])
6 col = np.array([0, 2, 1, 2, 3, 1, 2, 3])
7 Acoo = coo_array((data, (row, col)), shape=(4, 4))
8 print("Acoo = \n", Acoo)
9 print("A = \n", Acoo.toarray())
```

Vantagens do formato COO

- Permite a entrada de dados duplicados (simplicidade).
- Conversão rápida para os formatos CSR e CSC⁹.

Desvantagens do formato COO

- complexidade em operações aritméticas.
- complexidade na extração de submatrizes.

Observação 1.1.1. (**Entradas Duplicadas.**) O formato COO permite a entrada duplicada de elementos da matriz. Na conversão para outros formatos (por exemplo, CSR ou CSC), as entradas duplicadas são somadas.

Formato CSR

Em revisão

O **formato CSR** (do inglês, *Compressed Sparse Row*) é uma variação do COO que busca diminuir a alocação de dados repetidos. Assim como o COO, o formato conta com três arranjos d , c , p :

- d é o arranjo contendo os elementos não-nulos da matriz, ordenados por linhas (i.e., da esquerda para direita, de cima para baixo);
- c é o arranjo contendo o índice das colunas das entradas não-nulas da matriz (como no formato COO);
- p é um arranjo cujos elementos são a posição no arranjo c em que cada linha da matriz começa a ser representada. O número de elementos de i -ésima linha da matriz dado por $p_{j+1} - p_j$.

O método `scipy.sparse.csr_array` permite a alocação de matrizes no formato CSR.

Exemplo 1.1.4. No Exemplo 1.1.3, alocamos a matriz

$$A = \begin{bmatrix} 2. & 0. & 1. & 0. \\ 0. & 3. & 2. & -1. \\ 0 & -1 & -2. & 0. \\ 0 & 0 & 0 & 1. \end{bmatrix} \quad (1.42)$$

no formato COO. Aqui, vamos converter a alocação para o formato CSR e, então, verificar seus atributos.

```
1 from scipy.sparse import csr_array
2 Acsr = Acoo.tocsr()
3 print(f'd = {Acsr.data}')
4 print(f'c = {Acsr.indices}')
5 print(f'p = {Acsr.indptr}')
```

```
d = [ 2.  1.  3.  2. -1. -1. -2.  1.]  
c = [0 2 1 2 3 1 2 3]  
p = [0 2 5 7 8]
```

Por exemplo, o elemento $p[i=2] = 5$ aponta para o $c[k=5] = 1$, o que fornece $A[i=2, j=1] = d[k] = -1$. Verifique!

Vantagens do formato CSR

- operações aritméticas eficientes;
- fatiamento por linhas eficiente;
- multiplicação matriz vetor eficiente.

Desvantagens do formato CSR

- fatiamento por colunas não eficiente;
- custo elevado de realocamento com alteração da esparsidade da matriz.

Formato CSC

Em revisão

O formato CSC (do inglês, *Compressed Sparse Column*) é uma variação análoga do CSR, mas para armazenamento por colunas. O formato conta com três arranjos d , l , p :

- d é o arranjo contendo os elementos não-nulos da matriz, ordenados por colunas (i.e., de cima para baixo, da esquerda para direita);
- l é o arranjo contendo o índice das linhas das entradas não-nulas da matriz;
- p é um arranjo cujos elementos são a posição no arranjo l em que cada coluna da matriz começa a ser representada. O número de elementos de j -ésima coluna da matriz dado por $p_{j+1} - p_j$.

O método `scipy.sparse.csc_array` permite a alocação de matrizes no for-

mato CSC.

Exemplo 1.1.5. No Exemplo 1.1.3, alocamos a matriz

$$A = \begin{bmatrix} 2. & 0. & 1. & 0. \\ 0. & 3. & 2. & -1. \\ 0 & -1 & -2. & 0. \\ 0 & 0 & 0 & 1. \end{bmatrix} \quad (1.43)$$

no formato COO. Aqui, vamos converter a alocação para o formato CSC e, então, verificar seus atributos.

```
1 from scipy.sparse import csc_array
2 Acsc = Acoo.tocsc()
3 l = Acsc.indices
4 p = Acsc.indptr
5 print(f'd = {Acsc.data}')
6 print(f'l = {Acsc.indices}')
7 print(f'p = {Acsc.indptr}')
```

```
d = [2.  3. -1.  1.  2. -2. -1.  1.]
l = [0 1 2 0 1 2 1 3]
p = [0 1 3 6 8]
```

Assim sendo, o elemento $p[j=2] = 3$ aponta para o $l[k=3] = 0$, o que informa que $A[i=0, j=2] = d[k] = 1$. Verifique!

Vantagens do formato CSC

- fatiamento por colunas eficiente;
- operações aritméticas eficientes;
- multiplicação matriz vetor eficiente¹.

Desvantagens do formato CSC

- fatiamento por linhas não eficiente;
- custo elevado de realocamento com alteração da esparsidade da matriz.

¹CSR é mais eficiente em muitos casos.

Observação 1.1.2. Além dos formatos COO, CSR e CSC, existem ainda vários outros que podem empregados e que são mais eficientes em determinadas aplicações. Recomendamos a leitura de [9, Seção 3.4] e da documentação do [scipy.sparse](#).

Exercícios

Em revisão

E.1.1.1. Considere o problema de Poisson dado no Exemplo 1.1.2.

- a) Armazene a matriz do problema discreto associado usando o formato COO.
- b) Converta a matriz armazenada para o formato CSR². Então, compute a solução do problema discreto com o método [spsolve](#)³.
- c) Converta a matriz armazenada para o formato CSC⁴. Então, compute a solução do problema discreto com o método [spsolve](#).
- d) Compare a eficiência da computação entre os itens b) e c) para tamanhos de malha $h = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$.

E.1.1.2. Considere o seguinte sistema linear

$$2x_1 - x_2 = 0 \tag{1.44}$$

$$x_{i-1} - 6x_i + 4x_{i+1} = \sin\left(i \frac{\pi}{2(n-1)}\right) \tag{1.45}$$

$$x_{n-1} + x_n = 1 \tag{1.46}$$

- a) Compute sua solução usando o Algoritmo de Thomas para $n = 3$.
- b) Compare a solução obtida no item anterior com a gerada pela função [scipy.linalg.solve](#).

²Use o método [coo_matrix.tocsr\(\)](#).

³[scipy.sparse.linalg.spsolve](#) é uma implementação do Método LU otimizado para matrizes esparsas.

⁴Use o método [coo_matrix.tocsc\(\)](#).

- c) Compare a solução com a obtida no item anterior com a gerada pela função `scipy.linalg.solve_banded`.
- d) Use o módulo `Python datetime` para comprar a demanda de tempo computacional de cada um dos métodos acima. Compute para $n = 10, 100, 1000, 10000$.

E.1.1.3. Considere que o problema de valor de contorno (PVC)

$$-u'' = \sin \pi x, \quad 0 < x < 1, \quad (1.47)$$

$$u(0) = 0, \quad (1.48)$$

$$u(1) = 0 \quad (1.49)$$

seja simulado com o Método das Diferenças Finitas⁵. Vamos assumir uma discretização espacial uniforme com n nodos e tamanho de malha

$$h = \frac{1}{n-1}. \quad (1.50)$$

Com isso, temos os nodos $x_i = (i-1)h$, $i = 1, 2, \dots, n$. Nos nodos internos, aplicamos a fórmula de diferenças central

$$u''(x_i) \approx \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2}, \quad (1.51)$$

onde, $u_i \approx u(x_i)$. Com isso, a discretização da EDO fornece

$$-\frac{1}{h^2}u_{i-1} + \frac{2}{h^2}u_i - \frac{1}{h^2}u_{i+1} = \sin \pi x_i \quad (1.52)$$

para $i = 2, 3, \dots, n-1$. Das condições de contorno temos $u_1 = u_n = 0$. Logo, o problema discreto lê-se: encontrar $u = (u_1, u_2, \dots, u_n) \in \mathbb{R}^n$ tal que

$$u_1 = 0 \quad (1.53)$$

$$-\frac{1}{h^2}u_{i-1} + \frac{2}{h^2}u_i - \frac{1}{h^2}u_{i+1} = \sin \pi x_i \quad (1.54)$$

$$u_n = 0 \quad (1.55)$$

- a) Calcule a solução analítica do PVC.

⁵Consulte mais em [Notas de Aula: Matemática Numérica](#).

- b) Use a função `scipy.linalg.solve_banded` para computar a solução do problema discreto associado para diferentes tamanhos de malha $h = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$. Compute o erro da solução discreta em relação à solução analítica.
- c) Compare a demanda de tempo computacional se a função `scipy.linalg.solve` for empregada na computação da solução discreta.

E.1.1.4. Consideremos o problema trabalho no Exemplo 1.1.2.

- a) Use a função `scipy.linalg.solve` para computar a solução do problema discreto associado para diferentes tamanhos de malha $h = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$. Compute o erro da solução discreta em relação à solução analítica. Compare as aproximações com a solução analítica

$$u(x, y) = \frac{1}{2} \sin(x) \sin(y). \quad (1.56)$$

- b) Compare a demanda de tempo e memória computacional se a função `scipy.linalg.solve_banded` for empregada na computação da solução discreta.
- c) Baseado no Algoritmo de Thomas, implemente o Método de Eliminação Gaussiana otimizado para a matriz banda deste problema. Compare com as abordagens dos itens a) e b).

1.2 Métodos Iterativos

Em revisão

1.2.1 GMRES

Em revisão

O **GMRES** (do inglês, *Generalized Minimal Residual Method*¹⁰) é um **método de subespaço de Krylov**¹¹ e é considerado uma das mais eficientes técnicas para a resolução de sistemas lineares gerais e de grande porte (esparcos).

Método de Subespaço de Krylov

Em revisão

A ideia básica é resolver o sistema linear

$$Ax = b \quad (1.57)$$

por um **método de projeção**. Mais especificamente, busca-se uma solução aproximada $x_m \in \mathbb{R}^n$ no subespaço afim $x_0 + \mathcal{K}_m$ de dimensão m , impondo-se a **condição de Petrov**¹²-**Galerkin**¹³

$$b - Ax_m \perp \mathcal{L}_m, \quad (1.58)$$

onde \mathcal{L}_m também é um subespaço de dimensão m . Quando \mathcal{K}_m é um **subespaço de Krylov**, i.e.

$$\mathcal{K}_m(A, r_0) = \text{span}\{r_0, Ar_0, A^2r_0, \dots, A^{m-1}r_0\}, \quad (1.59)$$

temos o **método de subespaço de Krylov**. Aqui, temos o **resíduo**

$$r_0 = b - Ax_0, \quad (1.60)$$

sendo x_0 uma aproximação inicial para a solução do sistema. Notamos que com isso, temos que a aproximação calculada é tal que

$$A^{-1}b \approx x_m = x_0 + q_{m-1}(A)r_0, \quad (1.61)$$

onde q_{m-1} é um dado polinômio de grau $m-1$. No caso particular de $x_0 = 0$, temos

$$A^{-1}b \approx q_{m-1}(A)b. \quad (1.62)$$

Diferentes versões deste método são obtidas pelas escolhas do subespaço \mathcal{L}_m e formas de condicionamento do sistema.

GMRES

Em revisão

O **GMRES** é um método de subespaço de Krylov assumindo $\mathcal{L}_m = A\mathcal{K}_m$, com

$$\mathcal{K}_m = \mathcal{K}_m(A, v_1) = \text{span}\{v_1, Av_1, \dots, A^{m-1}v_1\}, \quad (1.63)$$

onde $v_1 = r_0/\|r_0\|$ é o vetor unitário do resíduo $r_0 = b - Ax_0$ para uma dada aproximação inicial x_0 da solução do sistema $Ax = b$.

Vamos derivar o método observando que qualquer vetor x em $x_0 + \mathcal{K}_m$ pode ser escrito como segue

$$x = x_0 + V_m y \quad (1.64)$$

onde, $V_m = [v_1, \dots, v_m]$ é a matriz $n \times m$ cujas colunas formam uma base ortogonal $\{v_1, \dots, v_m\}$ de \mathcal{K}_m e $y \in R^m$. Aqui, V_m é computada usando-se o seguinte **método de Arnoldi**¹⁴ - **Gram**¹⁵ - **Schmidt**¹⁶ modificado [9, Subseção 6.3]:

1. Dado v_1 de norma 1

2. Para $j = 1, \dots, m$:

(a) $w_j \leftarrow Av_j$

(b) Para $i = 1, \dots, j$:

i. $h_{i,j} \leftarrow (w_j, v_i)$

ii. $w_j \leftarrow w_j - h_{i,j}v_i$

(c) $h_{j+1,j} \leftarrow \|w_j\|$

(d) Se $h_{j+1,j} \leftarrow 0$, então pare.

(e) $v_{j+1} \leftarrow w_j/h_{j+1,j}$

Seja, então, $\bar{H}_m = [h_{i,j}]_{i,j=1}^{m+1,m}$ a **matriz de Hessenberg**¹⁷ cujas entradas não nulas são computadas pelo algoritmo acima (Passos 2(a)i-ii). Pode-se mostrar que [9, Proposição 6.5]

$$J(y) = \|b - Ax\| \quad (1.65)$$

$$= \|b - A(x_0 + V_m y)\| \quad (1.66)$$

$$= \|\beta e_1 - \bar{H}_m y\| \quad (1.67)$$

onde, $\beta = \|r_0\|$.

A aproximação GMRES x_m é então computada como

$$x_m = x_0 + V_m y_m, \quad (1.68)$$

$$y_m = \min_y \|\beta e_1 - \bar{H}_m y\| \quad (1.69)$$

Observamos que este último é um pequeno problema de minimização, sendo que requer a solução de um sistema $(m+1) \times m$ de mínimos quadrados, sendo m normalmente pequeno.

Em resumo, a solução GMRES x_m é computada seguindo os seguintes passos:

1. Escolhemos uma aproximação inicial x_0 para a solução de $Ax = b$.
2. Computamos o resíduo $r_0 = b - Ax_0$.
3. Computamos o vetor unitário $v_1 = r_0/\|r_0\|$.
4. Usamos o método de Arnoldi-Gram-Schmidt modificado para calculamos uma base ortogonal V_m de \mathcal{K}_m e a matriz de Hessenberg \bar{H}_m associada.
5. Computamos $y_m = \min_y \|\beta e_1 - \hat{H}_m y\|$.
6. Computamos $x_m = x_0 + V_m y$.

Observação 1.2.1 (Convergência). Pode-se mostrar que o GMRES converge em ao menos n passos.

Observação 1.2.2 (GMRES com a ortogonalização de Householder). No algoritmo acima, o método modificado de Gram-Schmidt é utilizado no processo de Arnoldi. Uma versão numericamente mais eficiente é obtida quando a **transformação de Householder**¹⁸ é utilizada. Consulte mais em [9, Subsection 6.5.2].

Observação 1.2.3 (GMRES com Reinicialização). O **restarted GMRES** é uma variação do método para sistemas que requerem uma aproximação GMRES x_m com m grande. Nestes casos, o método original pode

demandar um custo muito alto de memória computacional. A ideia consiste em assumir m pequeno e, caso não suficiente, recalculer a aproximação GMRES com $x_0 = x_m$. Este algoritmo pode ser descrito como segue.

1. Computamos $r_0 = b - Ax_0$, $\beta = \|r_0\|$ e $v_1 = r_0/\beta$
2. Computamos V_m e \hat{H}_m pelo método de Arnoldi
3. Computamos

$$y_m = \min_y \|\beta e_1 - \hat{H}_m y\| \quad (1.70)$$

$$x_m = x_0 + V_m y_m \quad (1.71)$$

4. Se $\|b - Ax_m\|$ é satisfatória, paramos. Caso contrário, setamos $x_0 := x_m$ e voltamos ao passo 1.

A convergência do restarted GMRES não é garantida para matrizes que não sejam positiva-definidas.

Exercícios

Em revisão

E.1.2.1. Considere o problema discreto do Exercício 1.1.3.

- a) Compute a solução com a implementação restarted GMRES

[scipy.sparse.linalg.gmres.](#)

- b) Por padrão, o intervalo de iterações entre as inicializações é `restart=20`. Compare o desempenho para diferentes intervalos de reinicialização.

- c) Compare o desempenho entre as abordagens dos itens a) e b) frente a implementação do método de eliminação gaussiana disponível em

[scipy.sparse.linalg.spsolve.](#)

E.1.2.2. Considere o problema discreto trabalhado no Exemplo 1.1.2.

a) Compute a solução com a implementação restarted GMRES

[scipy.sparse.linalg.gmres](#).

b) Por padrão, o intervalo de iterações entre as inicializações é `restart=20`. Compare o desempenho para diferentes intervalos de reinicialização.

c) Compare o desempenho entre as abordagens dos itens a) e b) frente a implementação do método de eliminação gaussiana disponível em

[scipy.sparse.linalg.spsolve](#).

E.1.2.3. Considere o seguinte problema de Poisson¹⁹ com condições de contorno não homogêneas.

$$-\Delta u = f(x, y), \quad (x, y) \in D, \quad (1.72)$$

$$u = g, \quad \text{em } \partial D \quad (1.73)$$

Para fixarmos as ideias, vamos assumir o domínio $D = (0, 1) \times (0, 1)$, a fonte

$$f(x, y) = 2\pi^2 \sin \pi(x + y) \quad (1.74)$$

e os valores no contorno

$$g = \sin \pi(x + y), \quad (x, y) \in \partial D. \quad (1.75)$$

Observamos que a solução analítica deste problema é

$$u(x, y) = \sin \pi(x + y). \quad (1.76)$$

Empregue o método de diferenças finitas para computar uma aproximação para a solução. Assumimos uma malha uniforme de n^2 nodos

$$x_i = (i - 1)h \quad (1.77)$$

$$y_j = (j - 1)h \quad (1.78)$$

com tamanho de malha $h = 1/(n - 1)$, $i = 1, 2, \dots, n$ e $j = 1, 2, \dots, n$. Empregando a fórmula de diferenças central encontramos o seguinte problema discreto associado

$$u_{i,1} = g(x_i, 0) \quad (1.79)$$

$$u_{1,j} = g(0, y_j) \quad (1.80)$$

$$\begin{aligned} & -\frac{1}{h^2}u_{i-1,j} - \frac{1}{h^2}u_{i,j-1} + \frac{4}{h^2}u_{i,j} \\ & -\frac{1}{h^2}u_{i+1,j} - \frac{1}{h^2}u_{i,j+1} = f(x_i, y_j) \end{aligned} \quad (1.81)$$

$$u_{i,n} = g(x_i, 1) \quad (1.82)$$

$$u_{n,j} = g(1, y_j) \quad (1.83)$$

Este pode ser escrito na forma matricial

$$Aw = b \quad (1.84)$$

onde, A é $(n-2)^2 \times (n-2)^2$ e assumindo a enumeração

$$u_{i,j} = w_{k=i-1+(j-2)(n-2)}, \quad i, j = 2, \dots, n-2. \quad (1.85)$$

Consulte a Figura 1.4.

1. Compute a solução do problema discreto associado usando a seguinte implementação [Python](#) do GMRES

[scipy.sparse.linalg.gmres](#)

2. Compare o desempenho com a aplicação do método LU implemento em

[scipy.sparse.linalg.spsolve](#)

E.1.2.4. Faça sua própria implementação do método GMRES. Valide-a e compare-a com a resolução do exercício anterior (Exercício 1.2.3).

1.2.2 Método do Gradiente Conjugado

Em revisão

O **método do gradiente conjugado** é uma das mais eficientes técnicas iterativas para a resolução de sistema linear com matriz esparsa, simétrica e

definida positiva. Vamos assumir que o sistema

$$Ax = b \quad (1.86)$$

onde, a A é **simétrica** e **definida positiva**.

O método pode ser derivado a partir do método de Arnoldi²⁰ [9, Seção 6.7] ou como uma variação do método do gradiente. Este é caminho que será adotado aqui.

Método do Gradiente

Em revisão

A ideia é reformular o sistema $Ax = b$ como um problema de minimização. Vamos começar definindo o funcional

$$J(y) = \frac{1}{2}y^T Ay - y^T b. \quad (1.87)$$

O vetor y que minimiza J é a solução de $Ax = b$. De fato, denotando x a solução de $Ax = b$, temos

$$J(y) = \frac{1}{2}y^T Ay - y^T b + \frac{1}{2}x^T Ax - \frac{1}{2}x^T Ax \quad (1.88)$$

$$= \frac{1}{2}(y - x)^T A(y - x) - \frac{1}{2}x^T Ax \quad (1.89)$$

O último termo é independente de y e, portanto, J é mínimo quando

$$\frac{1}{2}(y - x)^T A(y - x) \quad (1.90)$$

é minimizado. Agora, como A é definida positiva²¹, o menor valor deste termo ocorre quando $y - x = 0$, i.e. $y = x$.

Observamos, também, que o gradiente de J é

$$\nabla J = Ay - b \quad (1.91)$$

i.e., é o oposto do resíduo $r = b - Ay$. Com isso, temos que $y = x$ é a única escolha tal que $\nabla J = 0$. Ainda, temos que ∇J é o vetor que aponta na

direção e sentido de maior crescimento de J . Isso nos motiva a aplicarmos a seguinte iteração²²

$$x^{(0)} = \text{aprox. inicial} \quad (1.92)$$

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla J(x^{(k)}) \quad (1.93)$$

onde, $\alpha_k > 0$ é um escalar que regula o tamanho do passo a cada iteração. Lembrando que $-\nabla J = r$, temos que a iteração é equivalente a

$$x^{(k+1)} = x^{(k)} + \alpha_k r^{(k)}. \quad (1.94)$$

Notamos que $x^{(k+1)}$ é um ponto na reta $\{x^{(k)} + \alpha r^{(k)} : \alpha \in \mathbb{R}\}$ que tem a mesma direção de $\nabla J(x^{(k)})$ e passa pelo ponto $x^{(k)}$. O procedimento de escolher um $\alpha^{(k)}$ entre todos os possíveis, é conhecido como pesquisa linear (em inglês, *line search*).

A cada iteração, queremos escolher α_k de forma que $J(x^{(k+1)}) \leq J(x^{(k)})$. Isso pode ser garantido fazendo a seguinte escolha⁶

$$J(x^{(k+1)}) = \min_{\alpha \in \mathbb{R}} J(x^{(k)} + \alpha r^{(k)}) \quad (1.95)$$

A fim de resolver este problema de minimização, vamos denotar

$$g(\alpha) = J(x^{(k)} + \alpha r^{(k)}). \quad (1.96)$$

Então, observamos que

$$\begin{aligned} g(\alpha) &= \frac{1}{2} (x^{(k)} + \alpha r^{(k)})^T A (x^{(k)} + \alpha r^{(k)}) - (x^{(k)} + \alpha r^{(k)})^T b \\ &= \frac{1}{2} x^{(k)T} A x^{(k)} + \frac{\alpha}{2} x^{(k)T} A r^{(k)} + \frac{\alpha}{2} r^{(k)T} A x^{(k)} \\ &\quad + \frac{\alpha^2}{2} r^{(k)T} A r^{(k)} - x^{(k)T} b - \alpha r^{(k)T} b \end{aligned} \quad (1.97)$$

Agora, usando o fato de A ser simétrica, obtemos

$$g(\alpha) = J(x^{(k)}) + \alpha r^{(k)T} A x^{(k)} + \frac{\alpha^2}{2} r^{(k)T} A r^{(k)} - \alpha r^{(k)T} b \quad (1.98)$$

⁶Chamada de pesquisa linear exata. Qualquer outra escolha para α é conhecida como pesquisa linear não exata.

$$= J\left(x^{(k)}\right) - \alpha r^{(k)T} r^{(k)} + \frac{\alpha^2}{2} r^{(k)T} A r^{(k)} \quad (1.99)$$

a qual, é uma função quadrática. Seu único mínimo, ocorre quando

$$0 = g'(\alpha) \quad (1.100)$$

$$= -r^{(k)T} r^{(k)} + \alpha r^{(k)T} b. \quad (1.101)$$

Logo, encontramos

$$\alpha = \frac{r^{(k)T} r^{(k)}}{r^{(k)T} A r^{(k)}} \quad (1.102)$$

Com isso, temos a iteração do Método do Gradiente

$$x^{(0)} = \text{aprox. inicial} \quad (1.103)$$

$$x^{(k+1)} = x^{(k)} + \alpha_k r^{(k)}, \quad (1.104)$$

$$\alpha_k = \frac{r^{(k)T} r^{(k)}}{r^{(k)T} A r^{(k)}} \quad (1.105)$$

Observação 1.2.4 (Detalhe de Implementação). Observamos que, a cada iteração, precisamos computar $A r^{(k)}$ (no cálculo de α_k) e $A x^{(k)}$ (no cálculo do resíduo). Essas multiplicações matriz-vetor são os passos computacionais mais custosos do método. Podemos otimizar isso usando o fato de que

$$r^{(k+1)} = r^{(k)} - \alpha_k A r^{(k)}. \quad (1.106)$$

Exercícios

Em revisão

E.1.2.5. Faça sua implementação do método do gradiente.

E.1.2.6. Use a implementação feita no exercício anterior (Exercício 1.2.6) nos seguintes itens.

- a) Compute a solução do problema discreto do Exemplo 1.1.2 pelo Método do Gradiente. Quantas iterações são necessárias para obter um resíduo com norma $\leq 10^{-14}$?
- b) Compute a solução do problema discreto do Exercício 1.2.3 pelo Método do Gradiente. Quantas iterações são necessárias para obter um resíduo com norma $\leq 10^{-14}$?
- c) Compare a aplicação do método GMRES²³ e do método LU²⁴ nos itens anteriores.

E.1.2.7. Considere o Exercício 1.2.3.

- a) Use sua implementação do método do gradiente para computar uma solução aproximada, cuja norma do resíduo $\leq 10^{-14}$.
- b) Compare o desempenho com a aplicação da implementação GMRES

[scipy.sparse.linalg.gmres](#)

Método do Gradiente Conjugado

Em revisão

O método do gradiente consiste em uma iteração da forma

$$x_0 = \text{aprox. inicial}, \quad (1.107)$$

$$x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}, \quad (1.108)$$

com $p^{(k)} = r^{(k)}$. Ou seja, a nova aproximação $x^{(k+1)}$ é buscada na direção de $p^{(k)}$. Aqui, a ideia é usar uma melhor direção para buscar a solução.

O **método do gradiente conjugado** é um método de gradiente que busca encontrar a solução de $Ax = b$ pela computação do mínimo do seguinte funcional²⁵

$$J(y) = \frac{1}{2} \langle y, y \rangle_A - \langle b, y \rangle, \quad (1.109)$$

onde, $\langle \cdot, \cdot \rangle$ denota o produto interno padrão e

$$\langle x, y \rangle_A := x^T A y \quad (1.110)$$

é o **produto interno induzido por A** , lembrando que A é positiva definida²⁶. Associada a este produto interno, temos a norma

$$\|x\|_A := \sqrt{\langle x, x \rangle_A}, \quad (1.111)$$

chamada de **norma da energia**. O produto interno associado é também conhecido como **produto interno da energia**. Com isso, definimos que dois vetores x e y são **conjugados**, quando eles são ortogonais com respeito ao produto interno da energia, i.e. quando

$$\langle x, y \rangle_A = 0. \quad (1.112)$$

Aqui, a ideia é desenvolver um método iterativo em que o erro a cada passo seja conjugado a todas as direções de busca anteriores. Consulte o desenvolvimento detalhado do método em [11, Seção 7.7].

Código 1.3: Algoritmo do gradiente conjugado

```
1 import numpy as np
2 import scipy as sp
3 from scipy.linalg import norm
4
5 def mgc(A, b, x, tol=1e-14):
6     n, = b.shape
7     r = b - A*x
8     p = r
9     nu = np.dot(r, r)
10    for it in np.arange(n):
11        q = A*p
12        mu = np.dot(p, q)
13        alpha = nu/mu
14        x = x + alpha*p
15        r = r - alpha*q
16        nu0 = np.dot(r, r)
17        beta = nu0/nu
18        p = r + beta*p
19        nu = nu0
20        if (norm(r) < tol):
21            print(it)
```

```
22     return x
23     raise ValueError("Falha de convergencia.")
```

Exercícios

Em revisão

E.1.2.8. Use o Código 1.3 na resolução dos seguintes itens.

1. Compute a solução do problema discreto do Exemplo 1.1.2 pelo método do gradiente conjugado. Quantas iterações são necessárias para obter um resíduo com norma $\leq 10^{-14}$?
2. Compute a solução do problema discreto do Exercício 1.2.3 pelo método do gradiente conjugado. Quantas iterações são necessárias para obter um resíduo com norma $\leq 10^{-14}$?
3. Compare a aplicação do método GMRES²⁷ e da implementação SciPy do método do gradiente conjugado²⁸

E.1.2.9. Considere o Exercício 1.2.3.

- a) Use sua implementação do método do gradiente conjugado Código ?? para computar uma solução aproximada, cuja norma do resíduo $\leq 10^{-14}$.
- b) Compare o desempenho com a aplicação de sua implementação do método do gradiente (Exercício 1.2.6).
- c) Compare o desempenho com a aplicação da implementação GMRES

[scipy.sparse.linalg.gmres](#)

1.2.3 Precondicionamento

Em revisão

Precondicionamento refere-se a modificar o sistema linear original de forma que a computação de sua solução possa ser feita de forma mais eficiente. No

lugar do sistema original

$$Ax = b \quad (1.113)$$

resolvemos o sistema equivalente

$$MAx = Mb, \quad (1.114)$$

onde $M = P^{-1}$ e a matriz P é chamada de **precondicionador** do sistema. De forma geral, a escolha do preconditionador é tal que $P \approx A$, mas com inversa fácil de ser computada. Além disso, uma característica esperada é que MA tenha esparsidade parecida com A .

Precondicionamento ILU

Em revisão

A ideia é tomar P igual a uma fatoração LU incompleta (ILU, do inglês, *Incomplete LU*). Incompleta no sentido que entradas de L e de U sejam adequadamente removidas, buscando-se uma boa esparsidade e ao mesmo tempo uma boa aproximação LU para A .

ILU(0)

Em revisão

O preconditionamento ILU(0) impõe que as matrizes L e U tenham o mesmo padrão de esparsidade da matriz A .

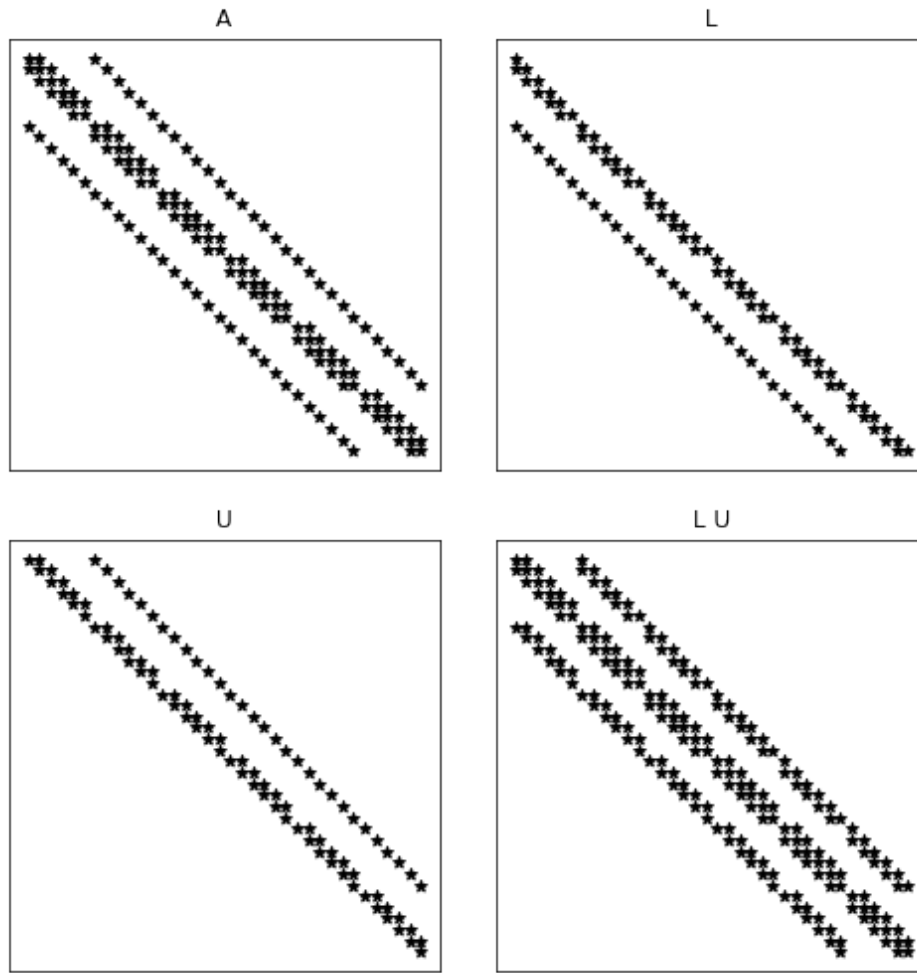


Figura 1.5: Representação das matrizes $ILU(0)$.

Exemplo 1.2.1. `ex:possoinDnhIlu0` Consideramos o sistema linear $Ax = b$ associado ao problema discreto trabalhado no Exercício 1.2.3. Para uma malha $n \times n = 8 \times 8$, obtemos as matrizes representadas na Figura 1.5.

Observamos que a matriz LU contém duas diagonais com elementos não nulos a mais que a matriz original A . Estes elementos são chamados de *fill-in*.

Código 1.4: Algoritmo $ILU(0)$

```
1 import scipy.sparse as sp
```

```
2
3 def ilu0(A):
4     n,n = A.get_shape()
5     LU = A.copy()
6     nz = A.nonzero()
7     for i in range(1,n):
8         for k in [_ for _ in range(i) if (i,_) in zip(
9             nz[0],nz[1])]:
10             LU[i,k] = LU[i,k]/LU[k,k]
11             LU[i,j] = LU[i,j] - LU[i,k]*LU[k,j]
12     return LU
```

Exercícios

Em revisão

E.1.2.10. Considere o problema discreto do Exercício 1.2.3.

- a) Compute a solução com o método GMRES com preconditionamento ILU(0).
- b) Compare com a resolução com o método GMRES sem preconditionamento.
- c) Compare com a resolução com o método CG sem preconditionamento.
- d) O preconditionamento ILU(0) é eficiente para o método CG?

Capítulo 2

Sistemas Não Lineares e Otimização

Em revisão

Neste capítulo, apresentam-se métodos numéricos para a resolução de sistemas não lineares e de problemas de otimização. Salvo explicitado ao contrário, assume-se que os problemas são bem definidos.

2.1 Sistemas Não-Lineares

Em revisão

Consideramos o seguinte problema: dada a função $F : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ encontrar $\mathbf{x}^* \in \mathbb{R}^n$ tal que

$$F(\mathbf{x}^*) = 0. \quad (2.1)$$

Salvo explicitado ao contrário, assumiremos que $F \in C^1(D)$, i.e. F é uma função continuamente diferenciável no domínio computacional $D \subset \mathbb{R}^n$.

Vamos, denotar por $J_F(\mathbf{x}) = [j_{i,j}(\mathbf{x})]_{i,j=1}^{n,n}$ a **matriz jacobiana**²⁹ da F com

$$j_{i,j}(\mathbf{x}) = \frac{\partial f_i(\mathbf{x})}{\partial x_j}, \quad (2.2)$$

onde $F(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_n(\mathbf{x}))$ e $\mathbf{x} = (x_1, x_2, \dots, x_n)$.

2.1.1 Método de Newton

Em revisão

A iteração básica do **método de Newton**³⁰ para sistemas de equações consiste em: dada uma aproximação inicial $\mathbf{x}^{(0)} \in \mathbb{R}^n$,

$$\text{resolver } J_F(\mathbf{x}^{(k)}) \boldsymbol{\delta}^{(k)} = -F(\mathbf{x}^{(k)}) \quad (2.3)$$

$$\text{computar } \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \boldsymbol{\delta}^{(k)} \quad (2.4)$$

para $k = 0, 1, 2, \dots$ até que um critério de parada seja satisfeito.

Observação 2.1.1 (Convergência). Para $\mathbf{x}^{(0)}$ suficientemente próximo da solução \mathbf{x}^* , **o método de Newton é quadraticamente convergente**. Mais precisamente, este resultado de convergência local requer que $J_F^{-1}(\mathbf{x}^*)$ seja não singular e que J_F seja Lipschitz³¹ contínua. Consulte [8, Seção 7.1] para mais detalhes.

Exemplo 2.1.1. Consideramos a **equação de Burgers**³²

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = \nu \frac{\partial^2 u}{\partial x^2} \quad (2.5)$$

com $\nu > 0$, condição inicial

$$u(0, x) = -\sin(\pi x) \quad (2.6)$$

e condições de contorno de Dirichlet³³ homogêneas

$$u(t, -1) = u(t, 1) = 0. \quad (2.7)$$

Aplicando o **método de Rothe**³⁴ com aproximação de Euler³⁵ implícita, obtemos

$$\frac{u(t + h_t, x) - u(t, x)}{h_t} + u(t + h_t, x) u_x(t + h_t, x) \approx \nu u_{xx}(t + h_t, x), \quad (2.8)$$

onde $h_t > 0$ é o passo no tempo. Agora, aplicamos diferenças finitas para obter

$$\begin{aligned} & \frac{u(t + h_t, x_i) - u(t, x_i)}{h_t} \\ & + u(t + h_t, x_i) \frac{u(t + h_t, x_{i+1}) - u(t + h_t, x_{i-1})}{2h_x} \\ & \approx \nu \frac{u(t + h_t, x_{i-1}) - 2u(t + h_t, x_i) + u(t + h_t, x_{i+1}))}{h_x^2}, \end{aligned} \quad (2.9)$$

onde, $x_i = (i - 1)h_x$, $i = 1, \dots, n_x$ e $h_x = 1/(n_x - 1)$ é o tamanho da malha.

Rearranjando os termos e denotando $u_i^{(k)} \approx u(t_k, x_i)$, $t_k = (k - 1)h$, obtemos o seguinte **problema discreto**

$$u_1^{(k+1)} = 0 \quad (2.10)$$

$$\begin{aligned} & \frac{1}{h_t} u_i^{(k+1)} - \frac{1}{h_t} u_i^{(k)} + \frac{1}{2h_x} u_i^{(k+1)} u_{i+1}^{(k+1)} \\ & - \frac{1}{2h_x} u_i^{(k+1)} u_{i-1}^{(k+1)} - \frac{\nu}{h_x^2} u_{i-1}^{(k+1)} \\ & + 2 \frac{\nu}{h_x^2} u_i^{(k+1)} - \frac{\nu}{h_x^2} u_{i+1}^{(k+1)} = 0, \end{aligned} \quad (2.11)$$

$$u_{n_x}^{(k+1)} = 0, \quad (2.12)$$

sendo $u_i^{(0)} = -\sin(\pi x_i)$, $i = 1, \dots, n_x$ e $k = 1, 2, \dots$

Este problema pode ser reescrito como segue: para cada $k = 0, 1, \dots$, encontrar $\mathbf{w} \in \mathbb{R}^{n_x}$, tal que

$$F(\mathbf{w}; \mathbf{w}^{(0)}) = 0, \quad (2.13)$$

onde $w_i = u_i^{(k+1)}$, $w_i^{(0)} = u_i^{(k)}$ e

$$f_1(\mathbf{w}; \mathbf{w}^{(0)}) = w_1, \quad (2.14)$$

$$\begin{aligned} f_i(\mathbf{w}; \mathbf{w}^{(0)}) &= \frac{1}{h_t} w_i - \frac{1}{h_t} w_i^{(0)} + \frac{1}{2h_x} w_i w_{i+1} \\ & - \frac{1}{2h_x} w_i w_{i-1} - \frac{\nu}{h_x^2} w_{i-1} + 2 \frac{\nu}{h_x^2} w_i - \frac{\nu}{h_x^2} w_{i+1}, \end{aligned} \quad (2.15)$$

$$f_{n_x}(\mathbf{w}; \mathbf{w}^{(0)}) = w_{n_x}. \quad (2.16)$$

A matriz jacobiana associada $J = [J_{i,j}]_{i,j}^{n_x, n_x}$ contém

$$J_{i,j} = 0, \quad j \neq i-1, i, i+1, \quad (2.17)$$

$$J_{1,1} = 1, \quad (2.18)$$

$$J_{1,2} = 0, \quad (2.19)$$

$$J_{i,i-1} = \frac{1}{2h_x}w_i - \frac{\nu}{h_x^2}, \quad (2.20)$$

$$J_{i,i} = \frac{1}{h_t} + \frac{1}{2h_x}w_{i+1} - \frac{2}{h_x}w_{i-1} + \frac{2\nu}{h_x^2}, \quad (2.21)$$

$$J_{i,i+1} = \frac{1}{2h_x}w_i - \frac{\nu}{h_x^2}, \quad (2.22)$$

$$J_{n_x, n_x-1} = 0 \quad (2.23)$$

$$J_{n_x, n_x} = 1. \quad (2.24)$$

Implemente este esquema numérico!

Exercícios

Em revisão

E.2.1.1. Considere o problema discreto apresentado no Exemplo 2.1.1 para diferentes valores do coeficiente de difusão $\nu = \nu_0/\pi$, $\nu_0 = 1., 0.1, 0.01, 0.001$. Simule o problema com cada uma das seguintes estratégias e as compare quanto ao desempenho computacional.

- Simule-o aplicando o Método de Newton com o *solver* linear `npla.solve`.
- Observe que a jacobiana é uma matriz tridiagonal. Simule o problema aplicando o Método de Newton com o *solver* linear `npla.solve_banded`.
- Aloque a jacobiana como uma matriz esparsa. Então, simule o problema aplicando o Método de Newton com *solver* linear adequado para matrizes esparsas.

E.2.1.2. Desenvolva um esquema *upwind* para simular a equação de Burgers do Exemplo 2.1.1.

2.1.2 Método Tipo Newton

Em revisão

Existem várias modificações do Método de Newton³⁶ que buscam reduzir o custo computacional. Há estratégias para simplificar as computações da matriz jacobiana³⁷ e para reduzir o custo nas computações das atualizações de Newton.

Atualização Cíclica da Matriz Jacobiana

Em revisão

Geralmente, ao simplificarmos a matriz jacobina J_F ou aproximarmos a atualização de Newton $\delta^{(k)}$, perdemos a convergência quadrática do método (consulte a Observação 2.1.1). A ideia é, então, buscar uma convergência pelo menos super-linear, i.e.

$$\|e^{(k+1)}\| \approx \rho_k \|e^{(k)}\|, \quad (2.25)$$

com $\rho_k \rightarrow 0$ quando $k \rightarrow \infty$. Aqui, $e^{(k)} := x^* - x^{(k)}$. Se a convergência é superlinear, então podemos usar a seguinte aproximação

$$\rho_k \approx \frac{\|\delta^{(k)}\|}{\|\delta^{(k-1)}\|} \quad (2.26)$$

ou, equivalentemente,

$$\rho_k \approx \left(\frac{\|\delta^{(k)}\|}{\|\delta^{(0)}\|} \right)^{\frac{1}{k}} \quad (2.27)$$

Isso mostra que podemos acompanhar a convergência das iterações pelo fator

$$\beta_k = \frac{\|\delta^{(k)}\|}{\|\delta^{(0)}\|}. \quad (2.28)$$

Ao garantirmos $0 < \beta_k < 1$, deveremos ter uma convergência superlinear.

Vamos, então, propor o seguinte método tipo Newton de atualização cíclica da matriz jacobiana.

1. Escolha $0 < \beta < 1$
2. $J := J_F(x^{(0)})$
3. $J\delta^{(0)} = -F(x^{(0)})$
4. $x^{(1)} = x^{(0)} + \delta^{(0)}$
5. Para $k = 1, \dots$ até critério de convergência:
 - (a) $J\delta^{(k)} = -F(x^{(k)})$
 - (b) $x^{(k+1)} = x^{(k)} + \delta^{(k)}$
 - (c) Se $\|\delta^{(k)}\|/\|\delta^{(0)}\| > \beta$:
 - i. $J := J_F(x^{(k)})$

E.2.1.3. Implemente uma versão do método tipo Newton apresentado acima e aplique-o para simular o problema discutido no Exemplo 2.1.1 para $\nu = 1., 0.1, 0.01, 0.001, 0.0001$. Faça uma implementação com suporte para matrizes esparsas.

2.2 Problemas de Minimização

Em revisão

Vamos considerar o seguinte **problema de minimização**: dada a **função objetivo** $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$, resolver

$$\min_{x \in D} f(x). \quad (2.29)$$

No que segue e salvo dito explicitamente ao contrário, vamos assumir que o problema está bem determinado e que f é suficientemente suave. Ainda, vamos assumir as seguintes notações:

- gradiente de f

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \dots, \frac{\partial f}{\partial x_n}(\mathbf{x}) \right) \quad (2.30)$$

- derivada direcional de f com respeito a $\mathbf{d} \in \mathbb{R}^n$

$$\frac{\partial f}{\partial \mathbf{d}}(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \mathbf{d} \quad (2.31)$$

- matriz hessiana de f , $H = [h_{i,j}]_{i,j=1}^{n,n}$

$$h_{i,j}(\mathbf{x}) = \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) \quad (2.32)$$

Observação 2.2.1 (**Condições de Otimização**). Se $\nabla f(\mathbf{x}^*) = 0$ e $H(\mathbf{x}^*)$ é positiva definida, então \mathbf{x}^* é um mínimo local de f em uma vizinhança não vazia de \mathbf{x}^* . Consulte mais em [8, Seção 7.2]. Um ponto \mathbf{x}^* tal que $\nabla f(\mathbf{x}^*) = 0$ é chamado de **ponto crítico**.

2.2.1 Métodos de Declive

Em revisão

Um **método de declive** consiste em uma iteração tal que: dada uma aproximação inicial $\mathbf{x}^{(0)} \in \mathbb{R}^n$, computa-se

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^{(k)} \mathbf{d}^{(k)}, \quad (2.33)$$

com **tamanho de passo** $\alpha^{(k)} > 0$, para $k = 0, 1, 2, \dots$ até que um dado critério de parada seja satisfeito. As **direções descendentes** são tais que

$$\mathbf{d}^{(k)} \cdot \nabla f(\mathbf{x}^{(k)}) < 0, \quad \text{se } \nabla f(\mathbf{x}^{(k)}) \neq 0, \quad (2.34)$$

$$\mathbf{d}^{(k)} = 0, \quad \text{noutro caso.} \quad (2.35)$$

Observação 2.2.2 (**Condição de Convergência**). Da Série de Taylor³⁸, temos que

$$f(\mathbf{x}^{(k)} + \alpha^{(k)} \mathbf{d}^{(k)}) - f(\mathbf{x}^{(k)}) = \alpha^{(k)} \nabla f(\mathbf{x}^{(k)}) \cdot \mathbf{d}^{(k)} + \varepsilon, \quad (2.36)$$

com $\varepsilon \rightarrow 0$ quando $\alpha^{(k)} \rightarrow 0$. Como consequência da continuidade da f , para $\alpha^{(k)}$ suficientemente pequeno, o sinal do lado esquerdo é igual a do direito desta última equação. Logo, para tais $\alpha^{(k)}$ e $\mathbf{d}^{(k)} \neq 0$ uma direção descendente, temos garantido que

$$f(\mathbf{x}^{(k)} + \alpha^{(k)} \mathbf{d}^{(k)}) < f(\mathbf{x}^{(k)}). \quad (2.37)$$

Notamos que um método de declive fica determinado pelas escolhas da direção de declive $\mathbf{d}^{(k)}$ e o tamanho do passo $\alpha^{(k)}$. Primeiramente, vamos a este último item.

Pesquisa linear

Em revisão

O **método de pesquisa linear** consiste em escolher $\alpha^{(k)}$ com base na resolução do seguinte problema de minimização

$$\min_{\alpha \in \mathbb{R}} \phi(\alpha) := f(\mathbf{x}^{(k)} + \alpha \mathbf{d}^{(k)}). \quad (2.38)$$

Entretanto, a resolução exata deste problema é muitas vezes não factível. Técnicas de aproximações para a resolução deste problema são, então, aplicadas. Tais técnicas são chamadas de **pesquisa linear não exata**.

A **condição de Armijo** é que a escolha de $\alpha^{(k)}$ deve ser tal que

$$f(\mathbf{x}^{(k)} + \alpha^{(k)} \mathbf{d}^{(k)}) \leq f(\mathbf{x}^{(k)}) + \sigma \alpha^{(k)} \nabla f(\mathbf{x}^{(k)}) \cdot \mathbf{d}^{(k)}, \quad (2.39)$$

para alguma constante $\sigma \in (0, 1/2)$. Ou seja, a redução em f é esperada ser proporcional à derivada direcional de f com relação a direção $\mathbf{d}^{(k)}$ no ponto $\mathbf{x}^{(k)}$. Em aplicações computacionais, σ é normalmente escolhido no intervalo $[10^{-5}, 10^{-1}]$.

A condição (2.39) não é suficiente para evitar escolhas muito pequenas de $\alpha^{(k)}$. Para tanto, pode-se empregar a **condição de curvatura**, a qual requer que

$$\nabla f(\mathbf{x}^{(k)} + \alpha^{(k)} \mathbf{d}^{(k)}) \cdot \mathbf{d}^{(k)} \geq \beta \nabla f(\mathbf{x}^{(k)}) \cdot \mathbf{d}^{(k)}, \quad (2.40)$$

para $\beta \in [\sigma, 1/2]$. Notemos que o lado esquerdo de (2.40) é igual a $\phi'(\alpha^{(k)})$. Ou seja, esta condição impõe que $\alpha^{(k)}$ seja maior que $\beta\phi'(0)$. Normalmente, escolhe-se $\beta \in [10^{-1}, 1/2]$. Juntas, (2.39) e (2.40) são conhecidas como **condições de Wolfe**³⁹.

2.2.2 Método do Gradiente

Em revisão

O **método do gradiente** (ou método do máximo declive) é um método de declive tal que as direções descendentes são o oposto do gradiente da f , i.e.

$$\mathbf{d}^{(k)} = -\nabla f(\mathbf{x}^{(k)}). \quad (2.41)$$

É imediato verificar que as condições (2.34)-(2.35) são satisfeitas.

Exemplo 2.2.1. Consideramos o problema de encontrar o mínimo da **função de Rosenbrock**⁴⁰

$$f(\mathbf{x}) = \sum_{i=1}^n 100 \left(x_{i+1} - x_i^2 \right)^2 + (1 - x_i)^2. \quad (2.42)$$

O valor mínimo desta função é zero e ocorre no ponto $\mathbf{x} = \mathbf{1}$. Esta função é comumente usada como caso padrão para teste de métodos de otimização.

Para o método do gradiente, precisamos das derivadas parciais

$$\frac{\partial f}{\partial x_1} = -400x_1 \left(x_2 - x_1^2 \right) - 2(1 - x_1) \quad (2.43)$$

$$\begin{aligned} \frac{\partial f}{\partial x_j} &= \sum_{i=1}^n 200 \left(x_i - x_{i-1}^2 \right) (\delta_{i,j} - 2x_{i-1}\delta_{i-1,j}) \\ &\quad - 2(1 - x_{i-1})\delta_{i-1,j} \end{aligned} \quad (2.44)$$

$$\frac{\partial f}{\partial x_n} = 200 \left(x_n - x_{n-1}^2 \right) \quad (2.45)$$

onde, $\delta_{i,j}$ é o delta de Kronecker⁴¹.

Código 2.1: Algoritmo do Gradiente

```
1 import numpy as np
```



```
2 import numpy.linalg as npla
3 import scipy.optimize as spopt
4
5 # fun obj
6 def fun(x):
7     '''
8     Função de Rosenbrock
9     '''
10    return sum(100.*(x[1:]-x[:-1]**2.))**2. + (1.-x
11    [:-1])**2.)
12
13 # gradiente da fun
14 def grad(x):
15     xm = x[1:-1]
16     xm_m1 = x[:-2]
17     xm_p1 = x[2:]
18     der = np.zeros_like(x)
19     der[1:-1] = 200*(xm-xm_m1**2) - 400*(xm_p1 - xm
20     **2)*xm - 2*(1-xm)
21     der[0] = -400*x[0]*(x[1]-x[0]**2) - 2*(1-x[0])
22     der[-1] = 200*(x[-1]-x[-2]**2)
23
24     return der
25
26 # problem dimension
27 n = 2
28
29 # num max iters
30 maxiter = 100000
31
32 # tolerancia
33 tol = 1e-10
34
35 # aprox. inicial
36 x = np.zeros(n)
37
38 for k in range(maxiter):
39     # direcao descendente
40     d = -grad(x)
```

```
38
39 # pesquisa linear
40 alpha = spopt.line_search(fun, grad, x, d,
41                           c1=0.0001, c2=0.9)[0]
42
43 # atualizacao
44 x = x + alpha * d
45
46 nad = npla.norm(alpha * d)
47 nfun = npla.norm(fun(x))
48
49 if ((k+1) % 10 == 0):
50     print(f"{k+1}: {alpha:1.2e} {nad:1.2e} {nfun
51 :1.2e}")
52
53 if (nfun < tol):
54     break
```

Exercícios

Em revisão

E.2.2.1. Aplique o método do gradiente para computar o ponto mínimo da função de Rosenbrock⁴²

$$f(\mathbf{x}) = \sum_{i=1}^n 100 \left(x_{i+1} - x_i^2 \right)^2 + (1 - x_i)^2 \quad (2.46)$$

com

- a) $n = 2$.
- b) $n = 3$.
- c) $n = 4$.
- d) $n = 5$.
- e) $n = 10$.

E.2.2.2. Aplique o método do gradiente para computar o ponto mínimo da função de Beale [1]

$$\begin{aligned} f(x, y) = & (1.5 - x + xy)^2 \\ & + (2.25 - x + xy^2)^2 \\ & + (2.625 - x + xy^3)^2. \end{aligned} \quad (2.47)$$

para $\mathbf{x} \in [-4.5, 4.5]^2$.

E.2.2.3. Aplique o método do gradiente para computar o ponto mínimo da função de Goldstein-Price [4]

$$\begin{aligned} f(x, y) = & \left[1 + (x + y + 1)^2 (19 - 14x \right. \\ & \left. + 3x^2 - 14y + 6xy + 3y^2) \right] \\ & \times \left[30 + (2x - 3y)^2 (18 - 32x \right. \\ & \left. + 12x^2 + 48y - 36xy + 27y^2) \right] \end{aligned} \quad (2.48)$$

E.2.2.4. Aplique o método do gradiente para computar o ponto mínimo da função de Booth

$$f(x, y) = (x + 2y - 7)^2 + (2x + y - 5)^2 \quad (2.49)$$

para $\mathbf{x} \in [-10, 10]^2$.

E.2.2.5. Aplique o método do gradiente para computar o ponto mínimo da função de Rastrigin

$$f(\mathbf{x}) = 10n + \sum_{i=1}^n \left[x_i^2 - 10 \cos(2\pi x_i) \right], \quad (2.50)$$

para $\mathbf{x} \in [-5.12, 5.12]^n$, com

a) $n = 2$.

b) $n = 3$.

- c) $n = 4$.
- d) $n = 5$.
- e) $n = 10$.

2.2.3 Método de Newton

Em revisão

O método de Newton⁴³ para problemas de otimização é um método de declive com direções descendentes

$$\mathbf{d}^{(k)} = -H^{-1}(\mathbf{x}^{(k)}) \nabla f(\mathbf{x}^{(k)}), \quad (2.51)$$

assumindo que a hessiana H seja definida positiva dentro de uma vizinhança suficientemente grande do ponto de mínimo \mathbf{x}^* . Esta escolha é baseada no polinômio de Taylor da função objetivo f

$$f(\mathbf{x}^{(k+1)}) \approx f(\mathbf{x}^{(k)}) + \nabla f(\mathbf{x}^{(k)}) \cdot \mathbf{d}^{(k)} + \frac{1}{2} \mathbf{d}^{(k)} \cdot H(\mathbf{x}^{(k)}) \mathbf{d}^{(k)}. \quad (2.52)$$

Com isso, escolhemos $\mathbf{x}^{(k+1)}$ de forma a minimizar o lado direito desta aproximação, i.e.

$$\frac{\partial}{\partial d_i^{(k)}} \left(f(\mathbf{x}^{(k)}) + \nabla f(\mathbf{x}^{(k)}) \cdot \mathbf{d}^{(k)} + \frac{1}{2} \mathbf{d}^{(k)} \cdot H(\mathbf{x}^{(k)}) \mathbf{d}^{(k)} \right) = 0 \quad (2.53)$$

para $i = 1, 2, \dots, n$. Ou seja, temos

$$\nabla f(\mathbf{x}^{(k)}) \cdot \mathbf{d}^{(k)} + H(\mathbf{x}^{(k)}) \mathbf{d}^{(k)} = \mathbf{0} \quad (2.54)$$

o que leva a (2.51).

Observação 2.2.3 (Computação da Direção). Na implementação computacional, não é necessário computar a inversa da hessiana, a direção $\mathbf{d}^{(k)}$ pode ser mais eficientemente computada resolvendo-se

$$H(\mathbf{x}^{(k)}) \mathbf{d}^{(k)} = -\nabla f(\mathbf{x}^{(k)}). \quad (2.55)$$

Observação 2.2.4 (*Solver linear*). O método usado para computar a solução de (2.55) é chamado de *solver linear*. Por exemplo, **Newton-Krylov**⁴⁴ é o nome dado ao método de Newton que utiliza um método de subespaço de Krylov como *solver linear*. Mais especificamente, **Newton-GMRES** quando o GMRES é escolhido como *solver linear*. Uma escolha natural é **Newton-GC**, tendo em vista que o método de gradiente conjugado é ideal para matriz simétrica e definida positiva.

Exemplo 2.2.2. Seguindo o Exemplo 2.2.1, temos que a hessiana associada é a matriz simétrica $H = [h_{i,j}]_{i,j=1}^{n,n}$ com

$$h_{1,1} = \frac{\partial^2 f}{\partial x_1^2} = 1200x_1^2 - 400x_2 + 2 \quad (2.56)$$

$$h_{1,2} = \frac{\partial^2 f}{\partial x_1 \partial x_2} = -400x_1 \quad (2.57)$$

$$\begin{aligned} h_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j} = & 200(\delta_{i,j} - 2x_{i-1}\delta_{i-1,j}) - 400(\delta_{i+1,j} - 2x_i\delta_{i,j}) \\ & - 400\delta_{i,j}(x_{i+1} - x_i^2) + 2\delta_{i,j} \end{aligned} \quad (2.58)$$

$$h_{n-1,n} = \frac{\partial^2 f}{\partial x_{n-1} \partial x_n} = -400x_{n-1} \quad (2.59)$$

$$h_{n,n} = \frac{\partial^2 f}{\partial x_n^2} = 200 \quad (2.60)$$

Notemos que a hessiana é uma matriz tridiagonal.

```
1 import numpy as np
2 import numpy.linalg as npla
3 import scipy.optimize as spopt
4
5 # fun obj
6 def fun(x):
7     '''
8     Funcao de Rosenbrock
9     '''
10    return sum(100.*(x[1:] - x[:-1])**2.)**2. + (1.-x
    [:-1])**2.)
```

```
11
12 # gradiente da fun
13 def grad(x):
14     xm = x[1:-1]
15     xm_m1 = x[:-2]
16     xm_p1 = x[2:]
17     der = np.zeros_like(x)
18     der[1:-1] = 200*(xm-xm_m1**2) - 400*(xm_p1 - xm
19     **2)*xm - 2*(1-xm)
20     der[0] = -400*x[0]*(x[1]-x[0]**2) - 2*(1-x[0])
21     der[-1] = 200*(x[-1]-x[-2]**2)
22
23     return der
24
25 def hess(x):
26     x = np.asarray(x)
27     H = np.diag(-400*x[:-1],1) - np.diag(400*x
28     [:-1],-1)
29     diagonal = np.zeros_like(x)
30     diagonal[0] = 1200*x[0]**2-400*x[1]+2
31     diagonal[-1] = 200
32     diagonal[1:-1] = 202 + 1200*x[1:-1]**2 - 400*x
33     [2:]
34     H = H + np.diag(diagonal)
35
36     return H
37
38 # dimensao
39 n = 2
40
41 # num max iters
42 maxiter = 100000
43 # tolerancia
44 tol = 1e-10
45
46 # aprox. inicial
47 x = np.zeros(n)
48
```

```
46 for k in range(maxiter):
47
48     # direcao descendente
49     d = npla.solve (hess(x), -grad(x))
50
51     # pesquisa linear
52     alpha = spopt.line_search(fun, grad, x, d)[0]
53
54     # atualizacao
55     x = x + alpha * d
56
57     nad = npla.norm(alpha * d)
58     nfun = npla.norm(fun(x))
59
60     print(f"{k+1}: {alpha:1.2e} {nad:1.2e} {nfun:1.2e}")
61
62     if ((nfun < tol) or np.isnan(nfun)):
63         break
```

Observação 2.2.5 (**Métodos Tipo Newton**). Métodos tipo Newton são aqueles que utilizam uma aproximação para a inversa da matriz hessiana. Uma estratégia comumente aplicada, é a de atualizar a hessiana apenas em algumas iterações, baseando-se em uma estimativa da taxa de convergência. Na Subseção 2.1.2, exploramos esta técnica no contexto de resolução de sistemas não lineares.

Exercícios

Em revisão

E.2.2.6. Aplique o método de Newton para computar o ponto mínimo da função de Rosenbrock⁴⁵

$$f(\mathbf{x}) = \sum_{i=1}^n 100 \left(x_{i+1} - x_i^2 \right)^2 + (1 - x_i)^2 \quad (2.61)$$

com

- a) $n = 2$.
- b) $n = 3$.
- c) $n = 4$.
- d) $n = 5$.
- e) $n = 10$.

E.2.2.7. Aplique o método de Newton para computar o ponto mínimo da função de Beale [1]

$$\begin{aligned} f(x, y) = & (1.5 - x + xy)^2 \\ & + (2.25 - x + xy^2)^2 \\ & + (2.625 - x + xy^3)^2. \end{aligned} \quad (2.62)$$

para $\mathbf{x} \in [-4.5, 4.5]^2$.

E.2.2.8. Aplique o método de Newton para computar o ponto mínimo da função de Goldstein-Price [4]

$$\begin{aligned} f(x, y) = & \left[1 + (x + y + 1)^2 (19 - 14x \right. \\ & \left. + 3x^2 - 14y + 6xy + 3y^2) \right] \\ & \times \left[30 + (2x - 3y)^2 (18 - 32x \right. \\ & \left. + 12x^2 + 48y - 36xy + 27y^2) \right] \end{aligned} \quad (2.63)$$

E.2.2.9. Aplique o método de Newton para computar o ponto mínimo da função de Booth

$$f(x, y) = (x + 2y - 7)^2 + (2x + y - 5)^2 \quad (2.64)$$

para $\mathbf{x} \in [-10, 10]^2$.

E.2.2.10. Aplique o método de Newton para computar o ponto mínimo da função de Rastrigin

$$f(x) = 10n + \sum_{i=1}^n [x_i^2 - 10 \cos(2\pi x_i)], \quad (2.65)$$

para $\mathbf{x} \in [-5.12, 5.12]^n$, com

- a) $n = 2$.
- b) $n = 3$.
- c) $n = 4$.
- d) $n = 5$.
- e) $n = 10$.

2.2.4 Método do Gradiente Conjugado

Em revisão

Métodos do gradiente conjugado são obtidos escolhendo-se as direções descendentes

$$d^{(k)} = -\nabla f(x^{(k)}) + \beta_k d^{(k-1)}, \quad (2.66)$$

onde β_k é um escalar escolhido de forma que as direções $\{d^{(k)}\}$ sejam mutuamente ortogonais com respeito a uma dada norma. Por exemplo, o **método de Fletcher-Reeves** consiste em escolher

$$\beta_k = \frac{\nabla f(x^{(k)}) \cdot \nabla f(x^{(k)})}{\nabla f(x^{(k-1)}) \cdot \nabla f(x^{(k-1)})}, \quad (2.67)$$

o que garante que as direções sejam mutuamente ortogonais com respeito ao produto interno euclidiano.

Observação 2.2.6 (**Variantes**). Outras variações comumente empregadas são o Método de Polak-Ribière e suas variantes. Consulte mais em [7, Seção 5.2].

Exemplo 2.2.3. Implementação do Método de Fletcher-Reeves para a minimização da função de Rosenbrock dada no Exemplo 2.2.1.

```
1 import numpy as np
2 import numpy.linalg as npla
3 import scipy.optimize as spopt
4
5 # fun obj
6 def fun(x):
7     '''
8     Funcao de Rosenbrock
9     '''
10    return sum(100.*(x[1:]-x[:-1]**2.))**2. + (1.-x
11    [:-1])**2.)
12 # gradiente da fun
13 def grad(x):
14     xm = x[1:-1]
15     xm_m1 = x[:-2]
16     xm_p1 = x[2:]
17     der = np.zeros_like(x)
18     der[1:-1] = 200*(xm-xm_m1**2) - 400*(xm_p1 - xm
19     **2)*xm - 2*(1-xm)
20     der[0] = -400*x[0]*(x[1]-x[0]**2) - 2*(1-x[0])
21     der[-1] = 200*(x[-1]-x[-2]**2)
22
23     return der
24 # dimensao do prob
25 n = 2
26
27 # num max iters
28 maxiter = 100000
29 # tolerancia
30 tol = 1e-10
31
32 # aprox. inicial
33 x = np.zeros(n)
```

```
34
35 # iteracoes CG
36 gf = grad(x)
37 d = -gf
38
39 for k in range(maxiter):
40
41     # pesquisa linear
42     alpha = spopt.line_search(fun, grad, x, d)[0]
43
44     # atualizacao
45     x = x + alpha * d
46
47     nad = npla.norm(alpha * d)
48     nfun = npla.norm(fun(x))
49
50     print(f"{k+1}: {alpha:1.2e} {nad:1.2e} {nfun:1.2e}")
51
52     if ((nfun < tol) or np.isnan(nfun)):
53         break
54
55     # prepara nova iter
56     ngf = grad(x)
57
58     beta = np.dot(ngf, ngf) / np.dot(gf, gf)
59     d = -ngf + beta * d
60
61     gf = ngf
```

Exercícios

Em revisão

E.2.2.11. Aplique o método um gradiente conjugado para computar o ponto

mínimo da função de Rosenbrock⁴⁶

$$f(\mathbf{x}) = \sum_{i=1}^n 100 (x_{i+1} - x_i^2)^2 + (1 - x_i)^2 \quad (2.68)$$

com

- a) $n = 2$.
- b) $n = 3$.
- c) $n = 4$.
- d) $n = 5$.
- e) $n = 10$.

E.2.2.12. Aplique o método um gradiente conjugado para computar o ponto mínimo da função de Beale [1]

$$\begin{aligned} f(x, y) = & (1.5 - x + xy)^2 \\ & + (2.25 - x + xy^2)^2 \\ & + (2.625 - x + xy^3)^2. \end{aligned} \quad (2.69)$$

para $\mathbf{x} \in [-4.5, 4.5]^2$.

E.2.2.13. Aplique o método um gradiente conjugado para computar o ponto mínimo da função de Goldstein-Price [4]

$$\begin{aligned} f(x, y) = & \left[1 + (x + y + 1)^2 (19 - 14x \right. \\ & \left. + 3x^2 - 14y + 6xy + 3y^2) \right] \\ & \times \left[30 + (2x - 3y)^2 (18 - 32x \right. \\ & \left. + 12x^2 + 48y - 36xy + 27y^2) \right] \end{aligned} \quad (2.70)$$

E.2.2.14. Aplique o método um gradiente conjugado para computar o ponto mínimo da função de Booth

$$f(x, y) = (x + 2y - 7)^2 + (2x + y - 5)^2 \quad (2.71)$$

para $\mathbf{x} \in [-10, 10]^2$.

E.2.2.15. Aplique um método do gradiente conjugado para computar o ponto mínimo da função de Rastrigin

$$f(x) = 10n + \sum_{i=1}^n [x_i^2 - 10 \cos(2\pi x_i)], \quad (2.72)$$

para $\mathbf{x} \in [-5.12, 5.12]^n$, com

- a) $n = 2$.
- b) $n = 3$.
- c) $n = 4$.
- d) $n = 5$.
- e) $n = 10$.

Capítulo 3

Autovalores e Autovetores

Em revisão

Neste capítulo, estudamos métodos numéricos para a computação de autovalores e autovetores de matrizes.

3.1 Método da Potência

Em revisão

O método da potência é uma técnica iterativa para computar o autovalor dominante de uma matriz e um autovetor associado. Modificações do método, tornam possível sua aplicação para a determinação de outros autovalores próximos a um dado número. Desta forma, pode ser utilizá-lo em conjunto com outra técnica de aproximação de autovalores e autovetores.

3.1.1 Autovalor dominante

Em revisão

Vamos denotar o **espectro** de uma dada matriz $A \in \mathbb{C}^{n \times n}$ por

$$\sigma(A) = \{\lambda_1, \lambda_2, \dots, \lambda_n\}, \quad (3.1)$$

sendo $\lambda_i \in \mathbb{C}$ o i -ésimo **autovalor** de uma **matriz diagonalizável**¹ $A = [a_{i,j}]_{i,j=1}^{n,n}$, i.e. existe um vetor $0 \neq v^{(i)} \in \mathbb{C}^n$ tal que

$$Av^{(i)} = \lambda_i v^{(i)}, \quad (3.2)$$

sendo $v^{(i)}$ chamado de **autovetor** associado a λ_i . No desenvolvimento do **Método da Potência**, vamos assumir que os autovalores podem ser ordenados da seguinte forma

$$|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n|. \quad (3.3)$$

Assim sendo, o λ_1 é chamado de **autovalor dominante**. Na sequência, vamos denotar por A^H a matriz adjunta² de A , i.e. $A^H = [\bar{a}_{j,i}]_{i,j=1}^{n,n}$.

A **iteração** do Método da Potência consiste em

$$z^{(k)} = Aq^{(k-1)}, \quad (3.4)$$

$$q^{(k)} = z^{(k)} / \|z^{(k)}\|, \quad (3.5)$$

$$\nu_k = (q^{(k)})^H A q^{(k)}, \quad (3.6)$$

com dada aproximação inicial $q^{(0)}$ para o autovetor associado a λ_1 . Quando o método é bem sucedido, tem-se $\nu_k \rightarrow \lambda_1$ quando $k \rightarrow \infty$.

Para mostrar a **convergência** do método, basta mostrar que $q^{(k)}$ converge para um autovetor associado a λ_1 . Primeiramente, notemos que³

$$q^{(k)} = \frac{A^k q^{(0)}}{\|A^k q^{(0)}\|}, \quad k \geq 1. \quad (3.7)$$

Como A é diagonalizável, temos que existe uma base $\{v^{(1)}, v^{(2)}, \dots, v^{(n)}\}$ de $\mathbb{C}^{n \times n}$ formada apenas de autovetores de A . Segue que

$$q^{(0)} = \sum_{i=1}^n \alpha_i v^{(i)}, \quad (3.8)$$

onde $\alpha_i \in \mathbb{C}$. Vamos assumir que $\alpha_1 \neq 0$ ⁴. Ainda, $Av^{(i)} = \lambda_i v^{(i)}$, donde

$$A^k q^{(0)} = \sum_{i=1}^n \alpha_i A^k v^{(i)} \quad (3.9)$$

¹Existe uma base de $\mathbb{C}^{n \times n}$ formada apenas de autovetores de A .

²Também, chamada de matriz conjugada transposta de A .

³Segue por indução matemática.

⁴Condição necessária para a convergência.

$$= \sum_{i=1}^n \alpha_i \lambda_i^k v^{(i)} \quad (3.10)$$

$$= \alpha_1 \lambda_1^k \left[v^{(1)} + \sum_{i=2}^n \frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k v^{(i)} \right] \quad (3.11)$$

Como $|\lambda_i|/|\lambda_1| < 1$, $i = 2, \dots, n$, temos que

$$y^{(k)} = \sum_{i=2}^n \frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k v^{(i)} \rightarrow 0, \quad (3.12)$$

quando $k \rightarrow \infty$. Logo, temos que

$$q^{(k)} = \frac{\alpha_1 \lambda_1^k (v^{(1)} + y^{(k)})}{\|\alpha_1 \lambda_1 (v^{(1)} + y^{(k)})\|} \quad (3.13)$$

$$= \frac{\lambda_1}{|\lambda_1|} \frac{(v^{(1)} + y^{(k)})}{\|v^{(1)} + y^{(k)}\|}. \quad (3.14)$$

Por fim, observamos que $\lambda_1/|\lambda_1|$ tem o mesmo sinal de λ_1 . Portanto, $q^{(k)}$ tende a um múltiplo não nulo de $v^{(1)}$ quando $k \rightarrow \infty$.

Observação 3.1.1 (Taxa de convergência.). Pode-se mostrar a seguinte taxa de convergência para o Método da Potência

$$\|\tilde{q}^{(k)} - v^{(1)}\| \leq C \left| \frac{\lambda_2}{\lambda_1} \right|^k, \quad k \geq 1, \quad (3.15)$$

onde

$$\tilde{q}^{(k)} = v^{(1)} + \sum_{i=2}^n \frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k v^{(i)}. \quad (3.16)$$

Consulte [8, Seção 5.3.].

Código 3.1: Algoritmo do Método da Potência

```
1 import numpy as np
2 import numpy.linalg as npla
3
4 def metPot(A, q0, maxiter=1000, tol=1e-14):
```



```
5 q = q0/npla.norm(q0)
6 nu0 = np.dot(q, A @ q)
7 print(f"{0}: {nu0}")
8
9 info = -1
10 for k in range(maxiter):
11     z = A @ q
12     q = z/npla.norm(z)
13     nu = np.dot(q, A @ q)
14
15     print(f"{k+1}: {nu}")
16     if (np.fabs(nu-nu0) < tol):
17         info = 0
18         break
19
20     nu0 = nu
21
22 return nu, q, info
```

E.3.1.1. Use o Método da Potência para computar o autovalor dominante da matriz de coeficientes do problema discreto associado ao Exercício 1.1.3. Estude sua convergência para diferentes tamanhos da matriz.

E.3.1.2. Use o Método da Potência para computar o autovalor dominante da matriz de coeficientes do problema discreto associado ao Exemplo 1.1.2. Estude sua convergência para diferentes tamanhos da matriz.

3.1.2 Método da Potência Inverso

Em revisão

Esta variação do Método da Potência nos permite computar o autovalor mais próximo de um dado número $\mu \in \mathbb{C}$, $\mu \notin \sigma(A)$. A ideia é aplicar o método para a matriz

$$M_{\mu}^{-1} = (A - \mu I)^{-1}. \quad (3.17)$$

Neste contexto, μ é chamado de **deslocamento** (em inglês, *shift*).

Notemos que se (λ_i, v_i) é um autopar de A , então $\xi_i = (\lambda_i - \mu)^{-1}$ é autovalor de M_μ^{-1} associado ao autovetor v_i . De fato,

$$(\lambda_i - \mu)v_i = (A - \mu I)v_i \quad (3.18)$$

$$M_\mu^{-1}(\lambda_i - \mu)v_i = v_i \quad (3.19)$$

$$M_\mu^{-1}v_i = (\lambda_i - \mu)^{-1}v_i. \quad (3.20)$$

Isso também mostra que A e M_μ^{-1} têm os mesmos autovetores.

Agora, se μ for suficientemente próximo de λ_m , autovalor simples de A , então

$$|\lambda_m - \mu| < |\lambda_i - \mu|, \quad i = 1, 2, \dots, n, \quad i \neq m. \quad (3.21)$$

Com isso, $\xi_i = (\lambda_m - \mu)^{-1}$ é o autovalor dominante de M_μ^{-1} e, portanto, a iteração do Método da Potência aplicada a M_μ^{-1} fornece este autovalor. Mais especificamente, como A e M_μ^{-1} tem os mesmos autovetores, a **iteração do Método da Potência Inverso** é dada como segue

$$(A - \mu I)z^{(k)} = q^{(k-1)}, \quad (3.22)$$

$$q^{(k)} = z^{(k)} / \|z^{(k)}\|, \quad (3.23)$$

$$\nu_k = (q^{(k)})^H A q^{(k)}. \quad (3.24)$$

E.3.1.3. Use o Método da Potência para computar diferentes autovalores da matriz de coeficientes do problema discreto associado ao Exercício 1.1.3. Estude sua convergência para diferentes tamanhos da matriz.

E.3.1.4. Use o Método da Potência para computar diferentes autovalores da matriz de coeficientes do problema discreto associado ao Exemplo 1.1.2. Verifique se há vantagem em aplicar os métodos GMRES e GC na resolução de (3.22).

3.2 Iteração QR

Em revisão

A iteração QR é um método para a computação aproximada de todos os autovalores de uma dada matriz A . A ideia é explorar um método iterativo para a computação da **decomposição de Schur**⁵ de A , i.e. encontrar uma **matriz unitária** U ⁶ tal que

$$U^H A U = \begin{bmatrix} \lambda_1 & t_{12} & \cdots & t_{1n} \\ 0 & \lambda_2 & & t_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & \vdots & 0 & \lambda_n \end{bmatrix} \quad (3.25)$$

Assumindo $A \in \mathbb{R}^{n \times n}$, sejam $Q^{(0)} \in \mathbb{R}^{n \times n}$ uma **matriz ortogonal**⁷ e $T^{(0)} = (Q^{(0)})^T A Q^{(0)}$. A iteração QR consiste em

determinar $Q^{(k)}, R^{(k)}$ tal que

$$Q^{(k)} R^{(k)} = T^{(k-1)} \quad (\text{fatoração QR}) \quad (3.26)$$

$$T^{(k)} = R^{(k)} Q^{(k)}, \quad (3.27)$$

para $k = 1, 2, \dots$

Ou seja, a cada iteração k , computa-se a fatoração da matriz $T^{(k-1)}$ como o produto de uma matriz ortogonal $Q^{(k)}$ com uma matriz triangular superior $R^{(k)}$. Então, computa-se uma nova aproximação $T^{(k)}$ pela multiplicação de $R^{(k)}$ por $Q^{(k)}$. Com isso, temos

$$T^{(k)} = R^{(k)} Q^{(k)} \quad (3.28)$$

$$= (Q^{(k)})^T (Q^{(k)} R^{(k)}) Q^{(k)} \quad (3.29)$$

$$= (Q^{(k)})^T T^{(k-1)} Q^{(k)} \quad (3.30)$$

$$= (Q^{(k)})^T R^{(k-1)} Q^{(k-1)} Q^{(k)} \quad (3.31)$$

$$= (Q^{(k)})^T (Q^{(k-1)})^T Q^{(k-1)} R^{(k-1)} Q^{(k-1)} Q^{(k)} \quad (3.32)$$

$$= (Q^{(k-1)} Q^{(k)})^T T^{(k-2)} (Q^{(k-1)} Q^{(k)}) \quad (3.33)$$

$$= (Q^{(0)} \dots Q^{(k)})^T A (Q^{(0)} \dots Q^{(k)}). \quad (3.34)$$

⁵Issai Schur, 1875 - 1941, matemático russo-alemão. Fonte: [Wikipédia](#).

⁶Uma matriz U é dita unitária quando $U^{-1} = U^H$.

⁷Uma matriz Q é dita ortogonal quando $Q^T Q = I$.

Observação 3.2.1 (Convergência do método QR). Se $A \in \mathbb{R}^{n \times n}$ for uma matriz com autovalores tais que

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|, \quad (3.35)$$

então

$$\lim_{k \rightarrow \infty} T^{(k)} = \begin{bmatrix} \lambda_1 & t_{12} & \cdots & t_{1n} \\ 0 & \lambda_2 & & t_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & \vdots & 0 & \lambda_n \end{bmatrix}. \quad (3.36)$$

Para a taxa de convergência, temos

$$|t_{i,i-1}^{(k)}| = \mathcal{O} \left(\left| \frac{\lambda_i}{\lambda_{i-1}} \right|^k \right), \quad i = 2, \dots, n, \quad (3.37)$$

quando $k \rightarrow \infty$.

Ainda, se A for uma **matriz simétrica**, então $T^{(k)}$ tende a uma matriz diagonal quando $k \rightarrow \infty$.

Observação 3.2.2. Variantes do método QR permitem sua aplicação para matrizes mais gerais. Consulte [5].

Uma forma eficiente do método QR é chamada de **iteração Hessenberg⁸-QR**. Consiste em inicializar $T^{(0)}$ como uma **matriz de Hessenberg superior**, i.e. $t_{i,j}^{(0)} = 0$ para $i > j + 1$. A computação de $T^{(0)}$ é feito com **matrizes de Householder⁹** e a fatoração QR de $T^{(k)}$ utiliza de **matrizes de Givens¹⁰**.

Código 3.2: Algoritmo da Iteração QR

```
1 import numpy as np
2 import numpy.linalg as npla
```

⁸Karl Adolf Hessenberg, 1904 - 1959, matemático e engenheiro alemão. Fonte: [Wikipédia](#).

⁹Alston Scott Householder, 1904 - 1993, matemático estadunidense. Fonte: [Wikipédia](#).

¹⁰James Wallace Givens Jr., 1910 - 1993, matemático estadunidense. Fonte: [Wikipédia](#).

```

3
4 def qr_iter(A, Q=None, maxiter=10, tol=1e-5):
5     Q = np.eye(A.shape[0]) if Q==None else Q
6     T0 = Q.T @ A @ Q
7     info = -1
8     for k in range(maxiter):
9         Q, R = npla.qr(T0)
10        T = R @ Q
11        if (npla.norm(T-T0) < tol):
12            info = 0
13            break
14        T0 = T
15    return T, info

```

Observação 3.2.3 (Computação dos autovalores). Se v é autovetor associado ao autovalor simples λ de A , então

$$Av = \lambda v \quad (3.38)$$

$$(Q^{(k)})^T A Q^{(k)} (Q^{(k)})^T v \approx \lambda (Q^{(k)})^T v \quad (3.39)$$

Denotando, $y = (Q^{(k)})^T v$, temos

$$T^{(k)} y = \lambda y. \quad (3.40)$$

Com isso, podemos computar y e, então, temos $v \approx Q^{(k)} y$.

E.3.2.1. Use a iteração QR para computar os autovalores da matriz de coeficientes do problema discreto associado ao Exercício 1.1.3.

E.3.2.2. Use a iteração QR para computar os autovalores da matriz de coeficientes do problema discreto associado ao Exemplo 1.1.2. Use `spla.hessenberg` para computar $T^{(0)}$.

Capítulo 4

Integração

[Vídeo] | [Áudio] | [\[Contatar\]](#)

Neste capítulo, estudamos métodos numéricos para a integração de funções reais.

4.1 Integração Autoadaptativa

Vamos considerar o problema de integrar

$$I(a, b) = \int_a^b f(x) dx \quad (4.1)$$

pela **Regra de Simpson**¹. Em um dado subintervalo $[\alpha, \beta] \subset [a, b]$, temos

$$I(\alpha, \beta) = \underbrace{\frac{h_0}{3} [f(\alpha) + 4f(\alpha + h_0) + f(\beta)]}_{S(\alpha, \beta)} - \frac{h_0^5}{90} f^{(4)}(\xi), \quad (4.2)$$

onde $h_0 = (\beta - \alpha)/2$ e $\xi \in (\alpha, \beta)$. Ou seja, temos que

$$I(\alpha, \beta) - S(\alpha, \beta) = -\frac{h_0^5}{90} f^{(4)}(\xi). \quad (4.3)$$

¹Consulte mais sobre a Regra de Simpson em [Seção 10.1 Regras de Newton-Cotes](#).

A ideia é explorarmos esta informação de forma a obtermos uma estimativa para o erro de integração no intervalo $[\alpha, \beta]$ sem necessitar computar $f^{(4)}$.

Aplicando a Regra de Simpson na partição $[\alpha, (\alpha + \beta)/2] \cup [(\alpha + \beta)/2, \beta]$, obtemos

$$I(\alpha, \beta) - S_2(\alpha, \beta) = -\frac{(h_0/2)^5}{90} \left(f^{(4)}(\xi) + f^{(4)}(\eta) \right), \quad (4.4)$$

onde $\xi \in (\alpha, (\alpha + \beta)/2)$, $\eta \in ((\alpha + \beta)/2, \beta)$ e

$$S_2(\alpha, \beta) = S(\alpha, (\alpha + \beta)/2) + S((\alpha + \beta)/2, \beta). \quad (4.5)$$

Agora, **vamos assumir que** $f^{(4)}(\xi) \approx f^{(4)}(\eta)$ de forma que temos

$$I(\alpha, \beta) - S_2(\alpha, \beta) \approx -\frac{1}{16} \frac{h_0^5}{90} f^{(4)}(\xi). \quad (4.6)$$

De (4.3) e (4.6), obtemos

$$\frac{h_0^5}{90} f^{(4)}(\xi) \approx \frac{16}{15} \underbrace{[S(\alpha, \beta) - S_2(\alpha, \beta)]}_{\mathcal{E}(\alpha, \beta)}. \quad (4.7)$$

Isto nos fornece a seguinte estimativa *a posteriori* do erro

$$|I(\alpha, \beta) - S_2(\alpha, \beta)| \approx \frac{|\mathcal{E}(\alpha, \beta)|}{15}. \quad (4.8)$$

Na prática, costuma-se utilizar a seguinte estimativa mais restrita

$$|I(\alpha, \beta) - S_2(\alpha, \beta)| \approx \frac{|\mathcal{E}(\alpha, \beta)|}{10}. \quad (4.9)$$

Para garantir uma precisão global em $[a, b]$ igual a uma dada tolerância, é suficiente impor que

$$\frac{|\mathcal{E}(\alpha, \beta)|}{10} \leq \epsilon \frac{\beta - \alpha}{b - a}. \quad (4.10)$$

Código 4.1: Algoritmo Simpson Autoadaptativo

```
1 def simpson(f, a, b):
2     return (b-a)/6. * (f(a) + 4*f((a+b)/2.) + f(b))
```

```
3
4 def simad(f,a,b,tol=1e-8,hmin=1e-10):
5     # intervalo calculado
6     S = (a,a)
7     # intervalo a ser calculado
8     N = (a,b)
9     # intervalo ativo
10    A = N
11    # aprox calculada em [a, b]
12    J = 0.
13    # aprox calculada em A
14    JA = 0.
15    # num f eval
16    nfe = 0
17    # info
18    info = 0
19    while (S != (a,b)):
20        print(S,N,A)
21        # tamanho do intervalo
22        h = A[1]-A[0]
23        while (h > hmin):
24            J0 = simpson(f,A[0],A[1])
25            JA = simpson(f,A[0],A[0]+h/2.)
26            JA += simpson(f,A[0]+h/2.,A[1])
27            nfe += 9
28            # est erro
29            est = np.fabs(J0-JA)/10.
30            # criterio de parada
31            if (est < tol*h/(b-a)):
32                break
33            else:
34                A = (A[0],A[0]+h/2.)
35                h = h/2.
36                print("\t",S,N,A,h,est)
37                if (h < hmin):
38                    print("Atencao! h < hmin !")
39                    info = -1
40                    break
```



```

41     J += JA
42     S = (a, A[1])
43     N = (A[1], b)
44     A = N
45     return J, nfe, info

```

Exemplo 4.1.1.

$$\int_{-3}^4 \operatorname{arctg}(10x) dx = -3 \operatorname{arctg}(30) - \frac{\ln(1601)}{20} + \frac{\ln(901)}{20} + 4 \operatorname{arctg}(40) \quad (4.11)$$

$$\approx 1.54203622 \quad (4.12)$$

E.4.1.1. Implemente uma abordagem autoadaptativa usando a Regra do Trapézio. Valide-a e compare com o exemplo anterior.

4.2 Integrais múltiplas

Vamos trabalhar com métodos para a computação de integrais múltiplas

$$\int \int_R f(x, y) dA. \quad (4.13)$$

Em uma região retangular $A = [a, b] \times [c, d]$, podemos reescrevê-la como uma **integral iterada**

$$\int \int_R f(x, y) dA = \int_a^b \int_c^d f(x, y) dy dx. \quad (4.14)$$

4.2.1 Regras de Newton-Cotes

Regra do Trapézio

A Regra do Trapézio² nos fornece

$$\int_c^d f(x, y) dy = \frac{h_y}{2} [f(x, c) + f(x, d)] - \frac{h_y^3}{12} f''(x, \eta) \quad (4.15)$$

²Notas de Aula - Matemática Numérica.

com $h_y = (d - c)$ e $\eta \in (c, d)$. De forma iterada, temos

$$\int_a^b \int_c^d f(x, y) dy dx = \frac{h_y}{2} \int_a^b f(x, c) dx + \frac{h_y}{2} \int_a^b f(x, d) dx \quad (4.16)$$

$$- \frac{h_y^3}{12} \int_a^b f''(x, \eta) dx. \quad (4.17)$$

Então, à exceção do termo do erro, aplicamos a Regra do Trapézio para as integrais em x . Obtemos

$$\int_a^b \int_c^d f(x, y) dy dx = \frac{h_y}{2} \frac{h_x}{2} [f(a, c) + f(b, c)] \quad (4.18)$$

$$+ \frac{h_y}{2} \frac{h_x}{2} [f(a, d) + f(b, d)] \quad (4.19)$$

$$- \frac{h_y}{2} \frac{h_x^3}{12} f''(\mu', c) \quad (4.20)$$

$$- \frac{h_y}{2} \frac{h_x^3}{12} f''(\mu'', d) \quad (4.21)$$

$$- \frac{h_y^3}{12} \int_a^b f''(x, \eta) dx, \quad (4.22)$$

com $h_x = (b - a)$, $\mu', \mu'' \in (a, b)$. Pelos Teorema do Valor Intermediário e pelo Teorema do Valor Médio, podemos ver que o erro é $O(h_x h_y^3 + h_x^3 h_y)$. Por fim, obtemos a Regra do Trapézio para Integrais Iteradas

$$\int_a^b \int_c^d f(x, y) dy dx = \frac{h_y}{2} \frac{h_x}{2} [f(a, c) + f(b, c) + f(b, d) + f(a, d)] \quad (4.23)$$

$$+ O(h_x h_y^3 + h_x^3 h_y). \quad (4.24)$$

Exemplo 4.2.1. A Regra do Trapézio fornece

$$\int_{1.5}^2 \int_1^{1.5} \ln(x + 2y) dy dx \approx 0.36. \quad (4.25)$$

Verifique!

Regra de Simpson

A Regra do Simpson³ nos fornece

$$\int_c^d f(x, y) dy = \frac{h_y}{3} [f(x, y_1) + 4f(x, y_2) + f(x, y_3)] \quad (4.26)$$

³Notas de Aula - Matemática Numérica.

$$- \frac{h_y^5}{90} f^{(4)}(x, \eta) \quad (4.27)$$

com $h_y = (d - c)/2$, $y_j = (j - 1)h_y$, $j = 1, 2, 3$, e $\eta \in (c, d)$. De forma iterada, temos

$$\int_a^b \int_c^d f(x, y) dy dx = \frac{h_y}{3} \left[\int_a^b f(x, y_1) dx + 4 \int_a^b f(x, y_2) dx + \int_a^b f(x, y_3) dx \right] \quad (4.28)$$

$$- \frac{h_y^5}{90} \int_a^b f^{(4)}(x, \eta) dx \quad (4.29)$$

Então, à exceção do termo do erro, aplicamos a Regra de Simpson para as integrais em x . Obtemos

$$\int_a^b \int_c^d f(x, y) dy dx = \frac{h_x h_y}{9} [f(x_1, y_1) + 4f(x_2, y_1) + f(x_3, y_1)] \quad (4.30)$$

$$+ \frac{4h_x h_y}{9} [f(x_1, y_2) + 4f(x_2, y_2) + f(x_3, y_2)] \quad (4.31)$$

$$+ \frac{h_x h_y}{9} [f(x_1, y_3) + 4f(x_2, y_3) + f(x_3, y_3)] \quad (4.32)$$

$$- \frac{h_x^5 h_y}{270} f^{(4)}(\mu_1, y_1) \quad (4.33)$$

$$- \frac{4h_x^5 h_y}{270} f^{(4)}(\mu_2, y_2) \quad (4.34)$$

$$- \frac{h_x^5 h_y}{270} f^{(4)}(\mu_3, y_3) \quad (4.35)$$

$$- \frac{h_y^5}{90} \int_a^b f^{(4)}(x, \eta) dx \quad (4.36)$$

com $h_x = (b - a)/2$, $\mu_1, \mu_2, \mu_3 \in (a, b)$. Pelos Teorema do Valor Intermediário e Teorema do Valor Médio, podemos ver que o erro é $O(h_x h_y^5 + h_x^5 h_y)$. Por fim, obtemos a **Regra de Simpson para Integrais Iteradas**

$$\int_a^b \int_c^d f(x, y) dy dx = \frac{h_x h_y}{9} [f(x_1, y_1) + 4f(x_2, y_1) + f(x_3, y_1)] \quad (4.37)$$

$$+ \frac{4h_x h_y}{9} [f(x_1, y_2) + 4f(x_2, y_2) + f(x_3, y_2)] \quad (4.38)$$

$$+ \frac{h_x h_y}{9} [f(x_1, y_3) + 4f(x_2, y_3) + f(x_3, y_3)] \quad (4.39)$$

$$+ O(h_x h_y^5 + h_x^5 h_y). \quad (4.40)$$

Exemplo 4.2.2. A Regra de Simpson fornece

$$\int_{1.5}^2 \int_1^{1.5} \ln(x + 2y) dy dx \approx 0.361003. \quad (4.41)$$

Verifique!

4.2.2 Regras Compostas de Newton-Cotes

A ideia é particionar a região de integração em células e o resultado da integração é a soma da aplicação da regra de quadratura em cada uma das células.

Regra Composta do Trapézio

Para uma região retangular $R = [a, b] \times [c, d]$, vamos construir a malha

$$M = \{c_k = [x_i, x_{i+1}] \times [y_j, y_{j+1}] : k = i + (j - 1)n_x\}, \quad (4.42)$$

onde $x_i = (i - 1)h_x$, $h_x = (b - a)/n_x$, $i = 1, 2, \dots, n_x + 1$ e $y_j = (j - 1)h_y$, $h_y = (d - c)/n_y$, $j = 1, 2, \dots, n_y + 1$. Consulte a Figura 4.1.

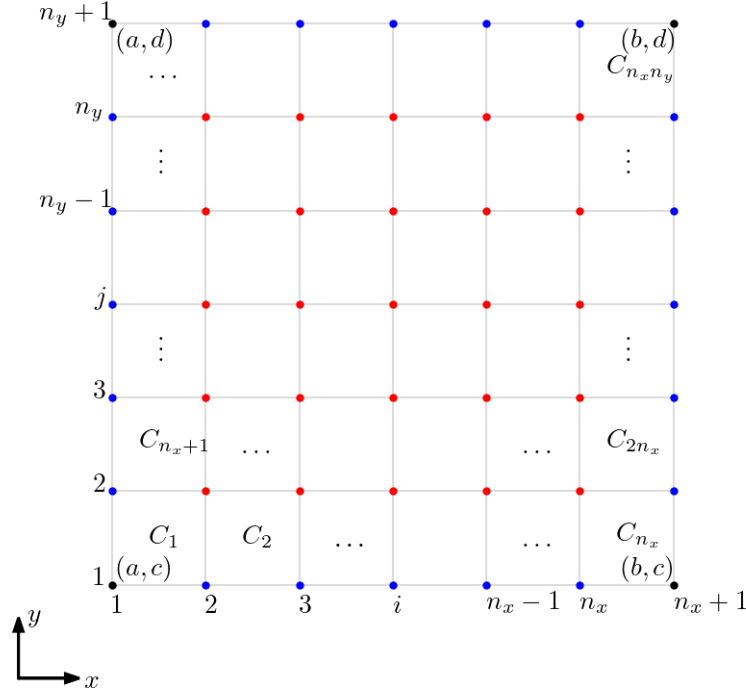


Figura 4.1: Representação da malha para a Regra Composta do Trapézio.

Aplicando a ideia, temos

$$\int_a^b \int_c^d f(x, y) dy dx = \sum_{k=1}^{n_x n_y} \int \int_{C_k} f(x, y) dy dx \quad (4.43)$$

$$= \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \int_{x_i}^{x_{i+1}} \int_{y_j}^{y_{j+1}} f(x, y) dy dx \quad (4.44)$$

Em cada integral em C_k , aplicamos a Regra do Trapézio, segue

$$\int_{x_i}^{x_{i+1}} \int_{y_j}^{y_{j+1}} f(x, y) dy dx \approx \frac{h_y}{2} \frac{h_x}{2} [f(x_i, y_j) + f(x_{i+1}, y_j) \quad (4.45)$$

$$+ f(x_{i+1}, y_{j+1}) + f(x_i, y_{j+1})] \quad (4.46)$$

Observamos que nos conjuntos de nodos (marcados em azul na Figura 4.1)

$$\{(i, j) : i = 2, \dots, n_x, j = 1 \text{ ou } j = n_y + 1\}, \quad (4.47)$$

$$\{(i, j) : i = 1 \text{ ou } i = n_x + 1, j = 2, \dots, n_y\} \quad (4.48)$$

a função integranda será avaliada 2 vezes. Já, em todos os nodos internos, $i = 2, \dots, n_x$, $j = 2, \dots, n_y$, a função será avaliada 4 vezes. Com isso, chegamos à Regra Composta do Trapézio

$$\begin{aligned} \int_a^b \int_c^d f(x, y) dy dx &= \frac{h_x h_y}{4} [f(x_1, y_1) + f(x_{n_x+1}, y_1) \\ &\quad + f(x_{n_x+1}, y_{n_y+1}) + f(x_1, y_{n_y+1})] \\ &\quad + \frac{h_x h_y}{2} \sum_{i=2}^{n_x} [f(x_i, y_1) + f(x_i, y_{n_y+1})] \\ &\quad + \frac{h_x h_y}{2} \sum_{j=2}^{n_y} [f(x_1, y_j) + f(x_{n_x+1}, y_j)] \\ &\quad + h_x h_y \sum_{i=2}^{n_x} \sum_{j=2}^{n_y} f(x_i, y_j) \\ &\quad + O(h_x^2 + h_y^2) \end{aligned} \quad (4.49)$$

Observamos que esta é uma quadratura de $(n_x + 1)(n_y + 1)$ nodos.

E.4.2.1. Verifique a aplicação da Regra Composta do Trapézio para computar

$$\int_{1.5}^2 \int_1^{1.5} \ln(x + 2y) dy dx. \quad (4.50)$$

Regra Composta de Simpson

Aqui, vamos construir uma malha

$$M = \{C_k = [x_i, x_{i+2}] \times [y_j, y_{j+2}] : i = 1, 3, \dots, n_x - 1, j = 1, 3, \dots, n_y - 1, \}, \quad (4.51)$$

onde $x_i = (i - 1)h_x$, $h_x = (b - a)/n_x$, $i = 1, 2, \dots, n_x + 1$ e $y_j = (j - 1)h_y$, $h_y = (d - c)/n_y$, $j = 1, 2, \dots, n_y + 1$. Com $n_x, n_y \leq 2$ números pares. Consulte a Figura 4.2.

$$\begin{aligned}
& + 4 \sum_{i=1}^{n_x/2-1} \sum_{j=1}^{n_y/2-1} f(x_{2j+1}, y_{2j+1}) \\
& + 8 \sum_{i=1}^{n_x/2-1} \sum_{j=1}^{n_y/2} f(x_{2i+1}, y_{2j}) \\
& + 8 \sum_{i=1}^{n_x/2} \sum_{j=1}^{n_y/2-1} f(x_{2i}, y_{2j+1}) \\
& + 16 \sum_{i=1}^{n_x/2} \sum_{j=1}^{n_y/2} f(x_{2i}, y_{2j}) \Bigg\} \\
& + O(h_x^4 + h_y^4).
\end{aligned} \tag{4.52}$$

E.4.2.2. Verifique a aplicação da Regra Composta de Simpson para computar

$$\int_{1.5}^2 \int_1^{1.5} \ln(x + 2y) \, dy \, dx. \tag{4.53}$$

Notas

¹Llewellyn Hilleth Thomas, 1903 - 1992, físico e matemático aplicado britânico. Fonte: [Wikipedia](#).

²Llewellyn Hilleth Thomas, 1903 - 1992, físico e matemático aplicado britânico. Fonte: [Wikipedia](#).

³Llewellyn Hilleth Thomas, 1903 - 1992, físico e matemático aplicado britânico. Fonte: [Wikipedia](#).

⁴Llewellyn Hilleth Thomas, 1903 - 1992, físico e matemático aplicado britânico. Fonte: [Wikipedia](#).

⁵Llewellyn Hilleth Thomas, 1903 - 1992, físico e matemático aplicado britânico. Fonte: [Wikipedia](#).

⁶Johann Carl Friedrich Gauss, 1777 - 1855, matemático alemão. Fonte: [Wikipédia: Carl Friedrich Gauss](#).

⁷Siméon Denis Poisson, 1781 - 1840, matemático francês. Fonte: [Wikipédia:Siméon Denis Poisson](#).

⁸Pierre-Simon Laplace, 1749 - 1827, matemático francês. Fonte: [Wikipédia: Pierre-Simon Laplace](#).

⁹CSR e CSC são formatos de matrizes esparsas mais eficientes para a computação matricial.

¹⁰Desenvolvido por Yousef Saad e H. Schultz, 1986. Fonte: [Wikipedia](#).

¹¹Alexei Nikolajewitsch Krylov, 1863 - 1945, engenheiro e matemático russo. Fonte: [Wikipédia](#).

¹²Georgi Iwanowitsch Petrov, 1912 - 1987, engenheiro soviético. Fonte: [Wikipedia](#).

¹³Boris Galerkin, 1871 - 1945, engenheiro e matemático soviético. Fonte: [Wikipédia](#).

¹⁴Walter Edwin Arnoldi, 1917 - 1995, engenheiro americano estadunidense. Fonte: [Wikipédia](#).

¹⁵Jørgen Pedersen Gram, 1850 - 1916, matemático dinamarquês. Fonte: [Wikipédia](#).

¹⁶Erhard Schmidt, 1876 - 1959, matemático alemão. Fonte: [Wikipédia](#).

¹⁷Karl Adolf Hessenberg, 1904 - 1959, engenheiro e matemático alemão. Fonte: [Wikipédia](#).

¹⁸Alston Scott Householder, 1904 - 1993, matemático americano estadunidense. Fonte: [Wikipédia](#).

¹⁹Siméon Denis Poisson, 1781 - 1840, matemático francês. Fonte: [Wikipédia:Siméon Denis Poisson](#).

²⁰Walter Edwin Arnoldi, 1917 - 1995, engenheiro americano estadunidense. Fonte: [Wikipédia](#).

²¹ $x^T Ax > 0$ para todo $x \neq 0$.

²²Iteração do método do máximo declive.

²³[scipy.sparse.linalg.gmres](#)

²⁴[scipy.sparse.linalg.spsolve](#)

²⁵Compare com o funcional J dado em (1.87).

²⁶Mostre que $\langle \cdot, \cdot \rangle_A$ é de fato um produto interno.

²⁷[scipy.sparse.linalg.gmres](#)

²⁸[scipy.sparse.linalg.cg](#)

²⁹Carl Gustav Jakob Jacobi, 1804 - 1851, matemático alemão. Fonte: [Wikipédia: Carl Gustav Jakob Jacobi](#).

³⁰Isaac Newton, 1642 - 1727, matemático, físico, astrônomo, teólogo e autor inglês. Fonte: [Wikipédia: Isaac Newton](#).

³¹Rudolf Otto Sigismund Lipschitz, 1832 - 1903, matemático alemão. Fonte: [Wikipédia](#).

³²Jan Burgers, 1895 - 1981, físico neerlandês. Fonte: [Wikipédia: Jan Burgers](#).

³³Johann Peter Gustav Lejeune Dirichlet, 1805 - 1859, matemático alemão. Fonte: [Wikipédia: Johann Peter Gustav Lejeune Dirichlet](#).

³⁴Erich Hans Rothe, 1895 - 1988, matemático alemão. Fonte: [Wikipédia](#).

³⁵Leonhard Paul Euler, 1707-1783, matemático e físico suíço. Fonte: [Wikipédia: Ronald Fisher](#).

³⁶Isaac Newton, 1642 - 1727, matemático, físico, astrônomo, teólogo e autor inglês. Fonte: [Wikipédia: Isaac Newton](#).

³⁷Carl Gustav Jakob Jacobi, 1804 - 1851, matemático alemão. Fonte: [Wikipédia: Carl Gustav Jakob Jacobi](#).

³⁸Brook Taylor, 1685 - 1731, matemático britânico. Fonte: [Wikipédia: Brook Taylor](#).

³⁹Philip Wolfe, 1927 - 2016, matemático estadunidense. Fonte: [Wikipédia](#).

⁴⁰Howard Harry Rosenbrock, 1920 - 2010, engenheiro britânico. Fonte: [Wikipedia: Howard Harry Rosenbrock](#).

⁴¹Leopold Kronecker, 1823 - 1891, matemático alemão. Fonte: [Wikipédia: Leopold Kronecker](#).

⁴²Howard Harry Rosenbrock, 1920 - 2010, engenheiro britânico. Fonte: [Wikipedia: Howard Harry Rosenbrock](#).

⁴³Isaac Newton, 1642 - 1727, matemático, físico, astrônomo, teólogo e autor inglês. Fonte: [Wikipédia: Isaac Newton](#).

⁴⁴Alexei Nikolajewitsch Krylov, 1863 - 1945, engenheiro e matemático russo. Fonte: [Wikipédia](#).

⁴⁵Howard Harry Rosenbrock, 1920 - 2010, engenheiro britânico. Fonte: [Wikipedia: Howard Harry Rosenbrock](#).

⁴⁶Howard Harry Rosenbrock, 1920 - 2010, engenheiro britânico. Fonte: [Wikipedia: Howard Harry Rosenbrock](#).

Referências

- [1] Beale, E. M. L.. *On Minimizing a Convex Function Subject to Linear Inequalities*, Journal of the Royal Statistical Society: Series B (Methodological), vol. 17, 1995, pp. 173-184. DOI: <https://doi.org/10.1111/j.2517-6161.1955.tb00191.x>
- [2] Burden, R. & Faires, J.D.. *Análise Numérica*, Cengage Learning, 3. ed., 2016. ISBN: 978-8522123407
- [3] Davis, J. P. & Rabinowitz, P.. *Methods of Numerical Integration*, Academic Press, Inc., 2. ed., 1984.
- [4] Goldstein, A. A. & Price, J. F.. *On Descent from Local Minima*, Mathematics of Computation, vol. 25, 1971, pp. 569-574. DOI: <https://doi.org/10.2307/2005219>
- [5] Golub, G. & Van Loan, C.F.. *Matrix Computations*, The Johns Hopkins University Press, 4. ed., 2013.
- [6] Kelley, C.T.. *Iterative Methods for Optimization*, Society for Industrial and Applied Mathematics, 1999.
- [7] Nocedal, J. & Wright, S. J.. *Numerical Optimization*, 2. ed., Springer, 2006. ISBN: 978-0387-30303-1
- [8] Quarteroni, A.; Sacco, R. & Saleri, F.. *Numerical Mathematics*, 2. ed., Springer, 2007. ISBN: 9783540346586
- [9] Saad, Y.. *Iterative Methods for Sparse Linear Systems*, Society for In-

dustrial and Applied Mathematics, 2003.

- [10] Stoer, J.. Approximate Calculation of Multiple Integrals, Prentice-Hall, Inc., 1971.
- [11] Watkins, D.. Fundamentals of Matrix Computations, John Wiley, 2002.

Índice de Comandos

<code>scipy</code>	<code>csc_array</code> , 14
<code>.sparse</code>	<code>csr_array</code> , 13
<code>coo_array</code> , 12	