

Cálculo numérico

S. C. Coutinho

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO, INSTITUTO DE MATEMÁTICA,
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO, P.O. BOX 68530, 21945-970 RIO
DE JANEIRO, RJ, BRAZIL.

Email address: `collier@dcc.ufrj.br`

Aviso e agradecimentos

Todos os capítulos nestas notas são apenas esboços e estão longe de representar uma versão final deste material. Correções e sugestões são muito bem-vindas. Obrigado a

- Alexandre Costard;
- Anna Carolina G. Bittencourt;
- Gabriel Conde;
- Gabriel Vargas Ferreira;
- Ingrid Canaane;
- João Pedro Gomes da Costa
- João Vitor de Oliveira Silva;
- Karina Pereira;
- Lucas Cavalcante Clarino;
- Lucas Clemente;
- Mariana Soares;
- Pedro Paulo Kastrup Ferreira;
- Rafael Vazquez Moraes;
- Stephanie Orazem;
- Thiago B. de Almeida Soares;
- Vinícius Lettieri Proença;
- Victor de Barros Melo;
- Luis Fernando Gonçalves de Faria;
- Eduarda de Souza Marques

pela ajuda em corrigir os erros.

Sumário

Aviso e agradecimentos	iii
Parte 1. Métodos diretos e problemas de valor de contorno	1
Capítulo 1. A ponte pênsil	3
1. A equação da ponte	3
2. Discretizando o problema	7
3. Resolução de sistemas lineares	13
4. Eliminação e substituição	18
5. Recapitulando e olhando adiante	23
Capítulo 2. Aproximação de números e funções	27
1. Representação de números	27
2. Decimais infinitas e erros	30
3. Ponto flutuante	33
4. Newton e a aproximação de funções	39
5. A fórmula de Taylor com resto	41
Exercícios	51
Capítulo 3. O problema de valor de contorno	55
1. Recapitulando e generalizando	55
2. Aproximando derivadas	58
3. O método de diferenças finitas	60
Exercícios	66
Capítulo 4. Decomposição de matrizes	67
1. Matrizes e eliminação	67
2. Matrizes elementares e decomposição LU	70
3. Sistemas lineares e decomposição LU	75
4. Decomposição PLU	78
5. Pivoteamento	84
Capítulo 5. Ajuste de curvas	89
1. Introdução	89
2. Interpolação	92
3. Mínimos quadrados: o enfoque analítico	97

4. Mínimos quadrados: o enfoque geométrico	104
Exercícios	109
Parte 2. Métodos iterativos e problemas de valor inicial	111
Capítulo 6. O pêndulo simples	113
1. O pêndulo	113
2. Problemas de valor inicial e o método de Euler	115
3. Aplicando o método de Euler ao pêndulo	120
4. A solução geral da equação do pêndulo	123
5. Funções e integrais elípticas	126
6. Integrais elípticas de primeira espécie	130
Capítulo 7. Sistemas dinâmicos	135
1. Iterações e pontos fixos	135
2. Existência de pontos fixos e de atratores	140
3. Zeros de funções	145
4. Métodos iterativos para sistemas lineares	152
Capítulo 8. Integração	161
1. Interpolação pelo método de Lagrange	161
2. Interpolação pelo método de Newton e erros	163
3. Integração: regra do trapézio	168
4. Regra de Simpson	173
Capítulo 9. O problema de valor inicial	181
1. O método de Euler Modificado	181
2. Convergência e ordem	187
3. O pêndulo revisitado	195
4. Aplicando o método de Euler ao pêndulo	197
Exercícios	199
Referências Bibliográficas	203

Parte 1

Métodos diretos e problemas de valor de contorno

CAPÍTULO 1

A ponte pênsil

Neste capítulo estudaremos o problema que servirá de inspiração para toda a primeira parte destas notas: a determinação da curva descrita pelo cabo de sustentação de uma ponte pênsil. Começaremos determinando a equação diferencial que descreve a curva desejada. Como esta equação, em geral, não tem solução analítica, veremos como determinar uma aproximação do problema, de modo que pontos ao longo da curva solução possam ser obtidos como soluções de um sistema linear. Para tornar isto viável, precisaremos desenvolver também um procedimento eficiente para resolver sistemas lineares. Encerraremos o capítulo identificando algumas questões relativas à precisão dos resultados obtidos, que serão discutidas em detalhe nos capítulos da primeira parte destas notas.

1. A equação da ponte

Uma ponte pênsil é aquela cujo deque é amarrado a dois cabos, suportados em colunas, nas cabeceiras da ponte. A mais famosa destas pontes é, provavelmente, *Golden Gate Bridge* em São Francisco, Estados Unidos. Nesta seção deduziremos a equação diferencial cuja solução é a curva descrita pelos cabos de sustentação de uma ponte pênsil. Mais precisamente, deduziremos a equação correspondente a um modelo de ponte pênsil. Neste contexto, um *modelo* é uma representação simplificada de um objeto ou de um fenômeno. Um modelo está para o fenômeno ou objeto que representa, assim como um mapa está para a região da qual é a imagem. É precisamente a ausência de detalhes do mapa que nos permite usá-lo para achar nosso caminho em uma cidade; um mapa tão detalhado quanto a cidade, nada ajudaria em nossa orientação. Da mesma maneira, um modelo do comportamento da atmosfera só é útil se nos permite simular o que vai acontecer ao longo do dia de amanhã em menos de vinte e quatro horas. Portanto, os modelos precisam ser suficientemente simples para que possam ser resolvidos com rapidez, mas suficientemente complexos para que sua solução represente uma boa aproximação do problema do qual são a solução.

Por exemplo, o modelo da ponte pênsil, que analisaremos a seguir, pressupõe a completa ausência de vento e de outros fenômenos atmosféricos. As consequências de uma tal suposição para uma ponte pênsil real são bem conhecidas: em 1940 a



FIGURA 1. Golden Gate Bridge

ponte pênsil em *Tacoma Narrows*, nos Estados Unidos, entrou em colapso meses depois de ser inaugurada, por causa das oscilações causadas por ventos de 64 Km/h . A bem da verdade, a ausência de fenômenos atmosféricos é o menor dos nossos problemas porque, para simplificar a equação diferencial, suporemos que o deque da ponte é apoiado em um único cabo, cujo peso é zero. Com isso, precisamos considerar unicamente o peso do deque e dos veículos que estão parados sobre a ponte. Assim, podemos analisar o efeito de um engarrafamento sobre a forma do cabo de sustentação, mas não seu comportamento quando os veículos se movem sobre a ponte. Na prática, a hipótese de que o cabo tem peso nulo é mais razoável do que pode parecer à primeira vista, porque o peso do deque, mesmo quando vazio, supera em muito o peso do cabo.

Suponhamos, então, uma ponte pênsil que consiste de um deque suportado por um único cabo de peso desprezível, como ilustrado na figura 2. Nosso objetivo é encontrar a equação diferencial que descreve a curva formada pelo cabo horizontal ao qual está amarrado o deque da ponte.

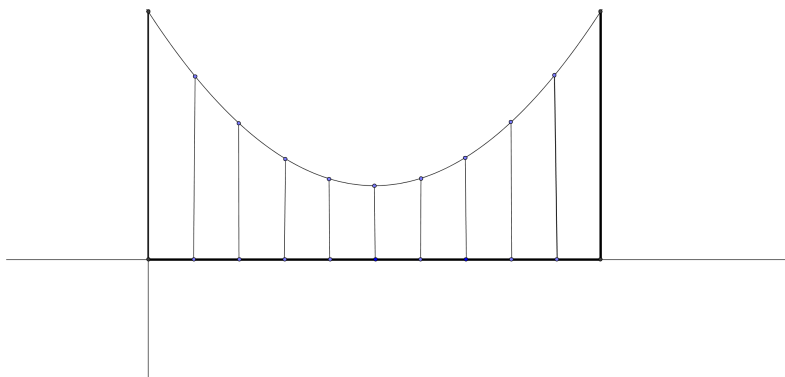


FIGURA 2. Diagrama de uma ponte pênsil

Começaremos a análise investigando as forças que atuam sobre um pequeno segmento do cabo que sustenta a ponte. Na figura 3, os vetores $\mathbf{T}(x)$ e $\mathbf{T}(x + \Delta x)$ denotam as tensões nos pontos de abscissas x e $x + \Delta x$, respectivamente, ao passo que \mathbf{F} corresponde à força exercida sobre o cabo pelo segmento do deque da ponte

entre x e $x + \Delta x$. Note que as tensões são tangentes ao cabo nas extremidades do segmento, ao passo que \mathbf{F} atua ao longo da vertical.

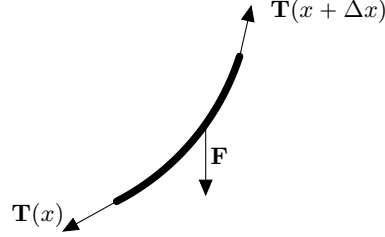


FIGURA 3. Forças no cabo de uma ponte pênsil

Nosso próximo passo consiste em projetar os vetores $\mathbf{T}(x)$, $\mathbf{T}(x + \Delta x)$ e \mathbf{F} ao longo das direções horizontal e vertical. Para isso precisamos conhecer a intensidade das tensões, que denotaremos por $T(x)$ e $T(x + \Delta x)$, além da massa $\rho(x)$ do deque por unidade de comprimento. Observe que estamos admitindo que $\rho(x)$ varie ao longo do deque, para que possamos analisar o que ocorre quando há veículos de diferentes pesos engarrafados ao longo da ponte. Suponhamos, finalmente, que $\varphi(x)$ seja o ângulo entre a tangente ao cabo no ponto de abscissa x e o eixo horizontal, como ilustrado na figura 4. Como \mathbf{F} é vertical, as projeções das tensões ao longo da horizontal devem ser iguais, o que nos dá

$$(1) \quad T(x) \cos(\varphi(x)) = T(x + \Delta x) \cos(\varphi(x + \Delta x)).$$

Por outro lado, o componente vertical de $T(x + \Delta x)$ deve anular a soma de \mathbf{F} com o componente vertical de $T(x)$, donde

$$(2) \quad T(x + \Delta x) \sin(\varphi(x + \Delta x)) = T(x) \sin(\varphi(x)) + g\rho(x)\Delta(x),$$

em que g é a aceleração da gravidade. Segue de (1) que os componentes horizontais das forças que atuam no cabo têm a mesma intensidade T_0 em todos os seus pontos. Reescrevendo (1) em termos de T_0 , obtemos

$$T(x) = \frac{T_0}{\cos(\varphi(x))} \quad \text{e} \quad T(x + \Delta x) = \frac{T_0}{\cos(\varphi(x + \Delta x))}.$$

Substituindo estas duas expressões em (2),

$$\frac{T_0}{\cos(\varphi(x + \Delta x))} \sin(\varphi(x + \Delta x)) = \frac{T_0}{\cos(\varphi(x))} \sin(\varphi(x)) + g\rho(x)\Delta(x);$$

que equivale a

$$(3) \quad T_0 \tan(\varphi(x + \Delta x)) = T_0 \tan(\varphi(x)) + g\rho(x)\Delta(x).$$

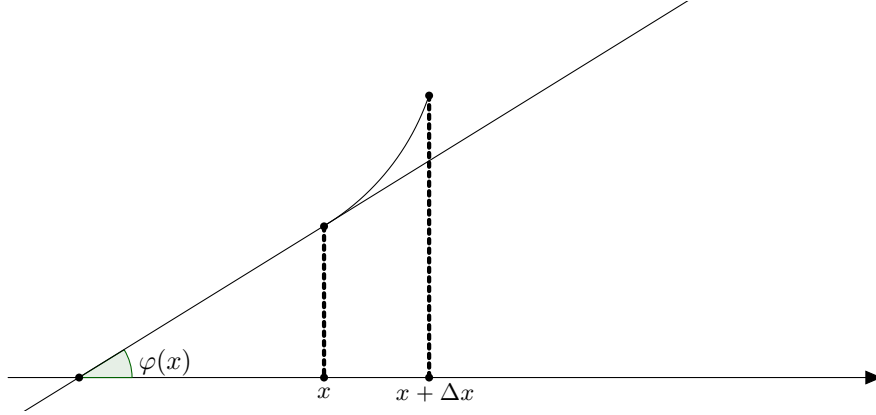


FIGURA 4. $\varphi(x)$ é o ângulo entre a tangente ao cabo em x e a horizontal.

Contudo, se $y = u(x)$ for a equação da curva descrita pelo cabo da ponte, então

$$u'(x) = \tan(\varphi(x)),$$

de modo que (3) pode ser escrita na forma

$$T_0 u'(x + \Delta x) - T_0 u'(x) = g\rho(x)\Delta(x).$$

Dividindo esta última equação por Δx e tomando o limite quando Δx tende a zero, obtemos

$$T_0 u''(x) = T_0 \lim_{\Delta x \rightarrow 0} \frac{u'(x + \Delta x) - u'(x)}{\Delta x} = g\rho(x).$$

Portanto, a função $u(x)$ cujo gráfico corresponde à curva descrita pelo cabo de sustentação da ponte satisfaz a equação diferencial

$$(4) \quad u''(x) = \frac{g}{T_0} \rho(x).$$

Vejamos o que acontece quando a distribuição de massa no deque da ponte é uniforme; isto é, quando $\rho(x) = \rho_0$ é constante. Neste caso a equação diferencial é

$$u''(x) = \frac{g}{T_0} \rho_0,$$

que pode ser facilmente resolvida. Integrando os dois lados desta equação duas vezes, obtemos

$$(5) \quad u(x) = \frac{g\rho_0}{2T_0} x^2 + c_1 x + c_0,$$

em que c_0 e c_1 são as constantes de integração. Para fixar completamente a curva descrita pelo cabo da ponte pênsil que estamos considerando, precisamos de duas condições que nos permitam calcular os valores de c_0 e c_1 . Em nosso modelo estas condições correspondem a dizer que o cabo é amarrado a duas torres verticais de altura h , situadas nas cabeceiras da ponte. Supondo que a ponte tem comprimento ℓ

e posicionando a origem dos eixos na cabeceira esquerda da ponte, as duas condições adicionais são

$$u(0) = a \quad \text{e} \quad u(\ell) = a.$$

Mas, por (5),

$$a = u(0) = c_0 \quad \text{e} \quad a = u(\ell) = \frac{g\rho_0}{2T_0}\ell^2 + c_1\ell + c_0;$$

donde

$$c_0 = a \quad \text{e} \quad c_1 = -\frac{g\rho_0}{2T_0}\ell.$$

Portanto, quando a distribuição de massa no deque da ponte é uniforme, nosso modelo tem como solução a parábola

$$u(x) = \frac{g\rho_0}{2T_0}x^2 - \frac{g\rho_0\ell}{2T_0}x + a.$$

Modelos descritos desta maneira surgem frequentemente em física e engenharia e são conhecidos como *problemas de valor de contorno*. Outros exemplos incluem a deflexão de uma barra cujas extremidades estão fixas e a distribuição de calor em uma barra metálica cujas extremidades são mantidas a uma temperatura constante. Todos os problemas de valor de contorno que estudaremos serão definidos por uma equação diferencial linear de segunda ordem e pelos valores constantes que a função assume nas extremidades do seu domínio. Infelizmente, mesmo uma equação como a da curva descrita pelo cabo da ponte em nosso modelo altamente simplificado pode não ter solução analítica. Isto ocorre, por exemplo, quando a ponte é muito longa e a massa dos veículos engarrafados sobre ela aumenta em direção ao centro da ponte, sendo descrita por $\rho(x) = \exp(-sx(x - \ell))$, em que s é um número real. Quanto maior for s , mais concentrado estará o peso no meio da ponte. Neste caso, a única saída para encontrar a curva descrita pela solução do problema de valor de contorno é apelar para aproximações numéricas da solução.

2. Discretizando o problema

Vimos na seção anterior que, sob algumas hipóteses simples, a curva descrita pelo cabo de sustentação de uma ponte pênsil é o gráfico da função $u(x)$ definida pelo problema de valor de contorno

$$(6) \quad u''(x) = \frac{g}{T_0}\rho(x), \quad u(0) = a \quad \text{e} \quad u(\ell) = a,$$

em que ℓ é o comprimento da ponte, a é a altura das colunas que apoiam o cabo de sustentação, g é a aceleração da gravidade, T_0 é a intensidade do componente horizontal da tensão no cabo e $\rho(x)$ é a distribuição de massa do deque por unidade de comprimento.

Para que a equação acima faça sentido, é necessário que $u(x)$ tenha segunda derivada; em particular, $u(x)$ tem que ser uma função contínua. Contudo, o espaço de memória em um computador é finito e não nos permite calcular $u(x)$ para todos os valores de x , a não ser nos raros casos em que a equação pode ser resolvida analiticamente. Contornaremos este problema calculando uma quantidade finita pontos que estejam suficientemente próximos da curva solução do problema. Quanto mais pontos calcularmos, melhor será a aproximação.

Infelizmente isto não resolve completamente nosso problema porque, na equação diferencial (6), aparece uma segunda derivada. Mas derivadas são calculadas usando limites que, por sua vez, requerem uma variável contínua. A saída é, mais uma vez, recorrer a aproximações finitas. Assim, como a primeira derivada de $u(x)$ é igual a

$$u'(x) = \lim_{\Delta x \rightarrow 0} \frac{u(x + \Delta x) - u(x)}{\Delta x},$$

sabemos que o quociente de Newton

$$(7) \quad \frac{u(x + \Delta x) - u(x)}{\Delta x}$$

estará tão mais próximo de $u'(x)$ quanto menor for o valor de Δx . Isto sugere que, tomando Δx suficientemente pequeno, o quociente (7) deveria nos dar uma boa aproximação para $u'(x)$. Quão pequeno Δx deve ser, vai depender da precisão com que o problema precisa ser resolvido. Por exemplo, se $u(x) = 10x^2$ a fórmula (7) nos dá

$$\frac{u(\Delta x) - u(0)}{\Delta x} = \frac{10(\Delta x)^2}{\Delta x} = 10\Delta x$$

como aproximação para

$$f'(0) = 0,$$

de modo que, para que o erro não seja maior que 10^{-k} , devemos tomar $\Delta x < 10^{-k-1}$.

Antes de poder continuar, precisamos organizar o que dissemos de maneira um pouco mais sistemática. Para começar, escolhemos em quantas partes o intervalo $[0, \ell]$ deve ser dividido. Digamos que sejam n partes, para que não precisemos nos comprometer com nenhum número específico. Tomando $h = \ell/n$, nossa meta é determinar aproximações y_j para os valores de $u(x_j)$ nos pontos $x_j = j \cdot h$, para $j = 0, \dots, n$. Lembre-se que, pelas condições de contorno.

$$u(x_0) = u(0) = a \quad \text{e} \quad u(x_n) = u(\ell) = a.$$

Em seguida, aproximamos as derivadas primeiras usando o quociente de Newton com $\Delta x = h$. Como

$$x_j + h = jh + h = (j + 1)h = x_{j+1},$$

temos que

$$u'(x_j) \approx \frac{u(x_j + h) - u(x_j)}{h} = \frac{u(x_{j+1}) - u(x_j)}{h} \approx \frac{y_{j+1} - y_j}{h}.$$

Contudo, o problema de valor de contorno que estamos resolvendo depende de uma equação diferencial de segunda ordem. Entretanto,

$$u''(x) = \lim_{\Delta x \rightarrow 0} \frac{u'(x + \Delta x) - u'(x)}{\Delta x}$$

de modo que, repetindo o que já fizemos para a primeira derivada, podemos aproximar $u''(x)$ pelo quociente de Newton

$$\frac{u'(x_j + h) - u'(x_j)}{h} = \frac{u'(x_{j+1}) - u'(x_j)}{h}$$

Como não conhecemos $u'(x_j)$, nem $u'(x_{j+1})$, o melhor que podemos fazer é substituí-las pelas aproximações

$$u'(x_j) \approx \frac{y_{j+1} - y_j}{h} \quad \text{e} \quad u'(x_{j+1}) \approx \frac{y_{j+2} - y_{j+1}}{h}.$$

Fazendo isto, obtemos

$$u''(x) \approx \frac{1}{h} \left(\frac{y_{j+2} - y_{j+1}}{h} - \frac{y_{j+1} - y_j}{h} \right) = \frac{y_{j+2} - 2y_{j+1} + y_j}{h^2}.$$

Admitindo que esta fórmula nos dê uma aproximação muito boa para $u''(x)$, temos de (6) que

$$(8) \quad \frac{y_{j+2} - 2y_{j+1} + y_j}{h^2} = \frac{g}{T_0} \rho(x_j)$$

para $j = 0, \dots, n-2$. Note que paramos em $j = n-2$, porque y_{j+2} não faz sentido quando $j > n-2$. Mais importante é que, em um ato com todos os sinais de pura bravata, substituímos o sinal \approx , usado para indicar aproximação, por uma igualdade. Esta maneira de resolver problemas de valor de contorno é conhecida como *método das diferenças finitas*.

À primeira vista, não parecemos ter feito grande coisa: mesmo aceitando que esta aproximação para $u''(x)$ não é um caso claro de abusar da sorte, apenas reduzimos o problema aos valores de $y_j \approx u(x_j)$; mas não são exatamente os y_j que queremos calcular? Para mostrar porque seu ceticismo é equivocado, basta juntarmos as equações para $j = 0, \dots, n-2$. Tomando $j = 0$ em (8), obtemos

$$\frac{y_2 - 2y_1 + y_0}{h^2} = \frac{g}{T_0} \rho(x_0).$$

Lembrando que $y_0 = a$ e $x_0 = 0$, esta equação pode ser reescrita na forma

$$\frac{y_2 - 2y_1}{h^2} = \frac{g}{T_0} \rho(x_0) - \frac{a}{h^2},$$

cujo lado direito contém apenas valores conhecidos. Tomando, agora, $j = 1$,

$$\frac{y_3 - 2y_2 + y_1}{h^2} = \frac{g}{T_0}\rho(x_1) = \frac{g}{T_0}\rho(h).$$

Continuando assim, obtemos um sistema de equações lineares cuja última equação é

$$\frac{-2y_{n-1} + y_{n-2}}{h^2} = \frac{g}{T_0}\rho(x_n) - \frac{y_n}{h^2}.$$

Como $y_n = a$ e $x_n = \ell$,

$$\frac{-2y_{n-1} + y_{n-2}}{h^2} = \frac{g}{T_0}\rho(\ell) - \frac{a}{h^2}.$$

Por exemplo, supondo que $\ell = 1$, que $a = 2$ e que $\rho(x)$ é constante e igual a T_0/g . Escolhendo $n = 5$, teremos $h = 1/5$, de modo que o sistema será

$$\begin{aligned} (9) \quad & 25(y_2 - 2y_1) = -49 \\ & 25(y_3 - 2y_2 + y_1) = 1 \\ & 25(y_4 - 2y_3 + y_2) = 1 \\ & 25(-2y_4 + y_3) = -49. \end{aligned}$$

Resolvendo este sistema linear obtemos

$$y_1 = \frac{48}{25}, \quad y_2 = \frac{47}{25}, \quad y_3 = \frac{47}{25} \quad \text{e} \quad y_4 = \frac{48}{25},$$

que correspondem às seguintes aproximações de pontos sobre a curva solução,

$$\left(\frac{1}{5}, \frac{48}{25}\right), \left(\frac{2}{5}, \frac{47}{25}\right), \left(\frac{3}{5}, \frac{47}{25}\right), \left(\frac{4}{5}, \frac{48}{25}\right),$$

além, naturalmente, de $(0, 2)$ e $(1, 2)$. A figura 5 ilustra a curva poligonal obtida ligando entre si os pontos da solução aproximada, juntamente com a solução analítica exata, que calculamos na seção anterior.

Na solução do exemplo passamos diretamente do sistema (9) à sua solução, sem dar nenhuma indicação de como foi resolvido. Por sorte este sistema específico pode ser reduzido a um sistema com duas equações e duas incógnitas que pode ser facilmente resolvido. Para isto, basta usar a primeira equação para escrever y_1 em função de y_2 e a quarta equação para escrever y_4 em função de y_3 . Substituindo estas expressões, respectivamente, na segunda e terceira equações, obtemos o sistema 2×2

$$\begin{aligned} 25y_3 - \frac{75y_2}{2} &= -\frac{47}{2} \\ -\frac{75y_3}{2} + 25y_2 &= -\frac{47}{2}. \end{aligned}$$

Infelizmente, a aproximação que fizemos produziu uma curva poligonal que está muito longe da solução correta. Isto aconteceu porque dividimos o intervalo em

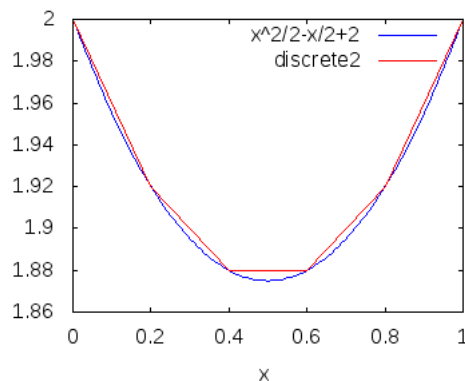


FIGURA 5. A solução exata e a aproximação poligonal da ponte com peso uniforme com $n = 5$.

poucas partes. Aumentando n de 5 para 20, obtemos um resultado muito melhor, como mostra a figura 6 onde, apesar das duas curvas terem sido desenhadas, mal conseguimos distinguir uma da outra.

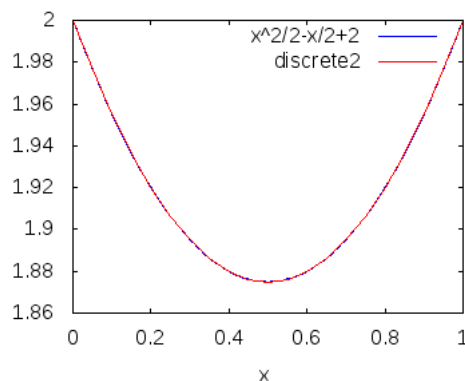


FIGURA 6. A solução exata e a aproximação poligonal da ponte com peso uniforme e $n = 20$.

Desta vez, o sistema tem 19 equações e 19 incógnitas e está fora de cogitação resolvê-lo através de um método ingênuo, como o utilizado quando $n = 5$. Na próxima seção descreveremos um método prático para resolver sistemas lineares, que funciona tão bem com lápis-e-papel, quanto quando usamos um computador. Antes, porém, faremos um exemplo em que a carga sobre a ponte não está uniformemente distribuída sobre o deque.

Suponhamos, como sugerido ao final da seção anterior, que a carga se acumula no centro da ponte de acordo com a função $y = \exp(-10x(x - 1))$, que é ilustrada

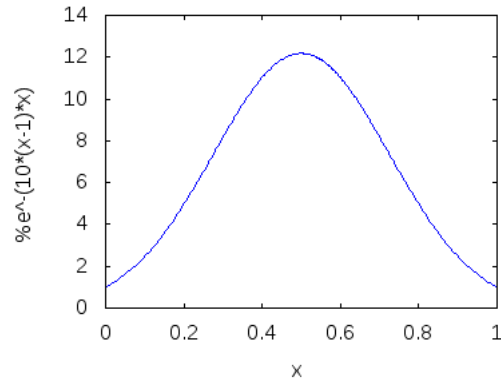


FIGURA 7. Distribuição normal de peso na ponte.

na figura 7 da página 12. Usando o método das diferenças finitas para resolver o problema de valor de contorno

$$u''(x) = \exp(-10x(x-1)), \quad u(0) = 2 \quad \text{e} \quad u(1) = 2,$$

com $n = 20$, obtemos um sistema com 19 equações e, portanto, grande demais para escrever aqui. Na figura 8 da página 8 desenhamos a curva poligonal, obtida a partir da solução do sistema linear, e a parábola que passa pelo vértice da poligonal e pelos pontos $(0, 2)$ e $(1, 2)$. Como você pode observar, mesmo tendo escolhido $n = 20$, a solução desta vez está muito longe de ser uma parábola — o que não é surpreendente, já que agora a massa está concentrada no centro da ponte.

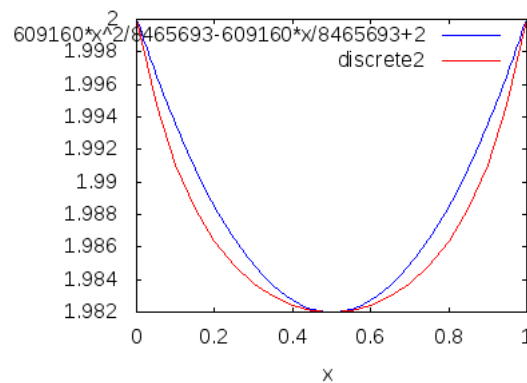


FIGURA 8. Curva descrita pelo cabo para a distribuição normal.

3. Resolução de sistemas lineares

Nos últimos anos do ensino fundamental aprendemos vários métodos para resolver sistemas lineares de duas variáveis. Um deles, é o *método de adição*, que consiste em multiplicar uma das equações por uma constante de modo que, quando as equações forem somadas, reste uma equação linear em apenas uma das variáveis. Por exemplo, se no sistema

$$(10) \quad \begin{aligned} x + 3y &= 1 \\ 2x + 5y &= 4, \end{aligned}$$

subtrairmos da segunda equação o dobro da primeira, obteremos $-y = 2$. Substituindo, então, $y = -2$ na primeira equação, encontramos

$$x = 1 - 3y = 1 - 3 \cdot (-2) = 7.$$

Portanto, a solução do (10) é $x = 7$ e $y = -2$. Podemos interpretar o que fizemos dizendo que transformamos o sistema (10) em

$$(11) \quad \begin{aligned} x + 3y &= 1 \\ -y &= 2, \end{aligned}$$

que é fácil de resolver porque a solução da segunda equação é totalmente óbvia.

Este procedimento pode ser estendido a sistemas de mais de duas equações. Por exemplo, considere o sistema

$$(12) \quad \begin{aligned} \mathbf{x} + \mathbf{3y} + z &= \mathbf{1} \\ \mathbf{2x} + \mathbf{5y} + 3z &= \mathbf{4} \\ 3x + 4y + 7z &= 5. \end{aligned}$$

cujas parcelas em negrito são iguais às do sistema 2×2 que consideramos anteriormente. Subtraindo da segunda equação o dobro da primeira, como fizemos acima, obtemos

$$\begin{aligned} x + 3y + z &= 1 \\ -y + z &= 2 \\ 3x + 4y + 7z &= 5; \end{aligned}$$

que, apesar de ainda ter três equações e três incógnitas, é um pouco mais simples que (12). Mas nada nos impede de utilizar um procedimento semelhante para simplificar a terceira equação. De fato, multiplicando a primeira equação por -3 e somando à terceira, encontramos

$$(13) \quad \begin{aligned} x + 3y + z &= 1 \\ -y + z &= 2 \\ -5y + 4z &= 2. \end{aligned}$$

Desta vez as duas últimas equações formam um sistema de duas equações em duas incógnitas, que podemos resolver utilizando o método de adição. Para isso, basta multiplicar a segunda equação de (13) por -5 e somá-la à segunda, que nos dá

$$(14) \quad \begin{aligned} x + 3y + z &= 1 \\ -y + z &= 2 \end{aligned}$$

$$(15) \quad -z = -8.$$

Mas segue da última equação de (14) que $z = 8$. Substituindo isto na segunda equação deste mesmo sistema, encontramos $y = 6$. Finalmente, substituindo $y = 6$ e $z = 8$ na primeira equação, obtemos $x = -25$, com o que resolvemos o sistema (12).

O método que utilizamos para resolver o sistema (12) é típico de um *procedimento recursivo*, que é o nome dado aos procedimentos que resolvem um dado problema reduzindo-o a uma instância mais simples do mesmo problema. No exemplo acima, ao eliminar as parcelas em x da segunda e terceira equações, transformamos

$$\begin{array}{ll} x + 3y + z = 1 & x + 3y + z = 1 \\ 2x + 5y + 3z = 4 & \text{em} \quad -y + z = 2 \\ 3x + 4y + 7z = 5 & -5y + 4z = 2. \end{array}$$

Mas, para resolver este último sistema, basta aplicar o mesmo procedimento a

$$\begin{aligned} -y + z &= 2 \\ -5y + 4z &= 2; \end{aligned}$$

que, como tem apenas duas equações e duas incógnitas, é um sistema menor que aquele com o qual começamos.

Nada nos impede de aplicar este mesmo procedimento a sistemas com mais equações e mais incógnitas, mas antes de fazer isto, convém simplificar um pouco a notação. A chave para isto é a observação de que o único papel que as incógnitas desempenham neste procedimento é o de marcadores de posição. Para tornar mais claro o que isto quer dizer, considere novamente o sistema dois por dois com o qual começamos

$$(16) \quad \begin{aligned} x + 3y &= 1 \\ 2x + 5y &= 4, \end{aligned}$$

A maneira mais comum de resolvê-lo é por substituição, e não por adição, e consiste em “tirar o valor de x da primeira equação e substituí-lo na segunda”. Em outras palavras, reescrevemos $x + 3y = 1$ na forma $x = 1 - 3y$ e aplicamos isto na segunda equação, obtendo

$$2(1 - 3y) + 5y = 4,$$

que tem uma única incógnita e pode ser facilmente resolvida. Neste método, uma das incógnitas é efetivamente escrita em função da outra, mas nada semelhante acontece

no método de adição *até que o sistema tenha sido completamente simplificado*; só então calculamos os valores das incógnitas, começando da última e acabando na primeira. De fato, para aplicar o método de adição, precisamos apenas saber quais coeficientes pertencem a quais incógnitas. A melhor maneira de lhe convencer disto é apagar as incógnitas e o sinal de igualdade em (16) e escrever os coeficientes em uma tabela, na qual suas posições relativas são mantidas, como abaixo

$$\begin{array}{ccc} 1 & 3 & 1 \\ 2 & 5 & 4 \end{array}$$

Para aplicar o método de adição basta multiplicar a primeira linha por -2 e somar o resultado à segunda linha, obtendo

$$\begin{array}{ccc} 1 & 3 & 1 \\ 0 & -1 & 2 \end{array}$$

Repondo as incógnitas em seus devidos lugares, obtemos o sistema (11), que pode ser facilmente resolvido. Tradicionalmente as tabelas acima são representadas como matrizes, mas a razão pela qual é preferível pensar nelas como matrizes e não simples tabelas só vai se tornar clara quando chegarmos ao capítulo 4.

Não posso lhe culpar se você estiver pensando que o parágrafo anterior é um caso típico de um matemático fazendo muito barulho por nada. A verdade é que esta simplificação só é perceptível quando o sistema tem muitas equações e muitas incógnitas. Para lhe convencer disto, faremos um outro exemplo. Considere o seguinte sistema de cinco equações e cinco incógnitas

$$(17) \quad x_1 + x_2 - x_3 + x_4 + 2x_5 = 1$$

$$-2x_1 - 3x_2 + x_3 - x_4 - 2x_5 = 0$$

$$2x_1 + 3x_2 + 2x_4 + 4x_5 = 7$$

$$(18) \quad -3x_1 - 6x_2 + 2x_3 + x_4 + 6x_5 = 16$$

$$-x_1 - 3x_2 + x_3 + x_4 + 11x_5 = 20.$$

Onde é mais fácil de identificar os coeficientes: no sistema acima, ou na matriz abaixo?

$$(19) \quad \begin{bmatrix} 1 & 1 & -1 & 1 & 2 & 1 \\ -2 & -3 & 1 & -1 & -2 & 0 \\ 2 & 3 & 0 & 2 & 4 & 7 \\ -3 & -6 & 2 & 1 & 6 & 16 \\ -1 & -3 & 1 & 1 & 11 & 20 \end{bmatrix}.$$

Já que temos a matriz, vamos aproveitar e simplificá-la usando o processo de eliminação introduzido na solução do sistema (12). Como cada linha da matriz corresponde a uma equação do sistema, multiplicar toda uma linha por uma constante

e somá-la a outra equivale a fazer esta mesma operação sobre as equações correspondentes. Começamos multiplicando a primeira linha da matriz 2 e somando o resultado à segunda linha, o que nos dá

$$\begin{bmatrix} 1 & 1 & -1 & 1 & 2 & 1 \\ 0 & -1 & -1 & 1 & 2 & 2 \\ 2 & 3 & 0 & 2 & 4 & 7 \\ -3 & -6 & 2 & 1 & 6 & 16 \\ -1 & -3 & 1 & 1 & 11 & 20 \end{bmatrix}.$$

Multiplicando, agora, a primeira linha por -2 e somando o resultado à terceira linha, resta

$$\begin{bmatrix} 1 & 1 & -1 & 1 & 2 & 1 \\ 0 & -1 & -1 & 1 & 2 & 2 \\ 0 & 1 & 2 & 0 & 0 & 5 \\ -3 & -6 & 2 & 1 & 6 & 16 \\ -1 & -3 & 1 & 1 & 11 & 20 \end{bmatrix}.$$

Procedendo de maneira semelhante para as duas últimas linhas, obtemos

$$\begin{bmatrix} 1 & 1 & -1 & 1 & 2 & 1 \\ 0 & -1 & -1 & 1 & 2 & 2 \\ 0 & 1 & 2 & 0 & 0 & 5 \\ 0 & -3 & -1 & 4 & 12 & 19 \\ 0 & -2 & 0 & 2 & 13 & 21 \end{bmatrix}.$$

Com isto eliminamos todas as entradas da primeira coluna da matriz, exceto a primeira. O processo pode, então, ser repetido para a submatriz

$$\begin{bmatrix} -1 & -1 & 1 & 2 & 2 \\ 1 & 2 & 0 & 0 & 5 \\ -3 & -1 & 4 & 12 & 19 \\ -2 & 0 & 2 & 13 & 21 \end{bmatrix}.$$

Na prática, não vale à pena destacar a submatriz, basta aplicar o processo de eliminação à segunda coluna de

$$\begin{bmatrix} 1 & 1 & -1 & 1 & 2 & 1 \\ 0 & -1 & -1 & 1 & 2 & 2 \\ 0 & 1 & 2 & 0 & 0 & 5 \\ 0 & -3 & -1 & 4 & 12 & 19 \\ 0 & -2 & 0 & 2 & 13 & 21 \end{bmatrix},$$

usando a primeira posição não nula da segunda linha (em negrito) para eliminar as posições da segunda coluna que ficam abaixo dela. Fazendo isto, obtemos

$$\begin{bmatrix} 1 & 1 & -1 & 1 & 2 & 1 \\ 0 & -1 & -1 & 1 & 2 & 2 \\ 0 & 0 & 1 & 1 & 2 & 7 \\ 0 & 0 & 2 & 1 & 6 & 13 \\ 0 & 0 & 2 & 0 & 9 & 17 \end{bmatrix},$$

que tem todas as entradas da segunda coluna nulas, exceto as duas primeiras. Note que esta segunda rodada da eliminação não afeta a primeira coluna, porque todas as entradas desta coluna já são nulas, exceto a primeira. Em seguida, eliminamos todas as posições na terceira coluna de

$$\begin{bmatrix} 1 & 1 & -1 & 1 & 2 & 1 \\ 0 & -1 & -1 & 1 & 2 & 2 \\ 0 & 0 & \mathbf{1} & 1 & 2 & 7 \\ 0 & 0 & 2 & 1 & 6 & 13 \\ 0 & 0 & 2 & 0 & 9 & 17 \end{bmatrix}$$

que ficam abaixo da diagonal. Com isto, obtemos

$$\begin{bmatrix} 1 & 1 & -1 & 1 & 2 & 1 \\ 0 & -1 & -1 & 1 & 2 & 2 \\ 0 & 0 & 1 & 1 & 2 & 7 \\ 0 & 0 & 0 & -\mathbf{1} & 2 & -1 \\ 0 & 0 & 0 & -2 & 5 & 3 \end{bmatrix}.$$

Finalmente, a posição abaixo da diagonal na quarta coluna é eliminada, resultando a matriz

$$\begin{bmatrix} 1 & 1 & -1 & 1 & 2 & 1 \\ 0 & -1 & -1 & 1 & 2 & 2 \\ 0 & 0 & 1 & 1 & 2 & 7 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 1 & 5 \end{bmatrix}.$$

Resta-nos apenas construir o sistema correspondente a esta última matriz e resolvê-lo. Tomando cuidado em usar o mesmo posicionamento das variáveis utilizado quando convertimos o sistema original na matriz (19), obtemos

$$\begin{aligned} x_1 + x_2 - x_3 + x_4 + 2x_5 &= 1 \\ -x_2 - x_3 + x_4 + 2x_5 &= 2 \\ x_3 + x_4 + 2x_5 &= 7 \\ -x_4 + 2x_5 &= -1 \\ x_5 &= 5, \end{aligned}$$

que pode ser facilmente resolvido, começando da última equação para a primeira, o que nos dá

$$x_5 = -5, \quad x_4 = -11, \quad x_3 = 14, \quad x_2 = -33, \quad x_1 = 67$$

Na próxima seção analisaremos em detalhe o procedimento que utilizamos ao resolver os sistemas acima, a fim de identificar suas principais etapas e explicar porque funciona.

4. Eliminação e substituição

Começaremos determinando as etapas que foram executadas para resolver os sistemas da seção anterior. A primeira coisa que fizemos foi introduzir a *matriz aumentada* do sistema, que é o nome dado à matriz cujas linhas contêm os coeficientes das variáveis de um dada equação do sistema. Lembre-se que os coeficientes de uma mesma variável devem todos aparecer na mesma coluna. Em seguida simplificamos a matriz, eliminando, em cada coluna, todas as entradas abaixo da diagonal. Ao final da simplificação, obtivemos uma matriz conhecida como *triangular superior*, porque todas as suas entradas abaixo da diagonal são nulas. Em seguida, construímos o sistema associado à matriz triangular superior, utilizando as mesmas convenções usadas para obter a matriz aumentada do sistema original. Finalmente, resolvemos o sistema associado à matriz triangular superior. Portanto, descontando as traduções entre sistemas e matrizes, que não tem nenhum conteúdo matemático, o método que usamos para resolver os sistemas da seção 3 consiste de duas etapas:

Etapa 1: *eliminação* sistemática das posições abaixo da diagonal, de modo a obter uma matriz triangular superior;

Etapa 2: solução do sistema correspondente à matriz triangular superior por *substituição reversa* (back substitution).

Estas duas etapas precisam ser analisadas em mais detalhe. Seja

$$(20) \quad \begin{array}{rcl} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n & = & b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,n}x_n & = & b_2 \\ & \vdots & \\ a_{n,1}x_1 + a_{n,2}x_2 + \cdots + a_{n,n}x_n & = & b_n \end{array}$$

um sistema *determinado*, cujo número de equações coincide com o número de incógnitas. Como vimos nos exemplos, a eliminação na **Etapa 1** é realizada usando

operações extremamente simples. A matriz aumentada do sistema (20) é

$$(21) \quad \left[\begin{array}{cccc|c} a_{1,1} & a_{1,2} & \cdots & a_{1,n} & b_1 \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} & b_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} & b_n \end{array} \right].$$

Supondo que $a_{1,1} \neq 0$ e tomando

$$(22) \quad c_j = -\frac{a_{j,1}}{a_{1,1}},$$

anulamos a primeira entrada da j -ésima linha usando a *operação elementar* entre esta e a primeira linha que consiste em substituir a entrada $a_{j,i}$ por $a_{j,i} + c_j a_{1,i}$, ao longo de toda a linha. A matriz resultante desta operação é

$$(23) \quad \left[\begin{array}{cccc|c} a_{1,1} & a_{1,2} & \cdots & a_{1,n} & b_1 \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} & b_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a_{j,2} + c_j a_{1,2} & \cdots & a_{j,n} + c_j a_{1,n} & b_j + c_j b_j \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} & b_n \end{array} \right].$$

Fazendo isto para todas as linhas abaixo da primeira, obtemos

$$\left[\begin{array}{cccc|c} a_{1,1} & a_{1,2} & \cdots & a_{1,n} & b_1 \\ 0 & a_{2,2} + c_2 a_{1,2} & \cdots & a_{2,n} + c_2 a_{1,n} & b_2 + c_2 b_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a_{j,2} + c_j a_{1,2} & \cdots & a_{j,n} + c_j a_{1,n} & b_j + c_j b_j \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a_{n,2} + c_n a_{1,2} & \cdots & a_{n,n} + c_n a_{1,n} & b_n + c_n b_n \end{array} \right],$$

pois

$$a_{j,1} + c_j a_{1,1} = a_{j,1} + -\frac{a_{j,1}}{a_{1,1}} a_{1,1} = 0.$$

A entrada $a_{1,1}$ é o *pivô* deste passo da eliminação e não pode ser nula, porque, se fosse, não poderíamos calcular c_j usando (22). Infelizmente não é possível garantir que as matrizes aumentadas de todos os sistemas que analisaremos tenham a entrada 1, 1 diferente de zero. Contudo, a primeira coluna da matriz (21) não pode ser toda nula porque, se isto acontecesse, teríamos um sistema com mais equações que incógnitas, o que contraria nossa suposição inicial. Portanto, tem que haver uma linha em (21) cuja primeira entrada é diferente de zero. Assim, podemos contornar o pivô nulo simplesmente trocando a primeira linha da matriz aumentada por outra linha cuja primeira posição é diferente de zero. Note que isto pode ser feito impunemente,

porque as linhas da matriz aumentada apenas nos dão uma maneira descomplicar as equações do sistema, e trocar equações de posição não altera o sistema.

Como nos exemplos, continuamos o processo de eliminação aplicando o mesmo procedimento à submatriz

$$\left[\begin{array}{ccc|c} a_{2,2} + c_2a_{1,2} & \cdots & a_{2,n} + c_2a_{1,n} & b_2 + c_2b_1 \\ \vdots & \vdots & \vdots & \vdots \\ a_{j,2} + c_ja_{1,2} & \cdots & a_{j,n} + c_ja_{1,n} & b_j + c_jb_1 \\ \vdots & \vdots & \vdots & \vdots \\ a_{n,2} + c_na_{1,2} & \cdots & a_{n,n} + c_na_{1,n} & b_n + c_nb_1 \end{array} \right],$$

e usando $a_{2,2} + c_2a_{1,2}$ como pivô. Caso $a_{2,2} + c_2a_{1,2} = 0$, trocamos a primeira linha desta submatriz com outra linha cuja primeira posição não seja nula, exatamente como fizemos quando $a_{1,1} = 0$. O procedimento continua desta maneira, até que tenhamos obtido uma matriz triangular superior

$$(24) \quad \left[\begin{array}{cccc|c} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,n} & \beta_1 \\ 0 & \alpha_{2,2} + c_2\alpha_{1,2} & \cdots & \alpha_{2,n} + c_2\alpha_{1,n} & \beta_2 + c_2\beta_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \alpha_{n,n} + c_n\alpha_{1,n} & \beta_n + c_n\beta_1 \end{array} \right],$$

com o quê a **Etapa 1** se encerra. Note que $\alpha_{1,i} = a_{1,i}$, para todo $1 \leq i \leq n$ e $\beta_1 = b_1$, mas, dependendo da matriz, os demais coeficientes podem ser todos diferentes dos seus correspondentes em (21).

A **Etapa 2** é, então, aplicada ao sistema

$$(25) \quad \begin{aligned} \alpha_{1,1}x_1 + \alpha_{1,2}x_2 + \cdots + \alpha_{1,n-1}x_{n-1} + \alpha_{1,n}x_n &= \beta_1 \\ \alpha_{2,2}x_2 + \cdots + \alpha_{1,n-1}x_{n-1} + \alpha_{2,n}x_n &= \beta_2 \\ &\vdots \\ \alpha_{n-1,n-1}x_{n-1} + \alpha_{n,n}x_n &= \beta_n \\ \alpha_{n,n}x_n &= \beta_n \end{aligned}$$

Note que, como estamos supondo que o sistema (20) é determinado, nenhuma das entradas $\alpha_{1,1}, \alpha_{2,2}, \dots, \alpha_{n,n}$ pode ser nula. Logo, a última equação de (25) nos dá

$$x_n = \frac{\beta_n}{\alpha_{n,n}}.$$

Substituindo isto nas demais equações de (25), obtemos o sistema

$$\begin{aligned}\alpha_{1,1}x_1 + \alpha_{1,2}x_2 + \cdots + \alpha_{1,n-1}x_{n-1} &= \beta_1 - \alpha_{1,n}\beta_n/\alpha_{n,n} \\ \alpha_{2,2}x_2 + \cdots + \alpha_{1,n-1}x_{n-1} &= \beta_2 - \alpha_{2,n}\beta_n/\alpha_{n,n} \\ &\vdots \\ \alpha_{n-1,n-1}x_{n-1} &= \beta_{n-1} - \alpha_{n-1,n}\beta_n/\alpha_{n,n},\end{aligned}$$

que tem uma variável e uma equação a menos que (25) e cuja última equação tem apenas uma variável. Fazendo o mesmo com este novo sistema, e assim sucessivamente, obteremos, por fim, os valores de todas as incógnitas.

Note que os procedimentos usados para resolver as duas etapas são recursivos; em outras palavras, consistem em efetuar operações que reduzem um problema $n \times n$ a um problema análogo de tamanho $(n-1) \times (n-1)$ que, por sua vez é reduzido a um problema $(n-2) \times (n-2)$, e assim por diante.

Tendo analisado as várias etapas do método que utilizamos para resolver sistemas lineares, somos confrontados com uma dura realidade. Embora nosso objetivo original fosse o de resolver (20), o sistema que efetivamente resolvemos foi (25). Mas *isto só nos dará uma solução do sistema original se os dois sistemas tiverem a mesma solução*. Por sorte, um sistema é transformado por uma sucessão de operações elementares por linhas que afetam apenas duas linhas de cada vez. Levando em conta a relação entre (20) e sua matriz aumentada (21), isto equivale a dizer que estamos transformando o sistema uma equação de cada vez. Portanto, basta mostrar que os sistemas correspondentes às matrizes aumentadas antes e depois da aplicação de *uma* operação elementar por linha têm a mesma solução. Por exemplo, se (20) tiver como solução

$$(26) \quad x_1 = s_1, \quad x_2 = s_2, \dots, x_n = s_n,$$

então

$$a_{1,1}s_1 + a_{1,2}s_2 + \cdots + a_{1,n}s_n = b_1 \quad \text{e} \quad a_{j,1}s_1 + a_{j,2}s_2 + \cdots + a_{j,n}s_n = b_j.$$

Multiplicando os dois lados da primeira igualdade por c_j e somando o resultado à segunda, obtemos

$$(a_{j,1} + c_j a_{1,1})s_1 + (a_{j,2} + c_j a_{1,2})s_2 + \cdots + (a_{j,n} + c_j a_{1,n})s_n = b_j + c_j a_{1,n}s_n,$$

o que mostra que (26) também é solução da equação correspondente à j -ésima linha da matriz (23). Como isto pode ser facilmente generalizado para os sistemas antes e depois de qualquer passo da eliminação, podemos concluir que (20) e (25) de fato têm a mesma solução.

Vamos encerrar com um exemplo em que ocorre o anulamento de um pivô. Considere o sistema

$$(27) \quad \begin{aligned} x_1 + x_2 - x_3 + x_4 &= 2 \\ -2x_1 - 2x_2 + 3x_3 &= 2 \\ -x_1 - 2x_2 + 3x_3 + x_4 &= 6 \\ 2x_1 + 4x_2 - 3x_3 + 2x_4 &= 16. \end{aligned}$$

cujas matriz aumentada é

$$\left[\begin{array}{cccc|c} 1 & 1 & -1 & 1 & 2 \\ -2 & -2 & 3 & 0 & 2 \\ -1 & -2 & 3 & 1 & 6 \\ 2 & 4 & -3 & 2 & 16 \end{array} \right].$$

Usando a entrada 1,1 como pivô, anulamos as entradas que ficam abaixo dela, obtendo

$$\left[\begin{array}{ccccc} 1 & 1 & -1 & 1 & 2 \\ 0 & 0 & 1 & 2 & 6 \\ 0 & -1 & 2 & 2 & 8 \\ 0 & 2 & -1 & 0 & 12 \end{array} \right].$$

Nosso próximo pivô deveria estar em 2,2. Entretanto, esta posição é nula, de modo que não podemos utilizá-la na eliminação. Neste exemplo, poderíamos trocar a segunda linha desta última matriz com a terceira ou quarta linhas. Digamos que a troca seja feita com a terceira linha, o que nos dá

$$\left[\begin{array}{ccccc} 1 & 1 & -1 & 1 & 2 \\ 0 & -1 & 2 & 2 & 8 \\ 0 & 0 & 1 & 2 & 6 \\ 0 & 2 & -1 & 0 & 12 \end{array} \right].$$

Continuando o processo de eliminação a partir desta matriz, obtemos ao final a matriz triangular superior

$$\left[\begin{array}{ccccc} 1 & 1 & -1 & 1 & 2 \\ 0 & 0 & 1 & 2 & 6 \\ 0 & -1 & 1 & 0 & 2 \\ 0 & 0 & 0 & -2 & 10 \end{array} \right],$$

que corresponde ao sistema

$$\begin{aligned} x_1 + x_2 - x_3 + x_4 &= 2 \\ x_3 + 2x_4 &= 6 \\ -x_2 + x_3 &= 2 \\ -2x_4 &= 10 \end{aligned}$$

cuja solução é dada por

$$x_4 = -5, x_3 = 16, x_2 = 14, x_1 = 9.$$

Como já sabemos que este último sistema tem as mesmas soluções que (27), resolvemos o sistema desejado.

5. Recapitulando e olhando adiante

É hora de recapitular o que fizemos e estabelecer metas para os próximos quatro capítulos. Nosso objetivo neste capítulo foi estudar a curva descrita pelo cabo de sustentação de uma ponte pênsil. Vimos que, na ausência de vento e supondo que a ponte fica suspensa em um único cabo, muito mais leve que o deque da ponte, a função $u(x)$ cujo gráfico descreve a forma do cabo é solução do problema de valor de contorno

$$u''(x) = \frac{g}{T_0} \rho(x) \quad \text{e} \quad u(0) = u(\ell) = a,$$

em que g é a aceleração da gravidade, T_0 é o componente horizontal (constante) da tensão no cabo, ℓ é o comprimento do deque, a é a altura das torres que sustentam o cabo e $\rho(x)$ é a massa por unidade de comprimento do deque, incluindo os veículos que estão sobre ele.

Quando esta equação não tem solução analítica, precisamos usar métodos numéricos para determinar a forma do cabo. Para isto, introduzimos, na seção 2 o método das diferenças finitas que nos permite aproximar a segunda derivada de $u(x)$, em um dado ponto, por uma fórmula que envolve apenas as aproximações dos valores de $u(x)$ para uma quantidade finita de valores de x .

Mais precisamente, escolhemos um inteiro positivo n , e dividimos o deque da ponte em n partes iguais, cada uma de comprimento $h = \ell/n$. Como estamos posicionando a origem dos eixos no ponto de interseção do deque com a torre esquerda, como na figura 2, os segmentos em que o deque fica dividido são da forma $[x_j, x_{j+1}]$, para $j = 0, \dots, n-1$. Usando y_j para denotar a aproximação de $u(x_j)$ que pretendemos calcular, argumentamos que é possível tomar

$$\frac{y_{j+2} - 2y_{j+1} + y_j}{h^2}$$

como aproximação de $u''(x_j)$. Substituindo estas aproximações na equação diferencial, e escolhendo as unidades de medida de maneira que $g/T_0 = 1$, obtemos

$$(28) \quad \frac{y_{j+2} - 2y_{j+1} + y_j}{h^2} = \rho(x_j) \quad \text{para} \quad j = 0, \dots, n-2.$$

Levando em conta que

$$y_0 = y(0) = a = y(\ell) = y_n,$$

as equações em (28) nos dão um sistema de $n - 1$ equações nas $n - 1$ incógnitas y_1, \dots, y_{n-1} , cuja matriz aumentada tem a forma

$$(29) \quad \frac{1}{h^2} \left[\begin{array}{cccccc|c} -2 & 1 & 0 & \cdots & 0 & 0 & \rho(x_0)h^2 - y_0 \\ 1 & -2 & 1 & \cdots & 0 & 0 & \rho(x_1)h^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -2 & 1 & \rho(x_{n-2})h^2 - y_n \end{array} \right].$$

Note que as únicas entradas não nulas desta matriz estão situadas em sua diagonal principal e nas duas subdiagonais, imediatamente acima e abaixo da diagonal principal. Em seguida, aplicamos operações elementares por linha à matriz (29) até transformá-la em uma matriz triangular superior. O sistema correspondente a esta última matriz pode, então, ser resolvido usando substituição reversa. A propósito, não é difícil ver que o sistema correspondente a (29) é sempre determinado, de modo que tudo o que aprendemos na seção 4 pode ser aplicado a ele. Tendo resolvido o sistema (28) obtemos números y_1, \dots, y_{n-1} , tais que os pontos

$$(x_1, y_1), \dots, (x_{n-1}, y_{n-1})$$

estão aproximadamente sobre a curva $y = u(x)$. Ligando os pontos consecutivos desta lista, juntamente com (x_0, y_0) e (x_n, y_n) , obtemos uma aproximação da curva $y = u(x)$.

O que fizemos até este ponto levanta tantas questões quanto as que responde. Para começar (1) quão boas são as aproximações que obtivemos desta maneira? Naturalmente gostaríamos que, quanto maior fosse o número de partes em que $[0, \ell]$ for dividido, tanto melhor seja a aproximação calculada para a solução. Mas isto é apenas um desejo: (2) como ter certeza de que realmente se verifica? Uma pergunta relacionada de perto às duas primeiras diz respeito ao procedimento de eliminação utilizado para simplificar a matriz aumentada. Para executá-lo, tivemos que fazer uma grande quantidade de cálculos com números que, em problemas de física, química ou engenharia, serão necessariamente aproximados: (3) como ter certeza de que estes cálculos não amplificam os erros inerentes a estes valores aproximados, a ponto de tornar inúteis as soluções do sistema linear? O mesmo problema pode ocorrer em casos em que não há números oriundos de medições de problemas físicos. Por exemplo, para resolver o problema de valor de contorno

$$u''(x) = \exp(10x(x-1)), \quad u(0) = 2 \quad \text{e} \quad u(1) = 2,$$

ao final da seção 2, tivemos que calcular $\exp(10x(x-1))$ para vários valores de x , mas (4) como controlar o erro cometido nestes cálculos? Mesmo tendo uma calculadora, como isto deveria ser feito? Finalmente, há a questão da *validação* do nosso modelo; isto é, (5) quão próxima da curva descrita pelo cabo de uma ponte de verdade está aquela calculada a partir de nosso modelo matemático?

As questões listadas acima podem ser resumidas nas seguintes perguntas, que associamos aos capítulos onde serão respondidas:

Capítulo 2: como estimar o erro cometido em cálculos aritméticos e no cálculo de valores aproximados de funções?

Capítulo 3: como aproximar $u''(x)$ cometendo um erro pequeno que não invalida o modelo que criamos?

Capítulo 4: como estimar o erro inerente aos cálculos executados na solução de um sistema linear por eliminação e substituição reversa?

Capítulo 5: como comparar o modelo matemático, codificado no problema de valor de contorno, a dados obtidos experimentalmente?

CAPÍTULO 2

Aproximação de números e funções

Neste capítulo veremos como aproximar números e funções. Começaremos tratando de números, onde o problema é mais familiar e mais fácil de resolver.

1. Representação de números

Ao longo da história da humanidade, diferentes civilizações adotaram diferentes maneiras de representar números. No ensino fundamental aprendemos a interpretar algarismos romanos; estes algarismos, na verdade, são derivados de um sistema de abreviações utilizados algumas vezes pelos gregos da época clássica. Neste sistema um número é representado pela primeira letra (maiúscula) do seu nome em grego. Por exemplo, $\Pi = 5$, porque cinco em grego é $\Pi\epsilon\nu\tau\epsilon$, de onde deriva nosso prefixo *penta*, e $\Delta = 10$, do grego $\Delta\epsilon\kappa\alpha$, que nos deu o prefixo *deca*.

A partir de 450 a.C. surgiu na Grécia uma segunda maneira de representar números, que consistia em usar as letras na ordem em que aparecem no alfabeto grego. Assim as nove primeiras letras vão de 1 a 9, as nove seguintes de 10 a 90 e as nove finais de 100 a até 900. Para que isto seja possível são necessários 27 símbolos; entretanto, o alfabeto grego clássico só tem 24 letras! Os gregos supriam os 3 símbolos extra usando letras que, na época clássica, haviam caído em desuso; por exemplo, o digama \varGamma era usado para representar o número seis.

Escrever um número menor que um milhão no sistema alfabético usado pelos gregos era relativamente fácil. Como $\Delta = 4$, $N = 50$ e $\Psi = 700$, o número 754 seria escrito como $\Psi N \Delta$. Mas a coisa se complicava quando o número a ser representado era muito grande. Em um de seus livros, conhecido pelo nome latino de *Arenarius*, Arquimedes se propõe a estimar a quantidade máxima de areia que caberia no universo. Para tornar isto possível, Arquimedes começa por inventar um sistema que lhe permita representar os números realmente enormes com que iria trabalhar.

Surpreendentemente, um sistema que permitia representar, sem maior dificuldade, números ainda maiores do que os que surgiram no *Arenarius* já havia sido inventado pelos babilônios mais de mil anos antes de Arquimedes. Trata-se do sistema posicional de base 60, que é atestado em inscrições datadas de 2000 a.C. Neste sistema há

60 símbolos, que representam os números de 1 a 60 e que funcionam como os dez algarismos de nosso sistema decimal. Por exemplo, $\nabla = 1$, ao passo que $\ll = 20$; portanto,

$$\nabla \ll = 1 \cdot 60 + 20 = 80.$$

Na verdade, como os babilônios não tinham nem zero, nem a vírgula (ou ponto), o símbolo acima podia denotar igualmente $60^2 + 20 \cdot 60$ e $1 + 20/60$. Já $60^2 + 20$ seria denotado deixando um espaço entre os símbolos para 1 e 60

$$\nabla \quad \ll$$

Como isso poderia facilmente criar confusão, um símbolo (semelhante a dois ∇ inclinados de 45° no sentido anti-horário) foi inventado para indicar uma casa vazia. Entretanto, ao contrário do nosso zero, esse símbolo não era considerado um número. O sistema de base 60 dos babilônios nunca se extinguiu completamente e, apesar dos esforços de decimalização que sucederam a Revolução Francesa, é utilizado até hoje em nossas medidas de ângulos em termos de grau.

Não custa lembrar que a maneira como estamos escrevendo números usando potências de 60 é totalmente anacrônica e nunca foi usada pelos babilônios, que pensavam em termos de 60 vezes maior, tal qual fizemos todos nós quando aprendemos a escrever números em notação decimal no ensino fundamental. O mesmo se aplica à versão original do sistema decimal e ao sistema maia, que descrevemos mais adiante.

O sistema posicional de base 10 que utilizamos originou-se na Índia. A recente datação do manuscrito Bakhshali, da Bodleian Library em Oxford, como tendo sido escrito entre o terceiro e quarto séculos d.C. mostrou que já naquela época uma versão primitiva do zero (um círculo cheio) era usada para indicar uma posição vazia. O sistema foi também adotado pelos matemáticos árabes. Tanto Al-Khwarizmi quanto Al-Kindi publicaram livros em que descreviam a utilização do sistema de cálculo com números decimais. Foi através das obras destes e de outros matemáticos árabes que este sistema chegou à Europa. Tão grande foi a influência das obras de Al-Khwarizmi no ocidente que, tanto *algarismo* quanto *algoritmo*, são derivadas de seu nome. Atestado na Europa a partir do ano 976, o sistema decimal foi promovido no *Liber Abaci* (Livro do Cálculo), publicado em 1202 pelo matemático italiano Fibonacci, mais conhecido hoje em dia pelos números que levam seu nome. Contudo, sua difusão no ocidente só ocorreu, verdadeiramente, a partir da invenção da imprensa.

A imprensa também foi a responsável pelo surgimento da vírgula (ou ponto) para separar as casas decimais. Os matemáticos indianos usavam uma barra sobre o algarismo das unidades, de modo que $2\overline{9}76$ correspondia a 29.76. Esta notação ainda era usada à época de Al-Khwarizmi, mas foi depois substituída por uma barra

vertical entre as unidades e os décimos; quando os livros começaram a ser impressos, a barra vertical passou a ser representada pela vírgula ou ponto atuais.

Generalizando o que vimos até agora, verificamos que, para que um sistema de representação de números seja *posicional de base β* é necessário que β seja um inteiro maior que 1 e que haja um conjunto \mathcal{S} formado por β símbolos, cada um dos quais representa um número de 0 a $\beta-1$. Por exemplo, $\mathcal{S} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, quando $\beta = 10$.

Digamos que r seja um número real. Uma vez que β e \mathcal{S} tenham sido fixados, existem um inteiro n e símbolos $\alpha_i \in \mathcal{S}$, um para cada $i < n$, tais que

$$(r)_\beta = \alpha_n \alpha_{n-1} \cdots \alpha_0 . \alpha_{-1} \alpha_{-2} \cdots .$$

Identificando α_i com o número (decimal) entre 0 e $\beta-1$ que ele representa, a expressão acima corresponde ao número decimal

$$(30) \quad r = \alpha_n \beta^n + \cdots + \alpha_0 + \alpha_{-1} \beta^{-1} + \alpha_{-2} \beta^{-2} + \cdots .$$

Quando $\beta = 10$, cada α s é um algarismo de 0 a 9.

Um exemplo menos óbvio é o sistema posicional de base $\beta = 20$, usado pelos maias. O zero, que eles inventaram independentemente dos indianos, era denotado pelo símbolo $|\textcircled{\cdot}|$. Os números de um a quatro eram representados por pontos dispostos horizontalmente; cinco, dez e quinze correspondiam a uma, duas ou três barras horizontais sobrepostas. Para obter os números entre 5 e 10, 10 e 15 e 15 e 20, os maias acrescentavam de um a quatro pontos sobre a barra superior. Assim, o número 17 seria representado na forma $|\textcircled{\cdot\cdot\cdot}|$. Resumindo temos que, no caso da numeração maia, $\beta = 20$ e

$$\mathcal{S} = \left\{ \underbrace{|\textcircled{\cdot}|}_0, \underbrace{|\cdot|}_1, \dots, \underbrace{|\equiv|}_{10}, \underbrace{|\textcircled{\cdot\cdot}|}_{11}, \dots, \underbrace{|\textcircled{\cdot\cdot\cdot}|}_{19} \right\} .$$

Isto basta para que, seguindo a receita dada em (30), possamos facilmente determinar

a representação decimal de qualquer número maia. Digamos que o número seja $\left| \begin{array}{c} \cdot \\ \textcircled{\cdot} \\ \dots \end{array} \right|$.

A descrição de \mathcal{S} feita acima, nos permite escrever

$$|\dots| = 3, \quad |\textcircled{\cdot}| = 0 \quad \text{e} \quad |\cdot| = 1.$$

Portanto, como $\beta = 20$, deduzimos de (30) que

$$\left| \begin{array}{c} \cdot \\ \textcircled{\cdot} \\ \dots \end{array} \right| = 1 \cdot 20^2 + 0 \cdot 20^1 + 3 \cdot 20^0.$$

Você tem toda razão, se estiver pensando que, apesar de serem posicionais, os sistemas de numeração dos babilônios e dos maias têm apenas interesse histórico. Apesar disto, há dois sistemas de numeração posicionais, além do decimal, que são extremamente importantes neste início de século XXI: os sistemas binário e hexadecimal. A importância desses sistemas decorre de seu uso em computação. No sistema *binário*

$$\beta = 2 \quad \text{e} \quad \mathcal{S} = \{0, 1\},$$

já no sistema *hexadecimal*

$$\beta = 16 \quad \text{e} \quad \mathcal{S} = \{0, 1, \dots, 9, A, B, C, D, E, F\},$$

em que as letras de A a F representam os números de 10 (= A) a 15 (= F). Por exemplo,

$$(403)_2 = 110010011 \quad \text{e} \quad (403)_{16} = 193.$$

Como vimos na definição geral, dada acima, os sistemas posicionais nos permitem representar quaisquer números reais; sempre, claro, com a ressalva que a expansão em base β de um dado número real pode ser infinita. Usando a fórmula (30) é fácil verificar que

$$(5/7)_2 = 0.101101\dots \quad \text{e} \quad (5/7)_{16} = 0.B6DB6D\dots$$

Veremos na seção 3 que, em um computador digital, números são representados em forma binária. Já o sistema hexadecimal é usado, entre outras coisas, na codificação de cores em html; por exemplo #FFFFFF corresponde a cor de número

$$15 + 15 \cdot 16 + 15 \cdot 16^2 + 15 \cdot 16^3 + 15 \cdot 16^4 + 15 \cdot 16^5 = 16777215,$$

que é o branco.

2. Decimais infinitas e erros

Um problema que surge ao usarmos um sistema de notação posicional para representar números é que quase todos requerem uma quantidade infinita de casas decimais. No caso das frações, isto pode ser facilmente contornado, porque a parte decimal é constituída por um padrão finito de números que se repete infinitamente. Mas, o que fazer no caso de números irracionais, como $\sqrt{2}$ ou π ? Neste caso não pode haver nenhum padrão; se houvesse, estes números seriam racionais. Mas mesmo Euclides, que viveu por volta em 300 a.C, já sabia que $\sqrt{2}$ não é um número racional. A única saída viável é usar uma aproximação.

Um número irracional cujas aproximações são muito bem documentadas no registro histórico é π . Isto não é surpreendente, dada a necessidade de medir o volume das vasilhas cilíndricas onde eram conservados grãos. O papiro Rhind, escrito no antigo Egito por volta de 1650 a.C. contém a seguinte receita: *subtraia do diâmetro*

do círculo sua nona parte, o quadrado deste resultado é a área do círculo. Denotando o raio do círculo por r , esta receita equivale à fórmula

$$\left(2r - \frac{2r}{9}\right)^2 = \frac{256}{81} r^2;$$

donde podemos deduzir que os antigos egípcios tomavam π como sendo aproximadamente igual a

$$\frac{256}{81} \approx 3.1604938271 \dots$$

Uma discussão de como, possivelmente, os egípcios chegaram a esta aproximação pode ser encontrada em [9, p. 44].

A mais famosa aproximação de π calculada na antiguidade foi provavelmente a obtida por Arquimedes. O método utilizado por Arquimedes consistia em aproximar a área do círculo como estando entre as áreas de um polígono inscrito e de um polígono circunscrito. Usando polígonos de 96 lados, Arquimedes determinou que

$$(31) \quad 3.1408 \approx \frac{223}{71} < \pi < \frac{22}{7} \approx 3.1429.$$

Um método completamente diferente para calcular π foi encontrado pelo matemático indiano Mādhava, que viveu por volta do ano 1450. Segundo ele, o comprimento de um círculo de raio r é aproximadamente igual a

$$(32) \quad \frac{8r}{1} - \frac{8r}{3} + \frac{8r}{5} + \dots + (-1)^{n-1} \frac{8r}{2n-1} + (-1)^n \frac{8nr}{(2n)^2 + 1}.$$

Tomando $n = 10$ na fórmula acima, obtemos

$$\frac{36658359104}{5834363535} r.$$

Levando em conta que o comprimento da circunferência é igual a $2\pi r$, isto nos dá que π é aproximadamente igual a

$$3.1415902423.$$

Mais detalhes sobre a fórmula de Mādhava podem ser encontrados em [8, p. 224].

Toda esta conversa sobre aproximações de π deixa em aberto uma questão fundamental: quão boas são estas aproximações? Para isto precisamos medir o quanto estas aproximações se afastam do valor correto para π . A maneira mais ingênua de fazer isto é usar o erro absoluto. Suponhamos que x_* seja o valor calculado para uma quantidade cujo valor exato é x , então o *erro absoluto* incorrido ao neste cálculo é igual a $|x - x_*|$. Por exemplo, usando o método que será estudado na seção 5, sabemos que as primeiros 10 casas decimais de π são

$$3.1415926535$$

Como

$$(33) \quad 3.1415926535 - 3.1415902423 = 0.0000024112 < 0.000003,$$

podemos afirmar que o erro absoluto cometido no cálculo de π usando a fórmula de Mādhava é inferior a 0.000003. Note que, como não estamos usando um valor exato para π no cálculo da diferença (33), não sabemos o valor exato do erro absoluto, de modo que o máximo que podemos fazer é dar uma estimativa para este valor.

O caso da aproximação dada por Arquimedes é ainda mais interessante porque, da desigualdade (31) podemos afirmar diretamente que

$$\pi - \frac{223}{71} < \frac{22}{7} - \frac{223}{71} = \frac{1}{497};$$

de modo que se Arquimedes usasse *qualquer número entre* $223/71$ e $22/7$ como aproximação para π , ele incorreria em um erro absoluto inferior a

$$\frac{1}{497} < 0.0021.$$

Embora tenhamos definido o erro absoluto comparando o valor *calculado* para uma quantidade com seu valor exato, também podemos usá-lo para determinar o erro cometido quando efetuamos uma *medida*. Entretanto, neste caso, não existe a menor possibilidade de encontrarmos o valor exato da quantidade que está sendo medida; caso contrário, não estaríamos interessados em medi-la. O que acontece, na prática, é que podemos estimar uma cota superior para o valor absoluto da medida a partir da escala do instrumento de medida. Por exemplo, usando uma régua razoavelmente precisa, podemos facilmente medir um segmento de reta desenhado em um papel com erro inferior a 0.5 mm. Já uma balança digital trará, como parte de suas especificações, o erro máximo cometido em uma pesagem.

Note que, embora o erro absoluto formalize o que a maioria das pessoas entende pelo erro de uma medida, ele não é uma maneira adequada de avaliar a precisão desta medida. Por exemplo, digamos que, ao efetuar duas medidas de distância, cometemos em ambas um erro absoluto de 3m; só que uma delas é a distância entre o Rio e Petrópolis (cerca de 44.9Km), ao passo que a outra é a distância entre o Sol e Próxima Centauri, a estrela mais próxima de nós (cerca de 40208000000000Km). Embora o erro absoluto cometido nas duas medidas seja exatamente igual, é óbvio que a segunda é muito mais precisa, porque 3m representa uma fração muito menor da distância do Sol à Próxima Centauri do que do Rio à Petrópolis. Isto sugere que, para auferir qual de duas medidas é a mais precisa, devemos comparar não os erros absolutos, mas sim que frações das medidas totais estes erros representam.

Inspirados por situações como a do exemplo acima, definimos o *erro relativo* de uma medida como sendo

$$\frac{|x - x_*|}{|x|},$$

em que, como acima, x corresponde ao valor exato e x_* ao valor medido. Embora esta definição do erro relativo seja bastante útil do ponto de vista teórico, ela tem pouca utilidade prática. Afinal, o único valor normalmente conhecido para uma dada grandeza é o valor medido x_* . A princípio isto parece invalidar a definição do erro relativo mas, na realidade, o que ocorre é que se o erro absoluto for pequeno, faz pouca diferença calcular

$$\frac{|x - x_*|}{|x|} \quad \text{ou} \quad \frac{|x - x_*|}{|x_*|},$$

mais detalhes podem ser encontrados nos exercícios 10 a 13. Usando

$$\frac{|x - x_*|}{|x_*|}$$

para estimar o erro relativo cometido na medida da distância do Rio a Petrópolis, obtemos

$$\frac{3}{44900} \approx 0.000067,$$

ao passo que o erro relativo referente à distância do Sol à Próxima Centauri é de

$$\frac{3}{40208000000000} \approx 0.75 \cdot 10^{-13}.$$

Embora o erro absoluto tenha a mesma unidade usada na medição da grandeza cujo valor estamos determinando, o erro relativo é adimensional, porque é a razão entre dois valores expressos na mesma unidade de medida.

3. Ponto flutuante

Ainda que os primeiros computadores só fossem capazes de cálculos numéricos, qualquer computador pessoal que não seja peça de museu pode perfeitamente calcular corretamente com expressões simbólicas, desde que tenha sido instalado um software adequado. Um tal computador determinará que $\cos(\pi) = -1$ e que $\arctan(1) = \pi/4$, sem a necessidade de utilizar uma aproximação para π . Isto significa que, de certa maneira, podemos ensinar o computador a calcular de forma exata com π tal qual faríamos nós. Por outro lado, como a memória de um computador é finita, não podemos esperar que armazene toda a expansão decimal de π . Do ponto de vista prático, esta não é uma grande restrição. Uma expansão de π com 100 decimais tem uma precisão muito além de qualquer necessidade prática, muito embora possa ser facilmente escrita em uma folha de papel. Porém, tendo aceito que, não importa

qual seja a base, apenas expansões posicionais finitas podem ser representadas em um computador, ainda nos resta decidir qual a melhor maneira de fazer isto.

A resposta a esta pergunta depende de que tipo de números estão sendo representados. Por exemplo, uma planilha que lide apenas com vendas feitas por uma loja não precisa de mais do que duas casas decimais e, provavelmente, não mais que cinco casas à esquerda da vírgula. A situação é mais complicada quando um astrônomo efetua cálculos relativos a um universo que contém galáxias cuja massa ultrapassa 10^{43}g , mas que estão separadas por grandes vazios cuja densidade é inferior a 10^{-29}g/cm^3 .

A solução atualmente adotada para este problema foi proposta pela primeira vez pelo engenheiro espanhol Leonardo Torres y Quevedo em seu *Ensayos sobre automática*, publicado em 1914, e foi adotada nos computadores eletro-mecânicos que construiu a partir de 1914. A mesma solução foi redescoberta, independentemente, pelo engenheiro alemão Konrad Zuse quando construiu o primeiro computador mecânico programável em 1938.

O sistema proposto por Torres y Quevedo e Zuse é conhecido hoje em dia como *representação de ponto flutuante*. Para entender a ideia por trás desta maneira de representar números, imagine que você tem uma calculadora muito básica, cujo visor tem espaço para apenas quatro algarismos: dois para a parte inteira e dois para a parte decimal. Se precisar usar esta calculadora para determinar quanta farinha vai ser necessária para fazer uma dada receita de bolo para dezessete pessoas, você provavelmente vai ter que usar o peso em quilogramas; mas se precisar saber quanto pesam dezessete fuscas, vai ter que tomar a unidade de peso como sendo a tonelada, porque um fusca pesa 800Kg e $17 \times 800 = 13600$ não cabe no visor de sua calculadora. Note que se trata de uma filosofia semelhante à que nos levou a introduzir o conceito de erro relativo.

Ao contrário da calculadora imaginária do parágrafo anterior, que lhe obrigou a fazer uma mudança na unidade de medida, as modernas calculadoras dão conta destas mudanças de escala automaticamente. Para isto seu visor tem dois tipos de casas: n casas usadas para armazenar alguns dos n algarismos mais significativos do número a ser representado, à direita das quais encontram-se k casas extras, usadas para armazenar o expoente de uma potência de 10. Por exemplo, calculando $2^{20} = 1048576$, em uma calculadora na qual $n = 4$ e $k = 2$ obteremos

.1	0	4	8	0	6
----	---	---	---	---	---

que equivale a

$$0.1048 \cdot 10^7 = 1048000.$$

Um efeito colateral desta maneira de representar números é que, quanto maior for o número a ser calculado, tanto maior será o erro absoluto entre o valor exato e o que

foi armazenado pela calculadora. No exemplo acima, o erro absoluto é igual a

$$1048576 - 1048000 = 576;$$

mas se usarmos a mesma calculadora para achar $2^{40} = 1099511627776$, teremos um erro absoluto igual a

$$1099511627776 - 0.1099 \cdot 10^{14} = 511627776.$$

Entretanto, o que parece uma falha desastrosa deste sistema de representação, desaparece se compararmos os erros relativos referentes aos dois cálculos, que são

$$\frac{576}{2^{20}} \approx 0.00054932$$

no primeiro exemplo e

$$\frac{511627776}{2^{40}} \approx 0.00046532$$

no segundo. Tendo descrito a ideia que norteia a representação em ponto flutuante, precisamos formulá-la de maneira suficientemente precisa, para que possa ser usada sem risco de erro. Embora este tipo de representação possa ser utilizado para qualquer base, vamos detalhá-lo apenas no caso em que a base é 10; a extensão a outras bases é fácil de fazer e fica por sua conta.

Fixados três inteiros positivos n , m e M , definimos o *conjunto de números de ponto flutuante* \mathbb{F} como sendo o conjunto de números reais r da forma

$$(34) \quad r = \pm 0.a_1 \dots a_n \cdot 10^e,$$

em que

- a_1, \dots, a_n são algarismos, isto é, números inteiros entre 0 e 9;
- $a_1 \neq 0$;
- $-m \leq e \leq M$ é um número inteiro.

O número $0.a_1 \dots a_n$ é a *mantissa* de r e a exigência de que a_1 seja diferente de zero garante que cada elemento de \mathbb{F} tenha apenas uma representação na forma dada em (34). É importante ter claro que \mathbb{F} é um conjunto finito contido no intervalo

$$I = [-0.9 \dots 9 \cdot 10^M, 0.9 \dots 9 \cdot 10^M]$$

e cujo menor elemento positivo é

$$\alpha = 0.10 \dots 0 \cdot 10^{-m}.$$

Como consequência disto, não somente há números reais grandes demais, ou pequenos demais, para serem representados em \mathbb{F} , como há infinitos números reais no intervalo

I que não são elementos de \mathbb{F} . Como se isso não bastasse, o produto e a soma da maior parte dos números de \mathbb{F} não pertence a este conjunto; por exemplo,

$$2 \times \underbrace{0.73 \dots 3}_{n+1 \text{ algarismos}} = \underbrace{0.146 \dots 6}_{n+2 \text{ algarismos}} \cdot 10^1 \notin \mathbb{F}$$

mesmo sendo um número entre $-\alpha$ e α .

Contornamos estes problemas introduzindo a função *arredondamento*

$$\text{fl} : \mathbb{R} \rightarrow \mathbb{F} \cup \{-\infty, +\infty\}$$

que, dado um número real r , retorna

$$\text{fl}(r) = \begin{cases} \text{o elemento de } \mathbb{F} \text{ mais próximo de } r & \text{se } r \in I \\ -\infty & \text{se } r < -0.9 \dots 9 \cdot 10^M \\ +\infty & \text{se } r > 0.9 \dots 9 \cdot 10^M. \end{cases}$$

Se você leu cuidadosamente a definição acima, deve ter notado que pode haver empate no caso em $r \in I$, porque r pode ser equidistante de dois elementos de \mathbb{F} . Caso isto aconteça, arredondaremos sempre para o elemento de \mathbb{F} mais próximo cuja n -ésima casa decimal é par. Por exemplo, se $n = 4$ e $m = M = 10$, então 0.12345 está a meio caminho entre 0.1234 e 0.1235, de modo que

$$\text{fl}(0.12345) = 0.1234;$$

por outro lado, 0.12355 está a meio caminho entre 0.1235 e 0.1236, de modo que

$$\text{fl}(0.12355) = 0.1236.$$

A estratégia de desempate que estamos usando não é a única possível; uma estratégia mais simples seria arredondar sempre para o elemento de \mathbb{F} menor, que equivale a *truncar* o número entre a n -ésima e $n + 1$ -ésima decimal. Mais detalhes podem ser encontrados em [5, Chapter 2].

A função arredondamento nos permite definir operações equivalentes às básicas da aritmética no conjunto \mathbb{F} . Para isto precisamos introduzir um pouco de notação. Sejam $+$, $-$, \times e \div as operações usuais entre números reais; as operações correspondentes em \mathbb{F} serão denotadas pelo mesmo símbolo dentro de um círculo. Por exemplo, a soma em \mathbb{F} será \oplus e a multiplicação será \otimes . Se $f_1, f_2 \in \mathbb{F}$ e \circ for $+$, $-$, \times ou \div , então

$$f_1 \odot f_2 = \text{fl}(f_1 \circ f_2).$$

Em outras palavras, $f_1 \odot f_2$ é o elemento de \mathbb{F} mais próximo $f_1 \circ f_2$, que é o resultado obtido aplicando a f_1 e f_2 a operação \circ de \mathbb{R} . Supondo, mais uma vez, que $n = 4$, temos que

$$0.1234 \oplus 0.9821 = \text{fl}(0.1234 + 0.9821) = \text{fl}(0.11055 \times 10) = 0.1106 \times 10^1.$$

Por outro lado,

$$0.1234 \otimes 0.9821 = \text{fl}(0.1234 \times 0.9821) = \text{fl}(0.97047 \times 10^{-1}) = 0.9705 \times 10^{-1}.$$

Infelizmente a maneira como estas operações foram definidas pode trazer problemas dos quais você precisa estar ciente. O primeiro deles é que algumas das propriedades mais básicas das operações aritméticas em \mathbb{R} podem falhar para suas correspondentes em \mathbb{F} . Por exemplo, a soma em \mathbb{R} é uma operação associativa; isto é, se r_1 , r_2 e r_3 são números reais, então

$$r_1 + (r_2 + r_3) = (r_1 + r_2) + r_3.$$

Você usa esta propriedade toda vez que soma vários números reais e espera obter o mesmo resultado não importando se começa a somar da esquerda para à direita ou da direita para à esquerda. No entanto \oplus não é associativa, como ilustra o seguinte exemplo. Suponhamos que $n = 4$ e que

$$f_1 = 10^{30}, \quad f_2 = -10^{30} \quad \text{e} \quad f_3 = 1.$$

Por um lado, como $f_2 = -f_1$, temos que

$$(f_1 \oplus f_2) \oplus f_3 = \text{fl}(f_1 + f_2) \oplus f_3 = 0 \oplus f_3 = \text{fl}(0 + f_3) = f_3,$$

pois $f_3 \in \mathbb{F}$. Contudo,

$$(f_2 \oplus f_3) = \text{fl}(-10^{30} + 1) = \text{fl}(-\underbrace{9 \dots 9}_{29 \text{ vezes}}) = \text{fl}(-0.9 \dots 9 \times 10^{30}) = -0.1 \times 10^{31};$$

donde

$$f_1 \oplus (f_2 \oplus f_3) = \text{fl}(10^{30} - 0.1 \times 10^{31}) = \text{fl}(0) = 0.$$

Assim,

$$f_1 \oplus (f_2 \oplus f_3) = 0 \neq -10^{30} = (f_1 \oplus f_2) \oplus f_3,$$

não têm sequer a mesma ordem de grandeza.

O problema ilustrado no exemplo acima é conhecido como *subtração catastrófica*. Para convencer você de que o adjetivo não é um exagero, vamos resolver a equação

$$(35) \quad 0.1 x^2 + 400.0 x + 0.1 = 0$$

utilizando a fórmula usual, em um computador com quatro casas decimais. Sob estas restrições,

$$400^2 - 4 \times 0.1 \times 0.1 = 159999.96 = 0.15999996 \times 10^6 \approx 0.16 \times 10^6$$

cuja raiz quadrada é igual a 0.4×10^3 . Com isso, aplicando a fórmula, obtemos

$$\frac{-400 \pm 400}{0.2},$$

segundo a qual as raízes de (35) são 0 e -4000 . Calculando as mesmas raízes, usando a mesma fórmula, mas com uma precisão melhor, obtemos

$$r_1 = -3999.99975, r_2 = -0.0002499818802$$

Para ter uma ideia de quão ruim são as aproximações calculadas com a mantissa de quatro algarismos, basta determinar os erros relativos, que são

$$\frac{|-4000 + 3999.99975|}{|-3999.99975|} \approx 0.9 \quad \text{e} \quad \frac{|0 + 0.0002499818802|}{|-0.0002499818802|} = 1.$$

Esses erros, você há de concordar, são ruins o suficiente para merecerem a qualificação de catastróficos. É claro que uma mantissa de quatro algarismos é inaceitavelmente pequena, mas exemplos semelhantes a este podem ser inventados qualquer que seja a quantidade (finita!) de algarismos na mantissa. Para isto, basta que, na equação $ax^2 + bx + c = 0$, o valor de b^2 seja muito maior que o de $4ac$.

Às vezes é possível diminuir o efeito danoso da subtração catastrófica no cálculo de $b^2 - 4ac$ reescrevendo ligeiramente a fórmula usual. Multiplicando o numerador e o denominador de

$$\frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

por $-b - \sqrt{b^2 - 4ac}$, obtemos

$$(36) \quad \frac{-b + \sqrt{b^2 - 4ac}}{2a} \left(\frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}} \right) = \frac{2c}{(-b - \sqrt{b^2 - 4ac})}$$

que nos dá,

$$\frac{0.2}{-400 - 400} = -0.00025 \approx -0.0002$$

quando aplicada à equação (35), no computador com a mantissa de quatro algarismos. Com isto o erro relativo cai de 1.0 para

$$\frac{|-0.0002 + 0.0002499818802|}{|-0.0002499818802|} \approx 0.19994175575981931485 \approx 0.2$$

que é, sem dúvida, melhor que o erro anterior. Observe que não podemos aplicar um truque semelhante ao cálculo da outra raiz de (35), por que se fizéssemos isto, obteríamos

$$\frac{-b - \sqrt{b^2 - 4ac}}{2a} \left(\frac{-b + \sqrt{b^2 - 4ac}}{-b + \sqrt{b^2 - 4ac}} \right) = \frac{2c}{(-b + \sqrt{b^2 - 4ac})}$$

que é uma fração cujo denominador é zero, por causa da subtração catastrófica.

Falta escrever pelo menos um parágrafo sobre o IEEE-754.

4. Newton e a aproximação de funções

Nas seções anteriores discutimos como aproximar números reais por números em ponto flutuante; a partir desta seção discutiremos como fazer o mesmo para funções reais. Talvez você esteja tentando imaginar o que quer dizer *aproximar uma função*: aproximar usando o quê? Na verdade a resposta remonta à invenção do cálculo por Newton, no século XVII.

O curioso é que, embora tenha estudado cálculo, você possivelmente não se deparou com o problema de como aproximar funções. A razão para isto é histórica: o cálculo foi inventado no século XVII, independentemente por Newton e Leibniz. Embora essencialmente equivalentes, as versões originais destes dois matemáticos são bastante diferentes. É na versão de Leibniz que se baseiam todas as exposições atuais do cálculo, que adotam tanto sua notação, quanto a apresentação do conteúdo como uma sequência de regras a partir das quais podemos derivar ou integrar qualquer *função elementar*, que são que podem ser obtidas combinando funções polinomiais, trigonométricas, exponenciais e logarítmicas, pelas operações de soma, subtração, multiplicação, divisão, radiciação e composição. Contudo, é à versão de Newton que devemos nos voltar para entender como aproximar funções.

A descoberta crucial do Newton, que o conduziu à sua versão do cálculo, foi que o binômio pode ser representado por uma soma infinita de potências, multiplicadas por números binomiais. Se esta estória de *soma infinita* lhe soa estranho, é porque a fórmula do *binômio de Newton* que você aprendeu no ensino médio vale apenas no caso em que o expoente é um número inteiro positivo. Ironicamente, este caso específico já era conhecido do matemático persa Al-Karaji, que viveu por volta do ano 1000. Newton, porém, percebeu que, qualquer que seja $\alpha \in \mathbb{R}$, teremos

$$(37) \quad (1+x)^\alpha = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k, \quad \text{em que} \quad \binom{\alpha}{k} = \frac{\alpha \cdot (\alpha-1) \cdots (\alpha-k+1)}{k!}.$$

Observe que, quando $\alpha = n > 2$ for um número inteiro e $k > n$, o número binomial $\binom{n}{k}$ será igual a zero, porque um dos termos do produto no numerador será igual a $n-n$; deste modo a fórmula geral reduz-se à que estudamos no ensino médio quando o expoente é um inteiro positivo. Usando a terminologia atual, o que Newton mostrou é que a função $g(x) = (1+x)^\alpha$ admite a expansão em série de potências da forma

$$\sum_{k=0}^{\infty} \binom{\alpha}{k} x^k = \binom{\alpha}{0} + \binom{\alpha}{1} x + \binom{\alpha}{2} x^2 + \cdots ;$$

uma *série de potências* é simplesmente uma soma infinita, cuja k -ésima parcela é o produto de uma constante por x^k . Uma outra função para a qual você conhece uma expansão em série de potências é $1/(1-x)$, com $x \in [0, 1)$; isto porque, sob esta

restrição para x temos que

$$(38) \quad 1 + x + x^2 + x^3 + \cdots = \frac{1}{1 - x},$$

pela fórmula da soma de uma progressão geométrica infinita cuja razão é menor que um.

Muito interessante; mas o que isto tem a ver com a aproximação de funções? Vamos deixar que o próprio Newton responda a esta pergunta. Em um manuscrito de outubro de 1666, Newton afirma que problemas como o de encontrar a expansão em série de potências do binômio deveriam ser tratados

como se você estivesse resolvendo a equação em número decimais, por divisão, ou extração de raízes [...]; esta operação pode ser continuada tão longe quanto desejado, quanto mais longe melhor[.] [13, p. 118]

Embora a citação originalmente refira-se, como dissemos, à expressão para o binômio, usaremos a função $h(x) = 1/(1 - x)$ para explicar o que Newton quis dizer. A pergunta, na verdade, é: como obter a expansão em série de $1/(1 - x)$. A resposta, segundo Newton, é que basta efetuar a divisão de 1 por $1 - x$ como polinômios, exceto que, neste caso, o grau do resto vai aumentar em vez de diminuir, como ilustrado abaixo:

$$\begin{array}{r} 1 \\ -1 \quad +x \\ \hline x \\ -x \quad +x^2 \\ \hline x^2 \\ -x^2 \quad +x^3 \\ \hline x^3 \\ -x^3 \quad +x^4 \\ \hline x^4 \end{array} \quad \left| \begin{array}{r} 1 - x \\ 1 + x + x^2 + x^3 \end{array} \right.$$

Observe que se continuarmos fazendo isto ao infinito, obteremos a expansão em série de $1/(1 - x)$ apresentada em (38).

O ponto crucial é a analogia que Newton estabelece entre achar a expansão em série de potência de $1/(1 - x)$ e a maneira como achamos a expansão decimal de $1/3$ usando o algoritmo de divisão longa. Para Newton, o polinômio $1 + x + \cdots + x^k$ que obtemos quando paramos, no k -ésimo passo, a execução do algoritmo de divisão de

polinômios aplicado a 1 e $1 - x$ é uma aproximação da função $1/(1 - x)$. Além disso, quanto maior k , melhor será a aproximação.

Em um manuscrito de 1669 intitulado *De Analysi per aequationes numero terminorum infinitas* Newton mostrou como usar argumentos semelhantes para encontrar as séries de potências para as funções exponencial, logaritmo, seno e cosseno. Entretanto, Newton não foi o primeiro a achar estas séries, expansões para algumas destas funções já haviam sido obtidas por Mādhava, o matemático indiano mencionado na seção 2, mas seu trabalho era desconhecido no ocidente à época de Newton. O passo final nesta história, porém, só foi dado em 1715 pelo matemático inglês Brook Taylor que mostrou como obter uma fórmula que possa ser aplicada a qualquer função elementar. Na próxima seção estudaremos uma versão modernizada do resultado de Taylor; mais precisamente, mostraremos como é construir um polinômio de grau k que aproxima uma dada função com erro tão pequeno quanto desejado.

5. A fórmula de Taylor com resto

A maioria das funções que estudamos nos cursos de cálculo são muito bem comportadas no intervalo em que estão definidas, uma das poucas exceções é a função valor absoluto (ou módulo); quase todas as demais têm derivada de qualquer ordem. No mundo real, entretanto, as coisas nem sempre são tão satisfatórias. Por isso é conveniente supor que a função cuja aproximação por polinômios estamos buscando não têm derivadas de todas as ordens possíveis. A isso você pode muito bem contrapor: por que supor que tem alguma derivada? Por que não admitir apenas que a função é contínua? A resposta é que este é um problema mais difícil, ao qual voltaremos no capítulo 5.

Seja, pois, $f : [a, b] \rightarrow \mathbb{R}$ uma função contínua, cujas $n + 1$ primeiras derivadas existem e são contínuas no intervalo (a, b) . O *teorema fundamental do cálculo* nos permite escrever

$$(39) \quad f(b) - f(a) = \int_a^b \frac{df}{dt} dt = \int_a^b f'(t) dt$$

A estratégia para obter o polinômio que aproxima esta função consiste em integrar por partes o lado direito da fórmula acima. Contudo, como no caso da divisão de polinômios calculada na seção 4, vamos efetuar a integração na “direção errada”, querendo dizer com isto que a integral do lado direito será substituída por uma expressão mais complicada ainda. Para isto tomamos,

$$du = -dt \quad \text{e} \quad v = -f'(t),$$

de modo que

$$u = b - t \quad \text{e} \quad dv = -f''(t) dt.$$

Lembrando que

$$\int_a^b v du = uv \Big|_a^b - \int_a^b u dv,$$

obtemos

$$\int_a^b f'(t) dt = -(b-t)f'(t) \Big|_a^b + \int_a^b (b-t)f''(t) dt = f'(a)(b-a) + \int_a^b (b-t)f''(t) dt.$$

Substituindo em (39),

$$(40) \quad f(b) - f(a) = f'(a)(b-a) + \int_a^b (b-t)f''(t) dt.$$

Será que você percebeu a malícia na escolha de $du = -dt$ e $u = b-t$, em vez de $du = dt$ e $u = t$? Fizemos isto para garantir que sobrasse apenas uma parcela em $uv|_a^b$, o que descomplica significativamente a fórmula final.

A próxima etapa é semelhante à primeira e consiste em integrar

$$\int_a^b (b-t)f''(t) dt$$

por partes, sempre escolhendo a “direção errada”. Neste caso, isto corresponde a tomar

$$du = -(b-t)dt \quad \text{e} \quad v = -f''(t),$$

donde

$$u = \frac{(b-t)^2}{2} \quad \text{e} \quad dv = -f'''(t)dt.$$

Assim,

$$\int_a^b f''(t) dt = -(b-t)^2 f''(t) \Big|_a^b + \frac{1}{2} \int_a^b (b-t)^2 f'''(t) dt = \frac{f''(a)}{2}(b-a)^2 + \int_a^b (b-t)^2 f'''(t) dt.$$

Substituindo em (40), obtemos

$$(41) \quad f(b) - f(a) = f'(a)(b-a) + \frac{f''(a)}{2}(b-a)^2 + \frac{1}{2} \int_a^b (b-t)^2 f'''(t) dt.$$

A esta altura você já deve estar começando a perceber um padrão; mas precisamos de mais uma integração para tornar claro como serão os denominadores das parcelas das forma $f^{(k)}(a)(b-a)^k$. Tomando, então,

$$du = -(b-t)^2 dt \quad \text{e} \quad v = -f'''(t),$$

obteremos

$$u = \frac{(b-t)^3}{3} \quad \text{e} \quad dv = -f^{(4)}(t)dt.$$

donde

$$f(b) - f(a) = f'(a)(b-a) + \frac{f''(a)}{2}(b-a)^2 + \frac{f'''(a)}{2 \cdot 3}(b-a)^3 + \frac{1}{2 \cdot 3} \int_a^b (b-t)^3 f^{(iv)}(t) dt.$$

Portanto, ao final da próxima integração por partes os denominadores da parcela $f^{(iv)}(b-a)^4$ e da constante que multiplica a integral serão ambos iguais a $2 \cdot 3 \cdot 4 = 4!$

Podemos continuar integrando por partes, à guisa das três etapas anteriores, enquanto f for diferenciável. Como estamos supondo que as primeiras $n+1$ derivadas de f existem e são contínuas, obteremos a seguinte fórmula ao final de n integrações por partes,

$$f(b) - f(a) = f'(a)(b-a) + \cdots + \frac{f^{(n)}(a)}{n!}(b-a)^n + \frac{1}{n!} \int_a^b (b-t)^n f^{(n+1)}(t) dt.$$

Passando $f(a)$ para o lado direito e usando somatórios, obtemos

$$(42) \quad f(b) = \sum_{j=0}^n \frac{f^{(j)}(a)}{j!} (b-a)^j + \frac{1}{n!} \int_a^b (b-t)^n f^{(n+1)}(t) dt.$$

Diremos que

$$P_n(x) = \sum_{j=0}^n \frac{f^{(j)}(a)}{j!} (x-a)^j$$

é o *polinômio de Taylor de grau n* de $f(x)$ em $x = a$. Reescrevendo (42) na forma

$$f(b) - P_n(b) = \frac{1}{n!} \int_a^b (b-t)^n f^{(n+1)}(t) dt,$$

verificamos que $P_n(b)$ é uma aproximação de $f(b)$ com erro igual a

$$E_n(b) = \frac{1}{n!} \int_a^b (b-t)^n f^{(n+1)}(t) dt.$$

Por isso (42) é conhecida como a *fórmula de Taylor com resto integral*; o *resto* é o nome pelo qual o erro $E_n(b)$ é usualmente conhecido.

O problema da forma de Taylor (42) é que ela não expressa $P_n(x)$ como uma aproximação da função $f(x)$ *todo o intervalo* $[a, b]$, mas sim do número $P_n(b)$ como aproximação para $f(b)$. Em outras palavras, para garantir que $P_n(x)$ seja uma aproximação de $f(x)$ precisamos de uma maneira de controlar o erro *em todo o intervalo* $[a, b]$. Entretanto, como $f^{(n+1)}(x)$ é contínua em $[a, b]$, ela atinge um valor máximo em $[a, b]$; digamos que

$$M_{n+1} = \max\{|f^{(n+1)}(x)| \mid x \in [a, b]\}.$$

Pela desigualdade triangular

$$\left| \int_a^x (x-t)^n f^{(n+1)}(t) dt \right| \leq \int_a^x (x-t)^n |f^{(n+1)}(t)| dt \leq M_{n+1} \left| \int_a^x (x-t)^n dt \right|.$$

Entretanto, como

$$\int_a^x (x-t)^n dt = \frac{(x-t)^{n+1}}{n+1} \Big|_a^x = \frac{(x-a)^{n+1}}{n+1}$$

podemos escrever

$$|E_n(x)| \leq \frac{M_{n+1}}{(n+1)!} (x-a)^{n+1}.$$

Finalmente, como $x-a < b-a$, para todo $x \in [a, b]$,

$$(43) \quad |E_n(x)| \leq \frac{M_{n+1}}{(n+1)!} (b-a)^{n+1},$$

que nos dá um valor máximo para o erro válido em todos os pontos do intervalo $[a, b]$. Para referência futura, vamos enunciar o que aprendemos até aqui como um teorema.

TEOREMA DE TAYLOR. *Seja $f : [a, b] \rightarrow \mathbb{R}$ uma função contínua, cujas primeiras $n+1$ derivadas existem e são contínuas em (a, b) . O n -ésimo polinômio de Taylor de $f(x)$ em $x = a$ é*

$$P_n(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n.$$

Se M_{n+1} for o valor máximo de $|f^{(n+1)}(x)|$ no intervalo $[a, b]$, então

$$|f(x) - P_n(x)| < \frac{M_{n+1} \cdot (b-a)^{n+1}}{(n+1)!}$$

para todo $x \in (a, b)$.

Antes de passar aos exemplos, é importante chamar sua atenção que, a despeito da maneira como foi formulado acima, o teorema de Taylor *não* se aplica apenas à extremidade inferior do intervalo de definição da função f . Se $f : [\alpha, \beta] \rightarrow \mathbb{R}$ for uma função e

$$\alpha \leq a < b \leq \beta,$$

podemos aplicar o teorema em $x = a$ desde que f seja contínua em $[a, b]$ e tenha suas primeiras $n+1$ derivadas contínuas em (a, b) .

Em nosso primeiro exemplo, queremos aproximar $f(x) = x \cos(x)$, no intervalo $[0, 5]$, pelo seu polinômio de Taylor de grau 3. Calculando as derivadas necessárias,

temos que

$$\begin{aligned}f'(x) &= -x \operatorname{sen}(x) + \cos(x) \\f''(x) &= -2 \operatorname{sen}(x) - x \cos(x) \\f'''(x) &= x \operatorname{sen}(x) - 3 \cos(x),\end{aligned}$$

donde

$$f(0) = 0, \quad f'(0) = 1, \quad f''(0) = 0 \quad \text{e} \quad f'''(0) = -3.$$

Logo, o polinômio de Taylor de grau 3 de $f(x)$ é

$$P_3(x) = x - \frac{3x^3}{3!} = x - \frac{x^3}{2}.$$

Por outro lado, como

$$f^{(\text{iv})}(x) = 4 \operatorname{sen}(x) + x \cos(x)$$

e, tanto o seno, quanto o cosseno, só tomam valores entre -1 e 1 temos, pela desigualdade triangular, que

$$|f^{(\text{iv})}(x)| \leq 4 |\operatorname{sen}(x)| + |x| |\cos(x)| \leq 4 + 5 = 9,$$

para todo $0 \leq x \leq 5$, de modo que $M_4 \leq 9$. Segue-se, então, de (43) que

$$|E_n(x)| \leq \frac{9 \cdot 5^4}{4!} = \frac{1875}{8}.$$

Mas este é um número muito grande, o que faz com que $P_4(x)$ seja, provavelmente, inútil como aproximação para $f(x)$ para qualquer propósito prático, como mostra a figura 1.

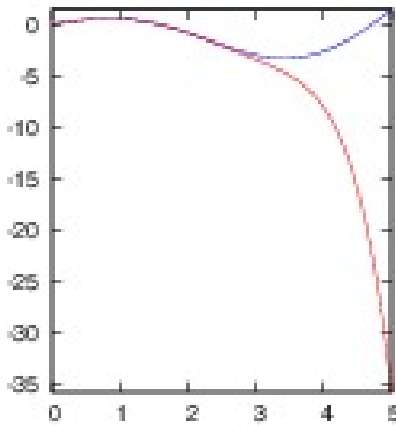


FIGURA 1. Aproximação de Taylor de grau 4 da função $y = x \cos(x)$

Talvez lhe ocorra que deveria ser possível obter uma cota superior melhor que 1 para $|\operatorname{sen}(x)|$ ou $|\cos(x)|$ no intervalo $[0, 5]$. Infelizmente, isto não ocorre, porque 0 e

$\pi/2$ pertencem a $[0, 5]$, mas $|\cos(0)| = 1$ e $|\sin(\pi/2)| = 1$. Portanto, neste exemplo, a única saída viável consiste escolher um valor de n grande o suficiente, para que tenhamos uma aproximação aceitável. Com isto em mente, a primeira pergunta que precisamos fazer é: o que é um erro aceitável para este problema? Na prática o erro vai depender do que pretendemos fazer com a aproximação; digamos que, no nosso exemplo, este erro não pudesse passar de 0.01. Começamos observando que (43) nos dá

$$(44) \quad |E_n(x)| \leq \frac{M_{n+1} \cdot 5^{n+1}}{(n+1)!} < 0.01.$$

Logo, nosso próximo passo deve consistir em determinar uma cota superior para M_{n+1} . Contudo, M_{n+1} depende de n , o que nos obriga a achar uma lei de formação para a derivada. Neste caso isto é bastante simples de fazer e nos dá

$$f^{(n)}(x) = \begin{cases} -n \sin(x) - x \cos(x) & \text{quando } n \text{ é par} \\ x \sin(x) - n \cos(x) & \text{quando } n \text{ é ímpar.} \end{cases}$$

Levando em conta que $|\sin(x)| \leq 1$ e $|\cos(x)| \leq 1$ e que $|x| \leq 5$ obtemos, nos dois casos, que

$$|f^{(n)}(x)| \leq n + 5.$$

Substituindo isto em (44),

$$|E_n(x)| \leq \frac{(n+5) \cdot 5^{n+1}}{(n+1)!} < 0.01.$$

Com a ajuda de um computador, é fácil verificar que $n = 18$ é o primeiro valor para o qual esta desigualdade é verificada, de modo que o polinômio de Taylor

$$P_{18}(x) = 17x^{17} - 15x^{15} + 13x^{13} - 11x^{11} + 9x^9 - 7x^7 + 5x^5 - 3x^3 + x$$

satisfaz

$$(45) \quad |f(x) - P_{18}(x)| < 0.01 \quad \text{para todo } x \in [0, 5].$$

Não é difícil dar um argumento menos brutal para achar n , mas é importante ter claro que, com ele, *não* há garantia de que encontraremos o menor valor possível para n . Para começar, determinamos, usando uma calculadora, que a desigualdade não é verificada para $n = 10$, o que nos permite supor que $n \geq 10$. Mas, sob esta hipótese, $n! \geq 9! \cdot 10^{n-9}$, donde

$$\frac{1}{n!} \leq \frac{1}{9! \cdot 10^{n-9}}.$$

Logo,

$$\frac{5^{n+1}}{n!} \leq \frac{5^{10} \cdot 5^{n-9}}{9! \cdot 10^{n-9}} = \frac{5^{10}}{9!} \left(\frac{1}{2}\right)^{n-9}.$$

Portanto, basta que

$$\left(\frac{1}{2}\right)^{n-9} \leq \frac{9! \cdot 0.01}{5^{10}} < 0.0003.$$

Aplicando logaritmos na base 10 à desigualdade mais à direita e levando em conta que o logaritmo é uma função crescente,

$$-(n-9) \log_{10}(2) \leq \log_{10}(3) - 4,$$

que nos dá

$$(n-9) \geq \frac{4 - \log_{10}(3)}{\log_{10}(2)} > 11,$$

donde $n \geq 20$ que, como seria de esperar, não produz o menor valor possível.

A desigualdade (45) admite uma interpretação geométrica bastante simples; segundo ela, o gráfico de $P_{18}(x)$ está contido em uma faixa de largura 0.02 cujo centro é a curva $y = x \cos(x)$. Como uma faixa com esta largura é estreita demais para ser visível quando $0 \leq x \leq 5$, ilustramos na figura 2 a faixa de largura 26, correspondente à desigualdade

$$|f(x) - P_{10}(x)| < 13 \quad \text{para todo } x \in [0, 5].$$

Os gráficos de $y = x \cos(x)$ e $y = P_{10}(x)$ aparecem em azul e verde, respectivamente; as curvas em vermelho correspondem aos limites superior e inferior da faixa de largura 26 com centro em $y = x \cos(x)$.

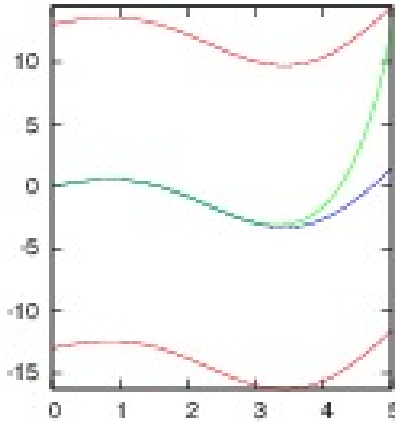


FIGURA 2. A aproximação de Taylor de grau 10 (em verde) da função $y = x \cos(x)$ (em azul) fica entre as curvas $y = x \cos(x) \pm 13$ (em vermelho).

Vejamos outro exemplo. Seja $f(x) = x \ln(x)$. Qual seria o valor de n necessário para que pudéssemos usar o polinômio de Taylor $P_n(x)$ para obter uma aproximação de $f(1.01)$ com erro absoluto inferior a 10^{-11} ? Neste caso não estamos tentando

aproximar o gráfico de $f(x)$ pelo de $P_n(x)$; queremos usar o polinômio de Taylor apenas para achar uma aproximação extremamente boa para $f(1.01)$. Como 1.01 está muito próximo de 1 e como $\ln(1) = 0$, construiremos $P_n(x)$ em $a = 1$. Calculando algumas derivadas, verificamos que

$$\begin{aligned}f'(x) &= \ln(x) + 1 \\f''(x) &= x^{-1} \\f'''(x) &= -x^{-2} \\f^{(iv)}(x) &= 2x^{-3} \\f^{(v)}(x) &= -2 \cdot 3 \cdot x^{-4},\end{aligned}$$

que nos permite concluir que

$$f^{(n)}(x) = (-1)^n (n-2)! \cdot x^{1-n}.$$

Levando em conta que x^{-n} é decrescente quando $x > 1$, obtemos

$$|f^{(n+1)}(x)| = |(n-1)! \cdot x^{-n}| < (n-1)!.$$

donde, $M_{n+1} < (n-1)!$. Logo, o teorema de Taylor nos garante que,

$$|f(x) - P_n(x)| = \frac{(n-1)!}{(n+1)!} (x-1)^{n+1} = \frac{1}{n(n+1)} (x-1)^{n+1}$$

para todo $1 \leq x \leq 1.01$. Quando $x = 1.01$, isto nos dá

$$|f(x) - P_n(x)| \leq \frac{1}{n(n+1)} (0.01)^{n+1} = \frac{1}{n(n+1)10^{2(n+1)}}.$$

O valor de n desejado deve, portanto, satisfazer

$$\frac{1}{n(n+1)10^{2(n+1)}} < 10^{-11},$$

o que ocorre quando $n = 4$, pois

$$\frac{1}{4 \cdot 5 \cdot 10^{10}} = 0.5 \cdot 10^{-11}.$$

Como,

$$P_4(x) = (x-1) + \frac{1}{2}(x-1)^2 - \frac{1}{6}(x-1)^3 + \frac{1}{12}(x-1)^4,$$

as primeiras 20 casas decimais da aproximação desejada são

$$P_4(1.01) \approx 0.0100498341666666667.$$

Para encerrar usaremos o teorema de Taylor para explicar como a aproximação de Mādhava para π pode ser obtida. Começaremos com o caso da fórmula para $n = 5$. Para isto, aproximamos a função $f(x) = \arctan(x)$ pelo seu polinômio de Taylor de

grau dez em $x = 0$. Infelizmente as derivadas do arco-tangente de ordem alta são dadas por funções complicadas; por exemplo, a quinta derivada é

$$\frac{120 x^4 - 240 x^2 + 24}{x^{10} + 5 x^8 + 10 x^6 + 10 x^4 + 5 x^2 + 1}.$$

Por isso deixaremos os cálculos a cargo do computador e escreveremos, diretamente, o polinômio de Taylor resultante

$$P_{10}(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \frac{x^9}{9}$$

e a cota

$$|E_{10}(x)| \leq \frac{2393802.79711}{11!} \approx 0.059969807$$

para seu resto. Para calcular π a partir deste polinômio basta lembrar que $\arctan(1) = \pi/4$ e usar

$$P_{10}(1) = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} = \frac{263}{315},$$

como aproximação de $\pi/4$. Observe que a soma que define $P_{10}(1)$ é a mesma da fórmula de Mādhava com $n = 5$, *exceto pelo último termo de (32)*. A aproximação de π decorrente do cálculo acima é

$$\pi \approx 3.3396825396.$$

Note que o erro cometido no cálculo de $\pi/4$, que é da ordem de 0.06, foi quadruplicado quando usamos $4 \cdot P_{10}(1)$ como aproximação para π .

Talvez você esteja se perguntando como Mādhava foi capaz de calcular o polinômio de Taylor de ordem n de $\arctan(x)$, quando precisamos de um computador para obter o polinômio de grau cinco. A resposta é que nem sempre a fórmula geral provê a maneira mais eficiente para se calcular o polinômio de Taylor. No caso do arco-tangente é preferível partir de

$$\frac{1}{1-x} = 1 + x + x^2 + \cdots + x^n + \frac{x^{n+1}}{1-x},$$

que pode ser facilmente encontrada usando o processo de divisão apresentado na seção 4. Substituindo x por $-t^2$, obtemos

$$\frac{1}{1+t^2} = 1 - t^2 + t^4 - t^6 + \cdots + (-1)^n t^{2n} + \frac{t^{2(n+1)}}{1+t^2},$$

e integrando ambos os lados entre $t = 0$ e $t = x$,

$$\arctan(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots + (-1)^n \frac{x^{2n+1}}{2n+1} + \int_0^x \frac{t^{2(n+1)}}{1+t^2} dt.$$

em que a integral que sobrou é o erro cometido quando

$$P_n(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots + (-1)^n \frac{x^{2n+1}}{2n+1}$$

é usado para aproximar $\arctan(x)$. Apesar de, em princípio, ser possível obter uma aproximação de $\arctan(1)$ tão boa quanto desejada usando $P_n(1)$, na prática isto requer que sejam escolhidos valores muito grandes para n . Por exemplo, mesmo tomando $n = 50$, o erro ainda é da ordem de 0.01. É por isso que a fórmula (32) inclui um termo extra, que não está entre as parcelas de $P_n(x)$, e que ajuda a reduzir o tamanho do erro.

⚡ Dizemos que uma função é *infinitamente diferenciável* em um intervalo se admite derivadas de toda ordem em todos os pontos de intervalo. A uma função $f : [a, b] \rightarrow \mathbb{R}$, infinitamente diferenciável em (a, b) e um ponto $x_0 \in (a, b)$, podemos associar a soma infinita

$$\tau_{x_0}(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}}{k!} (x - x_0)^k,$$

conhecida como *série de Taylor* de $f(x)$ em x_0 . Quando

$$f(x) = \tau_{x_0}(x), \text{ para todo } x \in (a, b),$$

dizemos que f é *analítica* em (a, b) . Quase todas as funções infinitamente diferenciáveis estudadas nos cursos de cálculo são analíticas. Uma das poucas exceções é

$$S(x) = \begin{cases} 0 & \text{se } x \leq 0 \\ \exp(-1/x) & \text{se } x > 0, \end{cases}$$

cujo gráfico é ilustrado na figura 3.

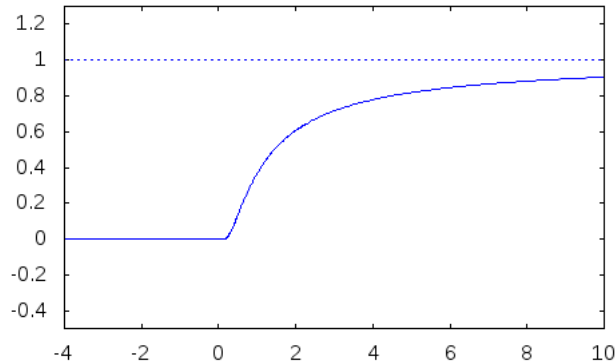


FIGURA 3. Gráfico da função $S(x)$.

Segue diretamente da definição de $S(x)$ que ela é infinitamente diferenciável em $(-\infty, 0)$ e $(0, \infty)$. Contudo, não é difícil mostrar que as derivadas (de qualquer ordem) de $\exp(-1/x)$ têm como limite 0, quando x tende a zero. Portanto, $S(x)$ tem derivadas de toda ordem na origem, o que mostra que é infinitamente diferenciável em toda a reta real. Contudo, $S^{(i)}(0) = 0$ implica que $\tau_0(x) = 0$. Como $S(x) \neq 0$ para todo número real positivo x , concluímos que S não pode ser analítica.

Exercícios

1. Em cada um dos itens abaixo são dados um número x e uma aproximação x_* de x . Determine, em cada caso, o erro absoluto, o erro relativo e a quantidade de algarismos corretos.
 - (a) $x = 123$ e $x_* = \frac{1106}{9}$;
 - (b) $x = 1/e$ e $x_* = 0.3666$;
 - (c) $x = 2^{10}$ e $x_* = 1000$;
 - (d) $x = 24$ e $x_* = 48$.
2. Para cada um dos números dados abaixo determine uma aproximação com erro absoluto 0.001 e uma aproximação com erro relativo 0.001.
 - (a) π ;
 - (b) $\sqrt{5}$;
 - (c) $\ln(3)$;
 - (d) $10/\ln(1.1)$.
3. Em cada item abaixo é dada uma aproximação x_* de um número x . Determine os possíveis valores de p quando o erro absoluto é 0.0005 e quando o erro relativo é 0.0005.
 - (a) $x_* = 0.2348263818643$;
 - (b) $x_* = 23.89627345677$;
 - (c) $x_* = -8.76257664363$.
4. Determine x e x_* , sabendo-se que x_* aproxima x com erro absoluto $1/100$ e erro relativo $3/100$.
5. Determine os valores de x e x_* sabendo-se que x_* é uma aproximação de x para a qual o erro relativo e o erro absoluto coincidem.
6. Considere a função $f(x) = \sin(\pi x/2)$ e seja $P_n(x)$ seu polinômio de Taylor de ordem n na origem. Determine n de modo que, para todo $0 \leq x \leq 2\pi$, o polinômio $P_n(x)$ seja uma aproximação de $f(x)$ para a qual as 6 primeiras casas decimais são corretas.
7. Seja $f : [-a, a] \rightarrow \mathbb{R}$ a função definida por $f(x) = e^x$.
 - (a) Calcule o polinômio de Taylor $P_2(x)$ de ordem dois de $f(x)$ na origem.
 - (b) Qual o maior valor de a para o qual podemos usar $P_2(x)$ para aproximar $f(x)$ com erro inferior a 0.005 em todo o intervalo $[-a, a]$?
8. Sejam $f(x) = x^{1/3}$ e $x_0 < x_0^*$ números reais positivos. Denote por E_x o erro relativo cometido ao usar x_0^* como aproximação de x_0 e por E_f o erro relativo cometido ao usar $f(x_0^*)$ como aproximação de $f(x_0)$.
 - (a) Determine uma cota superior para $E_f - E_x/3$ em função de x_0 e de E_x .

- (b) Qual a relação que E_x e x_0 devem satisfazer para que $E_x/3$ seja uma aproximação de E_f com erro inferior a 10^{-6} ?

9. A *função erro* é definida por

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt.$$

- (a) Calcule o polinômio de Taylor de ordem três de $\operatorname{erf}(x)$ na origem.
 (b) Determine $a > 0$ de modo que este polinômio aproxime o valor correto de $\operatorname{erf}(x)$ com erro inferior a 10^{-3} para todo $0 \leq x \leq a$?
10. Sejam a , b e e números reais não nulos.

- (a) Mostre, usando a desigualdade triangular, que

$$|a - b| \geq |a| - |b| \quad \text{e que} \quad |a - b| \geq |b| - |a|.$$

- (b) Mostre, usando (a) que

$$|a - b| \geq ||a| - |b|| = ||b| - |a||.$$

- (c) Mostre, usando (b), que se $||a| - |b|| \leq e$, então

$$-e \leq |a| - |b| \leq e.$$

11. Sejam a , b e e números reais não nulos tais que

$$\frac{|a - b|}{|b|} \leq e < 1.$$

Use o exercício 7 para provar as seguintes desigualdades

- (a)

$$1 - e \leq \frac{|a|}{|b|} \leq 1 + e.$$

- (b)

$$\frac{1}{1 + e} \leq \frac{|b|}{|a|} \leq \frac{1}{1 - e}.$$

- (c)

$$\frac{|a - b|}{|a|} \leq \frac{e}{1 - e}.$$

Sugestão para (c):

$$\frac{|a - b|}{|a|} = \frac{|b|}{|a|} \frac{|a - b|}{|b|}$$

e use (b) e a condição $|a - b|/|b| < e$.

12. Seja $e < 1$ um número real. Mostre, usando a soma de uma progressão geométrica, que

$$1 < \frac{1}{1-e} = 1 + e + e^2 + e^3 + \dots$$

e conclua disto que se $e < 0.5$, então

$$\frac{e}{1-e} \leq e + 2e^2.$$

Por exemplo, se $e < 10^{-1}$, então

$$\frac{1}{1-e} \leq 0.0102.$$

13. Seja a o valor exato e b o valor medido de uma determinada variável. Combine os exercícios 8 e 9 para mostrar que se

$$\frac{|a-b|}{|b|} \leq e < 1,$$

então o erro relativo correspondente é aproximadamente igual a e .

CAPÍTULO 3

O problema de valor de contorno

Neste capítulo veremos como utilizar a fórmula de Taylor para estimar o erro cometido ao aproximar a primeira e segunda derivadas de uma função pelas diferenças finitas introduzidas no capítulo 1. Esta análise nos permitirá obter de diferenças finitas melhores para estas derivadas. Além disso, generalizamos nosso estudo dos problemas de valores de contorno para cobrir quaisquer equações diferenciais lineares de ordem dois.

1. Recapitulando e generalizando

No capítulo 1 estudamos o problema de valor de contorno

$$y''(x) = c\rho(x), \quad \text{e} \quad y(0) = y(\ell) = a,$$

em que c , ℓ e a são números reais, obtido ao modelar a forma do cabo de sustentação de um ponte pênsil. Naturalmente, nem todos os problemas de valor de contorno são tão simples. Na verdade, os mais interessantes são modelados baseados em equações diferenciais parciais e estão fora do escopo deste livro. Contudo, podemos abordar, pelo método das diferenças finitas, problemas que envolvem equações diferenciais lineares de segunda ordem gerais, que são aquelas da forma

$$(46) \quad u''(x) = g_1(x)u'(x) + g_2(x)u(x) + g_3(x),$$

em que $g_1(x)$, $g_2(x)$ e $g_3(x)$ são funções contínuas definidas no intervalo $[a, b]$. Portanto, os problemas de valor de contorno que estudaremos neste capítulo consistem de uma equação (46) e dos valores de sua solução $u(x)$ nas extremidades a e b do intervalo de definição de $u(x)$.

Antes de prosseguir convém relembrar a estratégia adotada no método das diferenças finitas e que pode ser resumida nas seguintes etapas;

Etapla 1: discretizar o problema, substituindo as derivadas por expressões finitas;

Etapla 2: resolver o sistema linear resultante da discretização;

Etapla 3: plotar a curva poligonal que passa pelos pontos obtidos na etapa 2.

De acordo com as metas traçadas na página 25 para a primeira parte do livro, nosso objetivo neste capítulo consiste em *descobrir como aproximar $u''(x)$ cometendo um erro pequeno que não invalide o modelo que estamos resolvendo*. Para atingir este objetivo, usaremos a fórmula de Taylor com resto.

Digamos que $u : [a, b] \rightarrow \mathbb{R}$ seja uma função que tem derivadas contínuas até a quarta ordem. Vimos, no capítulo 1, que sua derivada em um ponto $x_0 \in (a, b)$ pode ser aproximada pelo quociente de Newton

$$(47) \quad \frac{u(x_0 + h) - u(x_0)}{h},$$

quando h é bastante pequeno. Contudo, a derivada de f em x_0 pode ser calculada também como o limite

$$u'(x_0) = \lim_{h \rightarrow 0} \frac{u(x_0) - u(x_0 - h)}{h};$$

de modo que podemos, igualmente, aproximá-la usando

$$(48) \quad \frac{u(x_0) - u(x_0 - h)}{h},$$

com h pequeno. Para podermos distinguir as duas formas do quociente de Newton, diremos que (47) é a *diferença posterior* e que (48) é a *diferença anterior* do quociente.

Talvez você esteja se perguntando se as duas diferenças não deveriam produzir sempre o mesmo resultado; afinal, a função tem uma única derivada, que pode ser calculada usando qualquer uma das duas diferenças. A resposta é que, no limite, as duas coincidirão, mas isto não acontece necessariamente para valores finitos de h . Considere, por exemplo, a função

$$\phi(x) = \begin{cases} \exp(-1/x) & \text{quando } x > 0 \\ 0 & \text{quando } x \leq 0, \end{cases}$$

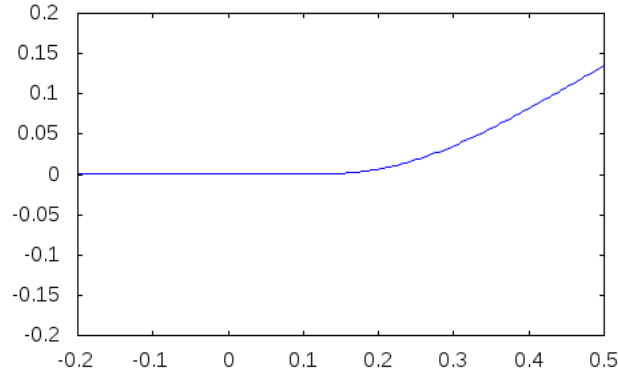
cujo gráfico é ilustrado na figura 1.

É claro que esta função tem derivadas de toda ordem fora da origem, e não é muito difícil mostrar que as derivadas na origem existem e são todas iguais a zero, *qualquer que seja a ordem*. Tomando $h = 0.1$, a diferença posterior da primeira derivada de $\phi(x)$ na origem é igual a

$$\frac{u(h) - u(0)}{h} = \frac{\exp(-1/h)}{h} = \frac{\exp(-10)}{0.1} = 0.000454,$$

ao passo que a diferença anterior é igual a

$$\frac{u(0) - u(-h)}{h} = 0,$$

FIGURA 1. Gráfico da função $\phi(x)$.

pois $u(-h) = u(0) = 0$. Portanto, a diferença anterior produz, neste exemplo, uma solução melhor que a posterior. Infelizmente isto não é sempre verdade; poderíamos igualmente ter escolhido uma função melhor para a diferença posterior que para a anterior.

Do ponto de vista das aplicações que desejamos fazer ao problema de valor de contorno, isto produz um impasse. Como decidir se é melhor usar a diferença anterior ou a posterior? A saída é não se comprometer com nenhuma das duas, usando a média entre elas,

$$\frac{1}{2} \left(\frac{u(x_0 + h) - u(x_0)}{h} + \frac{u(x_0) - u(x_0 - h)}{h} \right) = \frac{u(x_0 + h) - u(x_0 - h)}{2h}.$$

que é conhecida como *diferença centrada*. No caso da função $\phi(x)$, isto nos dá

$$\frac{\phi(h) - \phi(-h)}{2h} = \frac{\exp(-1/h) - 0}{2h} = 0.000455,$$

que é apenas ligeiramente melhor que o resultado da diferença posterior e claramente pior que a diferença anterior. Contudo, não é difícil dar um exemplo em que a diferença centrada ganha das outras duas; para isto, basta tomar $\psi(x) = x^2$. Neste caso, as diferenças anterior e posterior são, respectivamente,

$$\frac{\psi(h) - \psi(0)}{h} = \frac{h^2}{h} = h \quad \text{e} \quad \frac{\psi(0) - \psi(h)}{h} = \frac{-h^2}{h} = -h,$$

ao passo que a diferença centrada é igual a

$$\frac{\psi(h) - \psi(-h)}{2h} = \frac{h^2 - h^2}{2h} = 0,$$

que coincide com o valor exato da derivada. O problema é que a evidência destes exemplos é a mesma de um copo com água pela metade: alguns vão achar que é a favor das diferenças centradas, outros que é contra. A saída é provar, analiticamente,

que há fortes razões para esperar que as diferenças centradas são mesmo melhores que as diferenças posterior e anterior.

2. Aproximando derivadas

Para provar que a diferença centrada é, em geral, melhor que a diferença posterior, basta analisar o erro cometido em cada caso e eleger a fórmula para a qual o erro decai mais rapidamente com h . No caso da diferença posterior este erro é igual ao módulo de

$$e_0 = \frac{u(x_0 + h) - u(x_0)}{h} - u'(x_0)$$

Porém, o polinômio de Taylor de grau um de $u(x)$ em x_0 é

$$P_1(x) = u(x_0) + u'(x_0)(x - x_0),$$

o que nos permite reescrever e_0 na forma

$$e_0 = \frac{u(x_0 + h) - P_1(x_0 + h)}{h}.$$

Contudo, pelo teorema de Taylor,

$$|u(x_0 + h) - P_1(x_0 + h)| \leq \frac{M_0 h^2}{2},$$

em que M_0 é o máximo de $u''(x)$ no intervalo $[a, b]$. Mas isto implica que,

$$|e_0| = \frac{|u(x_0 + h) - P_1(x_0 + h)|}{h} \leq \frac{M_0 h}{2};$$

mostrando que o erro cometido decai linearmente em h , quando usamos diferenças posteriores para aproximar a primeira derivada. A análise referente à diferença anterior é análoga e vamos deixá-la por sua conta escrever os detalhes.

Como já sabemos que a fórmula de Taylor é o instrumento adequado ao estudo do erro, começaremos nossa análise da diferença centrada aproximando $u(x_0 + h)$ e $u(x_0 - h)$ pelo polinômio de Taylor de grau dois, o que nos dá

$$u(x_0 + h) = u(x_0) + u'(x_0)h + \frac{u''(x_0)h^2}{2} + e_1,$$

no primeiro caso, e

$$u(x_0 - h) = u(x_0) - u'(x_0)h + \frac{u''(x_0)h^2}{2} + e_2,$$

no segundo. Os módulos dos números e_1 e e_2 representam os erros cometidos nas aproximações de cada uma das fórmulas. Subtraindo uma fórmula da outra e cancelando os termos comuns,

$$u(x_0 + h) - u(x_0 - h) = 2u'(x_0)h + e_1 - e_2;$$

donde

$$\frac{u(x_0 + h) - u(x_0 - h)}{2h} - u'(x_0) = \frac{e_1 - e_2}{2h}.$$

Logo, o erro cometido quando usamos a diferença centrada para aproximar a derivada é igual a

$$\frac{|e_1 - e_2|}{2h}.$$

Mas, pela desigualdade triangular,

$$(49) \quad \frac{|e_1 - e_2|}{2h} \leq \frac{|e_1| + |e_2|}{2h}.$$

Denotando por M_1 e M_2 os máximos de $f'''(x)$ nos intervalos $[x_0, x_0 + h]$ e $[x_0 - h, x_0]$, respectivamente, o teorema de Taylor nos permite afirmar que

$$|e_1| \leq \frac{M_1 h^3}{6} \quad \text{e que} \quad |e_2| \leq \frac{M_2 h^3}{6}.$$

Substituindo estas desigualdades em (49),

$$\frac{|e_1 - e_2|}{2h} \leq \frac{|e_1| + |e_2|}{2h} \leq \frac{(M_1 + M_2)h^2}{12}.$$

Portanto, ao passo que os erros referentes às diferenças posteriores variam linearmente, aqueles referentes às diferenças centradas variam quadraticamente. Grosso modo, isto significa que, quando dividimos h por 10, o primeiro erro é dividido por 10, ao passo que o segundo é dividido por 100, o que explica porque as diferenças centradas merecem à boa fama que têm.

◈ Se você reler o primeiro exemplo que fizemos, no início de seção, verá que o erro referente à diferença anterior havia dado igual a zero, ao passo que o erro referente à diferença centrada havia sido positivo. Mas, como isto é possível, se o primeiro erro é linear e o outro é quadrático? A resposta é que a função ϕ tem derivadas nulas à esquerda da origem, de modo que o valor máximo de suas derivadas em $(-\infty, 0]$ é igual a zero. Portanto, existem situações em que, a despeito do que provamos acima, o erro associado à diferença anterior (ou posterior) pode ser menor que o que corresponde à diferença centrada.

Como as equações diferenciais que nos interessam são de segunda ordem, ainda precisamos achar a expressão para a diferença centrada da segunda derivada em x_0 da função f com que começamos esta seção. Como no caso da derivada primeira, nosso ponto de partida são as fórmulas de Taylor, desta vez de grau três, de $u(x_0 + h)$ e de $u(x_0 - h)$; isto é,

$$u(x_0 + h) = u(x_0) + u'(x_0)h + \frac{u''(x_0)h^2}{2} + \frac{f'''(x_0)h^3}{3!} + e_1,$$

e

$$u(x_0 - h) = u(x_0) - u'(x_0)h + \frac{u''(x_0)h^2}{2} - \frac{f'''(x_0)h^3}{3!} + e_2,$$

em que, como antes, os módulos de e_1 e e_2 denotam os respectivos erros. Somando as duas fórmulas,

$$u(x_0 + h) + u(x_0 - h) = 2u(x_0) + u''(x_0)h^2 + (e_1 + e_2);$$

donde

$$\frac{u(x_0 + h) + u(x_0 - h) - 2u(x_0)}{h^2} - u''(x_0) = \frac{e_1 + e_2}{h^2}.$$

Aplicando a desigualdade triangular, seguida da estimativa do erro do teorema de Taylor, obtemos

$$\left| \frac{e_1 + e_2}{h^2} \right| \leq \frac{|e_1| + |e_2|}{4!} \leq \frac{(M_1 + M_2)h^2}{4!},$$

em que M_1 e M_2 são os máximos da quarta derivada de f nos intervalos $[x_0, x_0 + h]$ e $[x_0 - h, x_0]$. Portanto, a diferença centrada

$$\frac{u(x_0 + h) + u(x_0 - h) - 2u(x_0)}{h^2},$$

é uma aproximação de $u''(x_0)$ com erro quadrático. Resumindo, mostramos nesta seção que, se h for suficientemente pequeno, então

$$u'(x_0) \approx \frac{u(x_0 + h) - u(x_0 - h)}{2h} \quad \text{e} \quad u''(x_0) \approx \frac{u(x_0 + h) - 2u(x_0) + u(x_0 - h)}{h^2}.$$

Resta-nos descobrir como estas fórmulas podem ser usadas para calcular aproximações para o tipo de problema de valor de contorno que nos propomos a resolver numericamente.

3. O método de diferenças finitas

Começaremos a seção usando diferenças centradas para resolver o problema de valor de contorno definido pela equação diferencial que define a forma do cabo de sustentação de uma ponte pênsil com carga uniforme

$$u'' = 2 \quad \text{e} \quad u(-1) = u(1) = 1.$$

Mais uma vez, nosso objetivo ao resolver este problema é que nos permite comparar facilmente a solução numérica obtida através do método de diferenças finitas com a solução analítica $u(x) = x^2$, que é exata.

A primeira coisa a fazer é escolher a quantidade n de partes em que o intervalo de definição da solução deve ser dividido. Neste exemplo, o intervalo é $[-1, 1]$ e, como não queremos nos comprometer com nenhum valor específico para n , deixaremos este valor indeterminado, pelo menos por enquanto. Como o intervalo de definição da função tem comprimento 2, cada um dos n segmentos em que $[-1, 1]$ será dividido tem tamanho $h = 2/n$. Se $x_i = -1 + ih$, então estes segmentos podem ser escritos na forma $[x_i, x_{i+1}]$, com $i = 0, \dots, n-1$. Observe que, com esta notação, $x_0 = -1$ e

$x_n = 1$. Usando a fórmula de diferenças centradas da seção anterior para aproximar $u''(x_i)$, temos que

$$u''(x_i) \approx \frac{u(x_i - h) - 2u(x_i) + u(x_i + h)}{h^2} = \frac{u(x_{i-1}) - 2u(x_i) + u(x_{i+1}))}{h^2}.$$

Para simplificar um pouco a notação, denotaremos $u(x_i)$ por y_i , o que nos permite reescrever a fórmula anterior na forma

$$u''(x_i) \approx \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2}.$$

Substituindo esta aproximação de $u''(x_i)$ na equação $u'' = 2$, obtemos

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} = 2;$$

isto é,

$$y_{i-1} - 2y_i + y_{i+1} = 2h^2.$$

Fazendo i variar entre 1 e $n - 1$, encontramos o sistema

$$\begin{aligned} y_0 - 2y_1 + y_2 &= 2h^2 \\ y_1 - 2y_2 + y_3 &= 2h^2 \\ &\vdots \\ y_{n-2} - 2y_{n-1} + y_n &= 2h^2 \end{aligned}$$

Como

$$y_0 = u(x_0) = u(-1) = 1 \quad \text{e} \quad y_n = u(x_n) = u(1) = 1$$

a primeira e última equações do sistema se simplificam como

$$-2y_1 + y_2 = 2h^2 - 1 \quad \text{e} \quad y_{n-2} - 2y_{n-1} = 2h^2 - 1.$$

Com isto, o sistema final tem a forma

$$\begin{aligned} -2y_1 + y_2 &= 2h^2 - 1 \\ y_1 - 2y_2 + y_3 &= 2h^2 \\ &\vdots \\ y_{n-2} - 2y_{n-1} &= 2h^2 - 1. \end{aligned}$$

Para que possamos efetuar os cálculos até o fim, precisamos escolher um valor numérico para n . Digamos que $n = 4$. Neste caso $h = 2/4 = 0.5$ e o sistema resume-se às equações

$$\begin{aligned} -2y_1 + y_2 &= -0.5 \\ y_1 - 2y_2 + y_3 &= 0.5 \\ y_2 - 2y_3 &= -0.5. \end{aligned}$$

A matriz aumentada correspondente é

$$\left[\begin{array}{ccc|c} -2 & 1 & 0 & -0.5 \\ 1 & -2 & 1 & 0.5 \\ 0 & 1 & -2 & -0.5 \end{array} \right]$$

Trocando as duas primeiras linhas de posição e aplicando eliminação gaussiana à primeira coluna, obtemos

$$\left[\begin{array}{ccc|c} 1 & -2 & 1 & 0.5 \\ 0 & -3 & 2 & 0.5 \\ 0 & 1 & -2 & -0.5 \end{array} \right]$$

Finalmente, trocando as duas últimas linhas de posição e aplicando eliminação à segunda coluna, resta a matriz

$$\left[\begin{array}{ccc|c} 1 & -2 & 1 & 0.5 \\ 0 & 1 & -2 & -0.5 \\ 0 & 0 & -4 & -1 \end{array} \right]$$

que equivale ao sistema

$$\begin{aligned} y_1 - 2y_2 + y_3 &= 0.5 \\ y_2 - 2y_3 &= -0.5 \\ -4y_3 &= -1 \end{aligned}$$

cujas soluções são

$$y_1 = 0.25, \quad y_2 = 0.0, \quad \text{e} \quad y_3 = 0.25,$$

que, neste caso, correspondem aos valores exatos de $y(x)$ em $x_1 = -0.5$, $x_2 = 0$ e $x_3 = 0.5$, respectivamente.

Refazendo os cálculos para $n = 8$, obtemos o sistema

$$\begin{aligned} -2y_1 + y_2 &= -0.5 & y_4 - 2y_5 + y_6 &= -0.125 \\ y_1 - 2y_2 + y_3 &= 0.875 & y_5 - 2y_6 + y_7 &= -0.125 \\ y_2 - 2y_3 + y_4 &= -0.125 & y_6 - 2y_7 &= 0.875 \\ y_3 - 2y_4 + y_5 &= -0.125. \end{aligned}$$

cujas matrizes aumentadas são

$$\left[\begin{array}{cccccc|c} 2 & -1 & 0 & 0 & 0 & 0 & 0.875 \\ -1 & 2 & -1 & 0 & 0 & 0 & -0.125 \\ 0 & -1 & 2 & -1 & 0 & 0 & -0.125 \\ 0 & 0 & -1 & 2 & -1 & 0 & -0.125 \\ 0 & 0 & 0 & -1 & 2 & -1 & -0.125 \\ 0 & 0 & 0 & 0 & -1 & 2 & -0.125 \\ 0 & 0 & 0 & 0 & 0 & -1 & 2 & 0.875 \end{array} \right].$$

Usando eliminação gaussiana, obtemos a solução

$$\begin{aligned} y_1 &= 0.5625, & y_2 &= 0.2500, & y_3 &= 0.0625, & y_4 &= -3.331 \cdot 10^{-17}, \\ y_5 &= 0.0625, & y_6 &= 0.25, & y_7 &= 0.5625. \end{aligned}$$

Desta vez os pontos, claramente, não caem exatamente sobre a parábola. A aproximação poligonal obtida ligando estes pontos por segmentos de retas é ilustrada na figura 2, juntamente com o gráfico da solução exata.

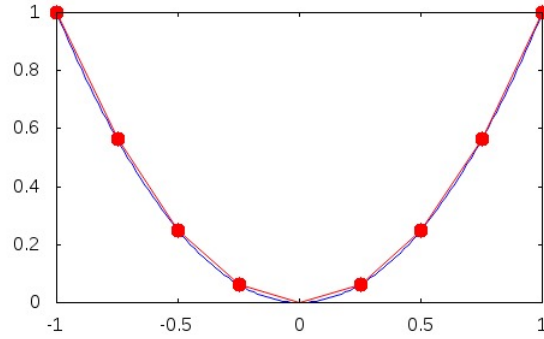


FIGURA 2. A curva azul é $y = x^2$, a vermelha é a solução aproximada de $u'' = 2$ com $n = 8$.

Nosso segundo exemplo consiste na equação

$$(50) \quad u'' = -u' + x^2,$$

sujeita às condições de contorno

$$u(0) = 2 \quad \text{e} \quad u(3) = 26.$$

Substituindo as aproximações

$$u'(x_i) \approx \frac{-y_{i-1} + y_{i+1}}{2h} \quad \text{e} \quad u''(x_i) \approx \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2}.$$

na equação, obtemos

$$2(y_{i-1} - 2y_i + y_{i+1}) + h(-y_{i-1} + y_{i+1}) = 2h^2 x_i^2$$

com $1 \leq i \leq n-1$. Reagrupando os termos,

$$(2-h)y_{i-1} - 4y_i + (2+h)y_{i+1} = 2h^2 x_i^2 \quad \text{com } 1 \leq i \leq n-1.$$

Finalmente, substituindo os valores de $y_0 = u(0) = 2$ e $y_n = u(3) = 26$, obtemos o sistema

$$\begin{aligned} -4y_1 + (2+h)y_2 &= -4 + 2h + 2h^2x_1^2 \\ (2-h)y_1 - 4y_2 + (2+h)y_3 &= 2h^2x_2^2 \\ &\vdots \\ (2-h)y_{n-2} - 4y_{n-1} &= 2h^2x_{n-1}^2 - 26(2+h). \end{aligned}$$

Quando $n = 8$, temos que $h = 3/8$, de modo que o sistema é

$$\begin{aligned} -4y_1 + (19/8)y_2 &= -6575/2048 \\ (13/8)y_1 - 4y_2 + (19/8)y_3 &= 81/512 \\ (13/8)y_2 - 4y_3 + (19/8)y_4 &= 729/2048 \\ (13/8)y_3 - 4y_4 + (19/8)y_5 &= 81/128 \\ (13/8)y_4 - 4y_5 + (19/8)y_6 &= 2025/2048 \\ (13/8)y_5 - 4y_6 + (19/8)y_7 &= 729/512 \\ (13/8)y_6 - 4y_7 &= -122495/2048, \end{aligned}$$

que tem matriz aumentada

$$\left[\begin{array}{ccccccc|c} -4 & \frac{19}{8} & 0 & 0 & 0 & 0 & 0 & -\frac{6575}{2048} \\ \frac{13}{8} & -4 & \frac{19}{8} & 0 & 0 & 0 & 0 & \frac{81}{512} \\ 0 & \frac{13}{8} & -4 & \frac{19}{8} & 0 & 0 & 0 & \frac{729}{2048} \\ 0 & 0 & \frac{13}{8} & -4 & \frac{19}{8} & 0 & 0 & \frac{81}{128} \\ 0 & 0 & 0 & \frac{13}{8} & -4 & \frac{19}{8} & 0 & \frac{2025}{2048} \\ 0 & 0 & 0 & 0 & \frac{13}{8} & -4 & \frac{19}{8} & \frac{729}{512} \\ 0 & 0 & 0 & 0 & 0 & \frac{13}{8} & -4 & -\frac{122495}{2048} \end{array} \right]$$

e soluções

$$\begin{aligned} y_2 &= 13.178, & y_1 &= 8.627, & y_3 &= 16.358, & y_4 &= 18.684, \\ y_5 &= 20.542, & y_6 &= 22.23, & y_7 &= 23.984 \end{aligned}$$

Como você pode constatar olhando a figura 3, a curva correta (azul) coincide a tal ponto com a curva poligonal obtida ligando os pontos calculados numericamente, que é difícil distinguir uma da outra. O algoritmo usado para resolver estes dois exemplos pode ser enunciado da seguinte maneira.

MÉTODO DAS DIFERENÇAS FINITAS. *Dado o problema de valor de contorno*

$$(51) \quad u'' = g_1u' + g_2y + g_3, \quad \text{em que} \quad u(a) = \alpha \quad \text{e} \quad u(b) = \beta,$$

e um inteiro positivo n , o algoritmo retorna uma curva poligonal que é uma aproximação da solução de (51).

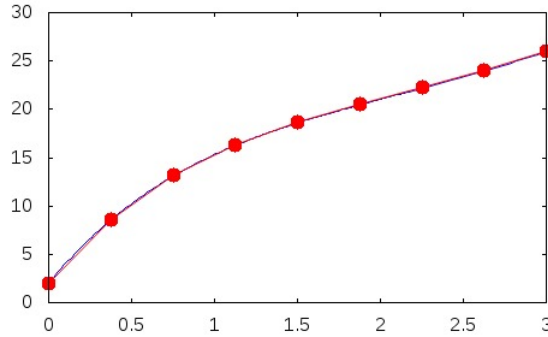


FIGURA 3. A curva azul é solução exata de $u'' = u' + x^2$, a vermelha é a solução numérica de com $n = 8$.

Etapa 1: calcule $h = (b - a)/n$;

Etapa 2: liste as diferenças centradas $(-y_{i-1} + y_{i+1})/2h$, e $(y_{i-1} - 2y_i + y_{i+1})/h^2$, para $i = 1, \dots, n-1$;

Etapa 3: escreva as equações obtidas substituindo as diferenças centradas listadas acima e x por $x_i = a + ih$ na equação diferencial, para $i = 1, \dots, n-1$;

Etapa 4: substitua y_0 por α e y_n por β no sistema da etapa 4;

Etapa 5: resolva o sistema linear da etapa 5;

Etapa 6: gere a lista \mathcal{P} cujos elementos são os pontos (x_i, y_i) , para todo $i = 1, \dots, n-1$, e acrescente $[a, \alpha]$ no início e $[b, \beta]$ ao final de \mathcal{P} ;

Etapa 7: retorne a curva poligonal que passa pelos pontos de \mathcal{P} .

Encerraremos com uma observação importante sobre o sistema linear obtido na etapa 5 do algoritmo. Se você comparar as três matrizes obtidas nos exemplos, verá que todas as suas entradas não nulas pertencem à sua diagonal ou às duas sub-diagonais imediatamente acima e abaixo dela. Estas matrizes são conhecidas como *tridiagonais* e é claro que são fáceis de simplificar, usando eliminação gaussiana, porque a maior parte das suas entradas já é igual a zero. A observação importante é que isto vale para qualquer matriz obtida a partir do método de diferenças finitas, porque as diferenças centradas usadas para aproximar as duas derivadas da i -ésima equação dependem apenas das variáveis y_{i-1} , y_i e y_{i+1} . Com isso, as únicas entradas não nulas de cada linha correspondem aos coeficientes destas três variáveis, que estão dispostos simetricamente em relação à diagonal da matriz.

Exercícios

1. Considere a função $f(x)$ cujos valores são dados na tabela abaixo:

x	0	0.1	0.2	0.3	0.4
$f(x)$	0.0000	0.0819	0.1341	0.1646	0.1797

Calcule $f'(x)$ e $f''(x)$ em $x = 0.1$ e $x = 0.3$ usando diferenças centradas.

2. Resolva os seguintes problemas de valores de contorno no intervalo $[0, 1]$ usando o método das diferenças finitas com o passo $h = 0.2$:
- (a) $u'' + u = 0$ com $u(0) = 0$ e $u(1) = 1$;
 - (b) $u'' = 4(u - x)$ com $u(0) = 0$ e $u(1) = 2$;
 - (c) $u'' + 2u' + u = x$ com $u(0) = 2$ e $u(1) = 0$;
 - (d) $u'' = -3u' + 2u + 2x + 3$, com $u(0) = 2$ e $u(1) = 1$;
 - (e) $u'' = -(x + 1)u' + 2u + (1 - x^2)e^{-x}$ com $u(0) = -1$ e $u(1) = 0$.

3. Considere o problema de valores de contorno no intervalo $[0, \pi/2]$ dado por

$$u'' + u = 0, \quad \text{com} \quad u(0) = 0 \quad \text{e} \quad u(\pi/2) = 2.$$

- (a) Resolva este problema pelo método das diferenças finitas com $h = \pi/8$.
- (b) Ache a solução exata deste problema usando o MAXIMA, ou outro sistema de computação algébrica, e determine os erros cometidos em cada ponto da malha.

4. Considere o problema de valores de contorno no intervalo $[0, \pi/2]$ dado por

$$u'' = u' + 2u + \cos(x), \quad \text{com} \quad u(0) = -0.3 \quad \text{e} \quad u\left(\frac{\pi}{2}\right) = -0.1.$$

- (a) Resolva este problema pelo método das diferenças finitas com $h = \pi/4$.
- (b) Ache a solução exata deste problema usando o MAXIMA, ou outro sistema de computação algébrica, e determine os erros cometidos em cada ponto da malha.

5. A temperatura $T(x)$ de uma barra uniformemente aquecida por uma fonte de calor é dada por

$$T''(x) = -f(x).$$

Supondo que $f(x) = 25^\circ C$ e que as temperaturas nas extremidades da barra são $T(0) = 40^\circ C$ e $T(8) = 200^\circ C$, esboce, usando splines lineares, a curva que descreve a distribuição de calor nesta barra quando tomamos $h = 2$.

CAPÍTULO 4

Decomposição de matrizes

Embora tenhamos usado matrizes para resolver sistemas lineares no capítulo 1, as matrizes serviram apenas para deixar mais claras os cálculos, porque nos permitiram escrever apenas os coeficientes das equações, sem incógnitas e sem os símbolos para as operações. Assim, não somamos ou multiplicamos quaisquer matrizes, apenas operamos com suas linhas, consideradas como uma maneira simplificada de representar as equações de um sistema linear. Entretanto, Alan Turing mostrou, em um artigo publicado em 1948, que é possível interpretar o método de eliminação como uma decomposição de matrizes. Mais precisamente, uma matriz quadrada pode ser escrita, a menos de troca de linhas, como o produto de uma matriz triangular inferior por uma matriz triangular superior. Para usar esta interpretação da eliminação na solução de sistemas lineares precisaremos reinterpretá-los como equações matriciais. Uma das vantagens deste enfoque é que nos permite resolver facilmente vários sistemas lineares que diferem apenas em seus termos constantes. Isto ocorre, por exemplo, quando precisamos resolver problemas de valor de contorno referentes a uma mesma equação diferencial, mas com condições de contorno distintas. Suporemos, ao longo de todo este capítulo, que os sistemas lineares de que estamos tratando são determinados; isto é, têm uma única solução.

1. Matrizes e eliminação

A partir desta seção vamos nos referir ao algoritmo de eliminação gaussiana da seção 3 do capítulo 1 como *clássico*, em oposição à versão moderna, que introduziremos agora. Começaremos nossa análise com a pergunta: é possível efetuar uma operação por linha, em uma dada matriz A , simplesmente multiplicando A por alguma outra matriz? A resposta é, predizivelmente, sim—ou a pergunta nem teria sido feita.

Vejamos o que acontece quando A é uma matriz 3×3 . Digamos que a operação elementar a ser aplicada a A consiste em substituir sua segunda linha por ela própria mais k_1 vezes a primeira linha, em que k_1 é um número real. Supondo que

$$A = \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{bmatrix}$$

nossa estratégia consiste em multiplicar A à esquerda por uma matriz L_1 , a ser determinada, de modo que

$$LA = \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 + k_1 a_1 & b_2 + k_1 a_2 & b_3 + k_1 a_3 \\ c_1 & c_2 & c_3 \end{bmatrix}.$$

Logo, L deve ser construída de tal maneira que o produto de sua primeira linha por A apenas reproduza a primeira linha de A . De maneira semelhante, do produto da terceira linha de L por A deve resultar apenas a última linha de A . Para que isto ocorra, basta que a primeira e a terceira linhas de L sejam iguais às linhas correspondentes da matriz identidade 3×3 . Em outras palavras,

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ ? & ? & ? \\ 0 & 0 & 1 \end{bmatrix}}_{L_1} \cdot \underbrace{\begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{bmatrix}}_A = \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 + k_1 a_1 & b_2 + k_1 a_2 & b_3 + k_1 a_3 \\ c_1 & c_2 & c_3 \end{bmatrix}.$$

Resta-nos decidir o que pôr no lugar das interrogações. Para que não apareça nenhum c na segunda linha, a terceira interrogação tem que ser igual a 0. De maneira análoga, para que apareçam bs , a segunda interrogação deve se igual a um, ao passo que os produtos dos as por k_1 , requerem que a primeira interrogação seja igual a k_1 . Temos, então, que

$$L = \begin{bmatrix} 1 & 0 & 0 \\ k_1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

De fato,

$$\begin{bmatrix} 1 & 0 & 0 \\ k_1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 + k_1 a_1 & b_2 + k_1 a_2 & b_3 + k_1 a_3 \\ c_1 & c_2 & c_3 \end{bmatrix},$$

como desejávamos. Um argumento semelhante mostra que a operação elementar que consiste em substituir a terceira linha de A por sua soma com k_2 vezes a primeira linha é obtida multiplicando-a à esquerda pela matriz

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ k_2 & 0 & 1 \end{bmatrix}.$$

Finalmente, para substituir a terceira linha de A por sua soma com k_3 vezes a segunda linha basta multiplicar A à esquerda por

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & k_3 & 1 \end{bmatrix}.$$

Uma coisa que você deve ter observado é que as três matrizes encontradas acima diferem da matriz identidade 3×3 em apenas uma posição, que é aquela onde está o k_i . Isto significa que estas matrizes podem ser caracterizadas completamente pelo número real que determina o valor e pelos dois inteiros positivos que determinam a posição da entrada em que a matriz difere da identidade. Tendo isto em vista, denotaremos por $L_{i,j}(c)$ a matriz que é igual à identidade 3×3 exceto pela posição ij que é igual a c . Assim,

$$L_{21}(k_1) = \begin{bmatrix} 1 & 0 & 0 \\ k_1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad L_{31}(k_2) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ k_2 & 0 & 1 \end{bmatrix} \quad \text{e} \quad L_{32}(k_3) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & k_3 & 1 \end{bmatrix}.$$

As matrizes $L_{i,j}(k)$ são conhecidas como *matrizes elementares*, porque será através delas que executaremos as operações elementares por linhas.

Os argumentos acima nos permitem apenas determinar a *forma* das matrizes elementares cujos produtos (sempre à esquerda!) nos permitem executar as operações por linhas requeridas na eliminação gaussiana. Mas, neste procedimento, nosso real objetivo ao calcular $L_{2,1}(k_1)A$ é anular a entrada 2, 1 da matriz. Para que isto aconteça, devemos escolher k_1 de modo que $b_1 + k_1 a_1 = 0$; o que nos dá

$$k_1 = -\frac{b_1}{a_1},$$

desde que $a_1 \neq 0$. Isto, naturalmente, não é novidade, porque o mesmo valia quando calculávamos a eliminação através do algoritmo clássico.

Um exemplo numérico vai ajudá-lo a entender melhor como efetuar eliminação usando matrizes elementares. Digamos que

$$A = \begin{bmatrix} 1 & 3 & 2 \\ -2 & -4 & 1 \\ 2 & 2 & 2 \end{bmatrix}$$

Para que a posição 2, 1 do produto $L_1 A$ seja nula, devemos escolher $k_1 = 2$, de modo que

$$L_{2,1}(2) = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{e} \quad L_{2,1}(2) \cdot A = \begin{bmatrix} 1 & 3 & 2 \\ 0 & 2 & 5 \\ 2 & 2 & 2 \end{bmatrix}$$

Escolhendo, agora, $k_2 = -2$, teremos que

$$L_{3,1}(-2) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} \quad \text{e} \quad L_{3,1}(-2) \cdot L_{2,1}(2)A = \begin{bmatrix} 1 & 3 & 2 \\ 0 & 2 & 5 \\ 0 & -4 & -2 \end{bmatrix}$$

Finalmente, tomando $k_3 = 4/2 = 2$,

$$L_{3,2}(2) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix} \quad \text{e} \quad L_{3,2}(2) \cdot L_{3,1}(-2) \cdot L_{2,1}(2)A = \begin{bmatrix} 1 & 3 & 2 \\ 0 & 2 & 5 \\ 0 & 0 & 8 \end{bmatrix},$$

que conclui o processo de eliminação.

2. Matrizes elementares e decomposição LU

A definição de matriz elementar da seção anterior pode ser facilmente generalizada. Diremos que uma matriz $n \times n$ é *elementar* se difere da matriz identidade apenas em uma posição fora de diagonal. Como no caso em que $n = 3$, denotaremos por $L_{i,j}(k)$ a matriz elementar cuja entrada ij é igual a k ; naturalmente isto pressupõe que $i \neq j$, ou a entrada estaria sobre a diagonal. Observe que a notação que estamos usando não especifica o tamanho da matriz; isto é, $L_{21}(1)$ pode denotar tanto a matriz

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \quad \text{quanto a matriz} \quad \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

dependendo do contexto em que estas matrizes aparecerem. Isto não causa nenhuma ambiguidade, porque as matrizes elementares têm que ter o mesmo tamanho da matriz à qual a eliminação está sendo aplicada.

Há uma outra maneira de escrever as matrizes elementares que é muito útil quando precisamos calcular com estas matrizes. Para obtê-la usaremos as matrizes $E_{i,j}$, que diferem da matriz nula apenas na posição ij , que é igual a 1. Por exemplo, quando $n = 3$,

$$E_{12} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad E_{23} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{e} \quad E_{32} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Denotando por I a matriz identidade $n \times n$ e supondo que $i \neq j$ são dois inteiros entre 1 e n , podemos escrever

$$L_{i,j}(k) = I + kE_{i,j}.$$

A grande vantagem de expressar as matrizes elementares em termos das matrizes $E_{i,j}$ é que isto torna os cálculos mais fáceis. De fato, como cada matriz $E_{i,j}$ tem apenas uma posição não nula, o produto de $E_{i,j}$ por E_{rs} só não dá zero quando o número 1 em $E_{i,j}$ é multiplicado pelo número 1 em E_{rs} . Levando em conta que o produto de matrizes é feito linha por coluna, isto acontece apenas quando o 1, que

aparece na coluna j de $E_{i,j}$, está na linha j de E_{rs} ; o que só é possível quando $r = j$. Resumindo,

$$(52) \quad E_{i,j} \cdot E_{rs} = \begin{cases} E_{is} & \text{quando } r = j \\ 0 & \text{quando } r \neq j. \end{cases}$$

Vejamos como usar isto para calcular

$$L_{i,j}(k_1)L_{rs}(k_2) = (I + k_1E_{i,j})(I + k_2E_{rs}),$$

sem esquecer que $i \neq j$ e $r \neq s$ são requisitos para que $L_{i,j}(k_1)$ e $L_{rs}(k_2)$ possam ser definidas. Usando a distributividade da multiplicação de matrizes

$$(I + k_1E_{i,j})(I + k_2E_{rs}) = I + k_1E_{i,j} + k_2E_{rs} + k_1k_2E_{i,j}E_{rs}.$$

Por (52) segue-se, então, que

$$(53) \quad L_{i,j}(k_1)L_{rs}(k_2) = \begin{cases} I + k_1E_{i,j} + k_2E_{rs} + k_1k_2E_{is} & \text{quando } r = j \\ I + k_1E_{i,j} + k_2E_{rs} & \text{quando } r \neq j. \end{cases}$$

Em particular, como $i \neq j$, então

$$L_{i,j}(k_1)L_{i,j}(k_2) = I + k_1E_{i,j} + k_2E_{i,j} = I + (k_1 + k_2)E_{i,j}.$$

Assim, quando $k_2 = -k_1$,

$$(54) \quad L_{i,j}(k_1)L_{i,j}(-k_1) = I + (k_1 - k_1)E_{i,j} = I;$$

de modo que $L_{i,j}(-k_1)$ é a matriz inversa de $L_{i,j}(k_1)$.

Voltando ao exemplo da seção anterior, vimos que

$$(55) \quad L_{32}(-2)L_{31}(-2)L_{21}(2) \cdot A = U,$$

em que U corresponde à matriz triangular superior

$$\begin{bmatrix} 1 & 3 & 2 \\ 0 & 2 & 5 \\ 0 & 0 & 8 \end{bmatrix}.$$

Mas sabemos de (54) que $L_{32}(2)$ é inversa de $L_{32}(-2)$. Isto significa que, multiplicando (55) à esquerda por $L_{32}(2)$, obtemos

$$L_{31}(-2)L_{21}(2) \cdot A = L_{32}(2) \cdot U,$$

pois $L_{32}(-2) \cdot L_{32}(2)$ é igual à matriz identidade 3×3 . Procedendo de maneira análoga para as outras duas matrizes elementares, concluímos que

$$(56) \quad A = L_{21}(2)L_{31}(2)L_{32}(2) \cdot U.$$

Usando a fórmula (53), é fácil calcular explicitamente o produto

$$L_{21}(2)L_{31}(2)L_{32}(-2).$$

De fato, como $1 \neq 3$,

$$L_{31}(2)L_{32}(-2) = I + 2E_{31} + (-2)E_{32}.$$

Contudo,

$$L_{21}(2)L_{31}(2)L_{32}(-2) = (I + 2E_{21})(I + 2E_{31} + (-2)E_{32}),$$

que, pela distributividade da multiplicação de matrizes, é igual a

$$I + (-2)E_{32} + 2E_{31} + 2E_{21} + 4E_{21}E_{31} + (-4)E_{21}E_{32}.$$

Finalmente, de (52) temos que

$$E_{21}E_{31} = E_{21}E_{32} = 0;$$

donde

$$L_{21}(2)L_{31}(2)L_{32}(-2) = I + 2E_{31} + (-2)E_{32} + 2E_{21} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & -2 & 1 \end{bmatrix}.$$

Denotando esta última matriz por L , podemos escrever (56) compactamente na forma

$$A = L \cdot U.$$

Diremos que $L \cdot U$ é a *decomposição LU* da matriz A . A importância da decomposição LU para a solução de sistemas está no fato de que U é uma matriz triangular superior (em inglês *Upper triangular*), ao passo que L é triangular inferior (em inglês *Lower triangular*). Antes de explicar em que isto nos ajuda, convém fazer um segundo exemplo.

Vejamos como nos saímos se aplicarmos o que aprendemos até aqui a uma matriz 4×4 . Por exemplo, digamos que

$$A = \begin{bmatrix} 1 & 2 & 3 & -1 \\ 3 & 7 & 9 & 0 \\ -1 & 1 & -1 & 7 \\ 1 & 4 & -1 & 16 \end{bmatrix}$$

Usando eliminação com pivô em 1, 1 e escrevendo as operações por linha em termos das matrizes elementares correspondentes, obtemos

$$L_{41}(-1)L_{31}(1)L_{21}(-3)A = \begin{bmatrix} 1 & 2 & 3 & -1 \\ 0 & 1 & 0 & 3 \\ 0 & 3 & 2 & 6 \\ 0 & 2 & -4 & 17 \end{bmatrix}.$$

Tomando, agora, a posição 2, 2 como pivô, teremos

$$L_{42}(-2)L_{32}(-3) \cdot L_{41}(-1)L_{31}(1)L_{21}(-3)A = \begin{bmatrix} 1 & 2 & 3 & -1 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 2 & -3 \\ 0 & 0 & -4 & 11 \end{bmatrix}.$$

Finalmente, quando o pivô está em 3, 3,

$$L_{43}(2) \cdot L_{42}(-2)L_{32}(-3) \cdot L_{41}(-1)L_{31}(1)L_{21}(-3)A = \begin{bmatrix} 1 & 2 & 3 & -1 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 2 & -3 \\ 0 & 0 & 0 & 5 \end{bmatrix}.$$

Denotando esta última matriz por U , mostramos que

$$L_{43}(2) \cdot L_{42}(-2)L_{32}(-3) \cdot L_{41}(-1)L_{31}(1)L_{21}(-3)A = U.$$

Procedendo como no exemplo em que a matriz era 3×3 , multiplicamos esta equação à esquerda, pelas inversas das várias matrizes elementares, uma de cada vez, o que nos dá

$$A = \underbrace{L_{21}(3)L_{31}(-1)L_{41}(1)L_{32}(3)L_{42}(2)L_{43}(-2)}_L \cdot U.$$

Resta-nos multiplicar as matrizes elementares para obter L . Por (53)

$$L_{42}(2)L_{43}(-2) = I + 2E_{42} - 2E_{43}.$$

Contudo, usando a distributividade do produto de matrizes e (52),

$$L_{32}(3)L_{42}(2)L_{43}(-2) = I + 2E_{42} - 2E_{43} + 3E_{32}.$$

Continuando desta maneira, obtemos finalmente que

$$L = I + 2E_{42} - 2E_{43} + 3E_{32} + 3E_{21} + (-1)E_{31} + E_{41};$$

isto é,

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

Observe que nos dois exemplos que fizemos, a entrada ij da matriz L foi preenchida com $k_{i,j}$ para cada $L_{i,j}(k_{i,j})$ no produto que define L . Encerraremos a seção mostrando que isto é sempre verdade. Seja

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix}$$

Então é possível escolher números reais $k_{i,j}$ de modo que

$$L_{n+1,n-1}(k_{n+1,n-1}) \cdots L_{3,1}(k_{3,1})L_{2,1}(k_{2,1})A = U$$

é uma matriz triangular superior. Seguindo o roteiro dos exemplos, multiplicamos esta equação à esquerda, sucessivamente, pelas inversas das matrizes elementares, obtendo

$$A = \underbrace{L_{2,1}(-k_{2,1})L_{3,1}(-k_{3,1}) \cdots L_{n+1,n-1}(-k_{n+1,n-1})}_L \cdot U.$$

Resta-nos apenas calcular as entradas de L . Copiando o que já fizemos nos exemplos, sabemos de (53), que

$$L_{21}(-k_{21})L_{31}(-k_{31}) = I - k_{21}E_{21} - k_{31}E_{31}.$$

Digamos que conseguimos verificar que, para algum $r < s < n$,

$$L_{21}(-k_{21})L_{31}(-k_{31}) \cdots L_{rs}(-k_{rs}) = I - k_{21}E_{21} - k_{31}E_{31} - \cdots - k_{rs}E_{rs}.$$

Precisamos apenas nos certificar de que seria possível continuar o procedimento, multiplicando a a matriz acima pela matriz elementar que vem imediatamente à direita de $L_{rs}(-k_{rs})$. Denotando esta matriz por $I - k_{i,j}E_{i,j}$, temos, pela distributividade da multiplicação, que

$$(I - k_{21}E_{21} - \cdots - k_{rs}E_{rs})(I - k_{i,j}E_{i,j})$$

é igual a

$$I - k_{21}E_{21} - \cdots - k_{i,j}E_{i,j} - k_{21}k_{i,j}E_{21}E_{i,j} - \cdots - k_{rs}k_{i,j}E_{rs}E_{i,j}.$$

Para concluir a fórmula esperada, falta apenas mostrar que

$$(57) \quad E_{21}E_{i,j} = \cdots = E_{rs}E_{i,j} = 0,$$

Lembrando que o processo de eliminação começa sempre na primeira coluna da matriz, temos que $j \geq s$; de modo que $i > r \geq s$. Levando em conta esta desigualdade, (57) segue diretamente de (52).

Continuando desta maneira até que todas as matrizes elementares tenham sido multiplicadas, teremos que

$$L = I - k_{2,1}E_{2,1} - \cdots - k_{n,n-1}E_{n,n-1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -k_{2,1} & 1 & 0 & \cdots & 0 & 0 \\ -k_{3,1} & -k_{3,2} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -k_{n,1} & -k_{n,2} & -k_{n,3} & \cdots & -k_{n,n-1} & 1 \end{bmatrix}.$$

Note que esta fórmula vale mesmo se houver entradas nulas, que não é preciso eliminar, porque neste caso a constante $k_{i,j}$ seria nula e não contribuiria nenhuma entrada nova à matriz L .

3. Sistemas lineares e decomposição LU

É hora de considerar como a decomposição LU pode ser usada para resolver sistemas lineares e quais são suas vantagens sobre a eliminação gaussiana usual. Na seção 3 do capítulo 1, vimos que é conveniente representar o sistema linear

$$(58) \quad \begin{aligned} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n &= b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,n}x_n &= b_2 \\ &\vdots \\ a_{n,1}x_1 + a_{n,2}x_2 + \cdots + a_{n,n}x_n &= b_n \end{aligned}$$

em termos de sua matriz aumentada

$$\left[\begin{array}{cccc|c} a_{1,1} & a_{1,2} & \cdots & a_{1,n} & b_1 \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} & b_n \end{array} \right].$$

Contudo, esta matriz provê apenas uma maneira abreviada de escrever os sistema linear; uma maneira na qual as variáveis podem ser descartadas, porque sabemos que coeficiente multiplica que variável por suas posições na matriz. Em outras palavras, apesar de termos construído uma matriz, nunca calculamos com ela, porque as operações elementares por linha não estavam sendo consideradas como operações com matrizes até introduzirmos as matrizes elementares na seção 1 deste capítulo.

Apesar de podermos efetuar as operações por linha multiplicando a matriz aumentada pelas matrizes elementares correspondentes, é preferível reescrever o sistema linear original como uma equação matricial e, então, usar decomposição LU para resolvê-lo. Para podermos escrever a equação matricial associada ao sistema (58), precisamos de

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix}, \quad X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

Na terminologia usual, A é conhecida como a *matriz do sistema*, X como a *matriz das variáveis* e b como a *matriz dos termos constantes*. Com isto o sistema (58) corresponde à equação matricial

$$(59) \quad AX = b.$$

Por exemplo, no caso em que o sistema linear é

$$(60) \quad \begin{aligned} x + 3y + 2z &= 11 \\ -2x - 4y + z &= 7 \\ 2x + 2y + 2z &= -3 \end{aligned}$$

teremos que

$$A = \begin{bmatrix} 1 & 3 & 2 \\ -2 & -4 & 1 \\ 2 & 2 & 2 \end{bmatrix}, \quad X = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} 11 \\ 7 \\ -3 \end{bmatrix}.$$

Para resolver um sistema linear usando decomposição LU, começamos por escrever $A = LU$, usando o algoritmo da seção 2. Substituindo isto em $AX = b$, obtemos

$$LUX = b,$$

que pode ser resolvido em dois passos. Primeiramente, resolvemos o sistema auxiliar

$$LY = b, \quad \text{em que} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

é uma nova matriz de variáveis. Digamos que a matriz coluna $c \in \mathbb{R}^n$ seja a solução de $LY = b$. O segundo passo consiste, então, em resolver o sistema $UX = c$, cuja solução $v \in \mathbb{R}^n$ satisfaz $AX = b$. Para confirmar que v é a solução do sistema $AX = b$, basta observar que

$$Av = L \cdot Uv = Lc = b,$$

pois $Uv = c$ e $Lc = b$. Observe que este procedimento só é viável porque L é uma matriz triangular inferior e U uma matriz triangular superior. Isto faz com que $LY = b$ possa ser resolvido por substituição direta e $UX = c$ por substituição reversa.

Um exemplo vai ajudá-lo a entender como isto funciona na prática. Vimos na seção 2 que a decomposição LU da matriz

$$A = \begin{bmatrix} 1 & 3 & 2 \\ -2 & -4 & 1 \\ 2 & 2 & 2 \end{bmatrix},$$

é dada por

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & -2 & 1 \end{bmatrix} \quad \text{e} \quad U = \begin{bmatrix} 1 & 3 & 2 \\ 0 & 2 & 5 \\ 0 & 0 & 8 \end{bmatrix}.$$

Tomando

$$Y = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix},$$

a equação

$$LY = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & -2 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} 11 \\ 7 \\ -3 \end{bmatrix} = b$$

corresponde ao sistema linear

$$\begin{aligned} \alpha &= 11 \\ 2\alpha + \beta &= 7 \\ 2\alpha - 2\beta + \gamma &= -3. \end{aligned}$$

Resolvendo este sistema por substituição direta, obtemos

$$\alpha = 11, \quad \beta = -15 \quad \text{e} \quad \gamma = -55;$$

donde

$$c = \begin{bmatrix} 11 \\ -15 \\ 33 \end{bmatrix}.$$

Portanto,

$$UX = \begin{bmatrix} 1 & 3 & 2 \\ 0 & 2 & 5 \\ 0 & 0 & 8 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 11 \\ -15 \\ 33 \end{bmatrix} = c,$$

que corresponde ao sistema

$$\begin{aligned} x + 3y + 2z &= 11 \\ 2y + 5z &= -15 \\ 8z &= 33, \end{aligned}$$

que, ao ser resolvido por substituição reversa, nos dá

$$x = -\frac{157}{16}, \quad y = \frac{67}{16} \quad \text{e} \quad z = \frac{33}{8}.$$

A esta altura você deve estar se perguntando qual a vantagem de utilizar duas soluções de sistemas triangulares para resolver um sistema linear, quando a eliminação gaussiana usual nos permite fazer isto resolvendo apenas um sistema triangular superior. A resposta, como já observamos na introdução deste capítulo, é que, em aplicações práticas, nos deparamos frequentemente com a necessidade de resolver vários sistemas lineares muito grandes que diferem apenas pelos termos constantes. Em

casos como este, podemos resolver todos os sistemas ao custo de apenas uma decomposição LU, ao passo que o procedimento clássico exigiria uma eliminação gaussiana para cada um dos sistemas.

4. Decomposição PLU

Até aqui passamos à margem do que fazer quando o pivô é nulo em algum momento do processo de eliminação. Digamos, por exemplo, que queremos resolver o sistema linear

$$(61) \quad \begin{aligned} 2y + 5z + 2w &= 1 \\ x + 2y + 3z + w &= 1 \\ -2x - 6y - 11z + w &= 3, \\ x + 4y + 11z + 6w &= 2 \end{aligned}$$

cujas matriz aumentada é

$$\left[\begin{array}{cccc|c} 0 & 2 & 5 & 2 & 1 \\ 1 & 2 & 3 & 1 & 1 \\ -2 & -6 & -11 & 1 & 3 \\ 1 & 4 & 11 & 6 & 2 \end{array} \right].$$

Como a primeira entrada desta matriz é nula, não podemos usá-la, na forma em que está, para eliminar as posições não nulas da primeira coluna. Entretanto, como as soluções de um sistema não mudam se trocamos suas linhas de posição, podemos concluir que a matriz aumentada

$$\left[\begin{array}{cccc|c} 1 & 2 & 3 & 1 & 1 \\ 0 & 2 & 5 & 2 & 1 \\ -2 & -6 & -11 & 1 & 3 \\ 1 & 4 & 11 & 6 & 2 \end{array} \right],$$

representa o mesmo sistema que a matriz com que começamos. Infelizmente este argumento *não* pode ser aplicado quando se trata de calcular a decomposição LU de uma matriz, porque uma troca de linhas altera a matriz cuja decomposição queremos achar.

Eliminando as entradas da primeira coluna, obtemos

$$\left[\begin{array}{cccc|c} 1 & 2 & 3 & 1 & 1 \\ 0 & 2 & 5 & 2 & 1 \\ 0 & -2 & -5 & 3 & 5 \\ 0 & 2 & 8 & 5 & 1 \end{array} \right],$$

que, depois de eliminadas as posições abaixo da diagonal na segunda coluna, torna-se

$$\left[\begin{array}{cccc|c} 1 & 2 & 3 & 1 & 1 \\ 0 & 2 & 5 & 2 & 1 \\ 0 & 0 & 0 & 5 & 6 \\ 0 & 0 & 3 & 3 & 0 \end{array} \right].$$

Mais uma vez nos deparamos com a necessidade de trocar linhas, desta vez as duas últimas. Fazendo isto, chegamos à matriz

$$\left[\begin{array}{cccc|c} 1 & 2 & 3 & 1 & 1 \\ 0 & 2 & 5 & 2 & 1 \\ 0 & 0 & 3 & 3 & 0 \\ 0 & 0 & 0 & 5 & 6 \end{array} \right],$$

que é triangular superior, o que nos permite resolver o sistema.

Para nossa sorte, existe uma maneira de sair do impasse. Para isso precisamos modificar o procedimento das seções 1 e 2 para que seja aplicável à matriz

$$A = \begin{bmatrix} 0 & 2 & 5 & 2 \\ 1 & 2 & 3 & 1 \\ -2 & -6 & -11 & 1 \\ 1 & 4 & 11 & 6 \end{bmatrix}$$

do sistema (61). A pergunta que precisamos responder é: existe uma matriz que, multiplicada à esquerda de A , troca entre si as duas primeiras linhas desta matriz? Se existir, uma tal matriz será muito parecida com a identidade. De fato, o produto de suas terceira e quarta linhas por A deve ter por único efeito copiar estas linhas tal qual estão em A . Contudo, embora as duas primeiras linhas também devam ser copiadas, isto é feito traspondo uma com a outra. Mas, para isto, basta trocar as respectivas linhas da matriz identidade. De fato,

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \underbrace{\begin{bmatrix} 0 & 2 & 5 & 2 \\ 1 & 2 & 3 & 1 \\ -2 & -6 & -11 & 1 \\ 1 & 4 & 11 & 6 \end{bmatrix}}_A = \begin{bmatrix} 1 & 2 & 3 & 1 \\ 0 & 2 & 5 & 2 \\ -2 & -6 & -11 & 1 \\ 1 & 4 & 11 & 6 \end{bmatrix}.$$

No que segue, denotaremos por $T_{i,j}$ a matriz $n \times n$ obtida trocando-se entre si as linhas i e j . Como no caso das matrizes elementares, o valor de n será determinado pelo contexto; no caso do exemplo atual, $n = 4$. Multiplicando

$$T_{1,2}A = \begin{bmatrix} 1 & 2 & 3 & 1 \\ 0 & 2 & 5 & 2 \\ -2 & -6 & -11 & 1 \\ 1 & 4 & 11 & 6 \end{bmatrix}$$

pelas matrizes elementares que realizam as necessárias eliminações, obtemos

$$L_{4,2}(-1)L_{3,2}(1)L_{4,1}(-1)L_{3,1}(2)T_{1,2}A = \begin{bmatrix} 1 & 2 & 3 & 1 \\ 0 & 2 & 5 & 2 \\ 0 & 0 & 0 & 5 \\ 0 & 0 & 3 & 3 \end{bmatrix};$$

o que produz um novo pivô nulo, desta vez na terceira linha. Mas, multiplicando esta matriz por $T_{3,4}$ para trocar suas duas últimas linhas de posição,

$$(62) \quad T_{3,4}L_{4,2}(-1)L_{3,2}(1)L_{4,1}(-1)L_{3,1}(2)T_{1,2}A = \begin{bmatrix} 1 & 2 & 3 & 1 \\ 0 & 2 & 5 & 2 \\ 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 5 \end{bmatrix} = U,$$

que é uma matriz triangular superior, como desejado.

Caso o produto por $T_{3,4}$, ao final da eliminação, não tivesse sido necessário, poderíamos interpretar o resultado acima como afirmando que o procedimento usual de eliminação funciona, *desde que comecemos fazendo uma troca de linhas na matriz inicial*. Naturalmente, poderíamos recorrer à mesma interpretação, se conseguíssemos, de alguma maneira, passar o $T_{3,4}$ do final para o início da eliminação. Isso seria fácil se $T_{3,4}$ comutasse com as demais matrizes, o que infelizmente não ocorre. Por exemplo,

$$T_{3,4}L_{4,2}(-1) \cdot T_{3,4} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = L_{3,2}(-1);$$

isto é,

$$T_{3,4}L_{4,2}(-1)T_{3,4} = L_{3,2}(-1).$$

Mas, multiplicando esta equação à direita por $T_{3,4}$ obtemos

$$(63) \quad T_{3,4}L_{4,2}(-1) = L_{3,2}(-1)T_{3,4},$$

já que $T_{3,4}^2 = I$. Note que a posição em **negrito** apenas trocou de linha, sem que houvesse nenhuma troca de coluna. Além disso, a troca de linhas corresponde ao fato de $T_{3,4}$ trocar as linhas 3 e 4 quando multiplicada à esquerda de uma dada matriz. Isto se dá porque o único efeito da multiplicação à direita por $T_{3,4}$ foi de recolocar na diagonal os 1s das colunas 3 e 4.

Podemos, então, usar (63) para passar $T_{3,4}$ da esquerda para à direita de $L_{4,2}(-1)$, ao custo de substituí-la por $L_{3,2}(-1)$; o que nos permite reescrever (62) na forma

$$(64) \quad L_{3,2}(-1)T_{3,4}L_{3,2}(1)L_{4,1}(-1)L_{3,1}(2)T_{1,2}A = U.$$

Com isto $T_{3,4}$ avançou uma casa para à direita. Como

$$T_{3,4}L_{3,2}(1) = L_{4,2}(1)T_{3,2},$$

um argumento análogo pode ser usado para passar $T_{3,4}$ para a direita de $L_{3,2}(1)$ em (64); como consequência disto, obtemos

$$L_{3,2}(-1)L_{4,2}(1)T_{3,4}L_{4,1}(-1)L_{3,1}(2)T_{1,2}A = U.$$

Continuando desta maneira,

$$T_{3,4}L_{4,1}(-1) = L_{3,1}(-1)T_{3,4}$$

nos dá

$$L_{3,2}(-1)L_{4,2}(1)L_{3,1}(-1)T_{3,4}L_{3,1}(2)T_{1,2}A = U;$$

ao passo que

$$T_{3,4}L_{3,1}(2) = L_{4,1}(2)T_{3,4}$$

nos permite chegar em

$$L_{3,2}(-1)L_{4,2}(1)L_{3,1}(-1)L_{4,1}(2)T_{3,4}T_{1,2}A = U.$$

Passando, então, as matrizes elementares da esquerda para à direita, como fizemos na seção 2, teremos

$$T_{3,4}T_{1,2}A = L_{4,1}(-2)L_{3,1}(1)L_{4,2}(-1)L_{3,2}(1)U.$$

Escrevendo

$$P = T_{3,4}T_{1,2} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

e $L = L_{4,1}(-2)L_{3,1}(1)L_{4,2}(-1)L_{3,2}(1)$, obtemos

$$PA = LU;$$

que é a *decomposição PLU* da matriz A . Voltando à motivação para os cálculos acima, a equação $PA = LU$ mostra que o algoritmo para decomposição LU da seção 2 pode ser aplicado sem sobressaltos, *desde que as linhas de A sejam inicialmente rearrumadas usando a matriz P* . Entretanto, apesar desta ser uma maneira muito satisfatória de interpretar o que fizemos, na prática *não há como determinar P antes de executar o algoritmo da decomposição LU*. Isto significa que só podemos ter certeza de que isto leva a um algoritmo para achar a decomposição PLU, se formos capazes de generalizar o truque de passar $T_{i,j}$ da esquerda para a direita de uma matriz elementar, obtendo, em troca, uma outra matriz elementar.

Começamos nossa análise com uma observação simples, que reduz bastante as várias combinações de índices que precisamos considerar. Supondo que $i < j$, verificamos que:

se, ao longo do processo de eliminação, nos deparamos com a necessidade de trocar as linhas i e j entre si então, todas as operações elementares usadas até este momento tiveram seu pivô em uma coluna anterior à i -ésima.

Como as matrizes elementares usadas para eliminar as posições abaixo da entrada s, s são todas da forma $L_{r,s}(k)$, com $r > s$, a tradução matricial da afirmação acima é:

se nos deparamos com a necessidade de trocar linhas usando $T_{i,j}$, então todas as matrizes elementares $L_{r,s}(k)$ usadas até este momento têm $s < i < j$.

As mudanças de posição das matrizes $T_{3,4}$ foram possíveis porque, embora tenhamos originalmente interpretado $T_{i,j}$ como a matriz obtida transpondo-se as *linhas* i e j da matriz identidade, ela pode ser igualmente considerada como a matriz obtida da identidade pela troca das *colunas* i e j . Como consequência disto, não é difícil ver que se A é uma matriz quadrada qualquer, $AT_{i,j}$ é a matriz obtida de A pela troca das colunas i e j . Supondo que $s < i < j$, isto significa que

$$(65) \quad T_{i,j}E_{r,s}T_{i,j} = \begin{cases} E_{j,s} & \text{quando } r = j; \\ E_{i,s} & \text{quando } r = i; \\ E_{r,s} & \text{quando } r \neq i, j. \end{cases}$$

Como estamos supondo que $s < i < j$, nem i , nem j , podem ser iguais a s , de modo que a multiplicação à esquerda por $T_{i,j}$ na fórmula acima não altera o resultado do produto $T_{i,j}E_{r,s}$. Levando em conta que

$$T_{i,j}L_{r,s}(k)T_{i,j} = (I + kT_{i,j}E_{r,s}T_{i,j})$$

segue de (65) que se $s < i < j$, então

$$(66) \quad T_{i,j}L_{r,s}(k) = \begin{cases} (I + kE_{j,s})T_{i,j} & \text{quando } r = j \\ (I + kE_{r,s})T_{i,j} & \text{quando } r \neq i, j. \end{cases}$$

Encerraremos aplicando o que aprendemos nesta seção para calcular a decomposição PLU da matriz

$$A = \begin{bmatrix} 1 & 2 & 3 & 2 & 1 \\ -2 & -4 & 0 & 12 & 8 \\ 1 & 2 & 6 & 10 & 5 \\ 3 & 13 & 14 & 6 & 6 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}.$$

Multiplicando A pelas devidas matrizes elementares, eliminamos as posições da primeira coluna que ficam abaixo da diagonal,

$$L_{4,1}(-3)L_{3,1}(-1)L_{2,1}(2)A = \begin{bmatrix} 1 & 2 & 3 & 2 & 1 \\ 0 & 0 & 6 & 16 & 10 \\ 0 & 0 & 3 & 8 & 4 \\ 0 & 7 & 5 & 0 & 3 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

Como a entrada 2,2 desta última matriz é nula, precisamos trocar a segunda linha com a quarta antes de continuar,

$$T_{2,4}L_{4,1}(-3)L_{3,1}(-1)L_{2,1}(2)A = \begin{bmatrix} 1 & 2 & 3 & 2 & 1 \\ 0 & 7 & 5 & 0 & 3 \\ 0 & 0 & 3 & 8 & 4 \\ 0 & 0 & 6 & 16 & 10 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

Prosseguindo, eliminamos a posição não nula da terceira coluna abaixo da diagonal,

$$L_{4,3}(-2)T_{2,4}L_{4,1}(-3)L_{3,1}(-1)L_{2,1}(2)A = \begin{bmatrix} 1 & 2 & 3 & 2 & 1 \\ 0 & 7 & 5 & 0 & 3 \\ 0 & 0 & 3 & 8 & 4 \\ 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

Finalmente as duas últimas linhas precisam ser trocadas entre si para que a matriz seja triangular superior,

$$T_{4,5}L_{4,3}(-2)T_{2,4}L_{4,1}(-3)L_{3,1}(-1)L_{2,1}(2)A = \begin{bmatrix} 1 & 2 & 3 & 2 & 1 \\ 0 & 7 & 5 & 0 & 3 \\ 0 & 0 & 3 & 8 & 4 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

Chamando esta última matriz de U , podemos escrever

$$(67) \quad T_{4,5}L_{4,3}(-2)T_{2,4}L_{4,1}(-3)L_{3,1}(-1)L_{2,1}(2)A = U.$$

Em seguida, movemos a matriz $T_{2,4}$ para à direita, um passo de cada vez. Usando (66),

$$T_{2,4}L_{4,1}(-3) = L_{2,1}(-3)T_{2,4},$$

obtemos

$$T_{4,5}L_{4,3}(-2)L_{2,1}(-3)T_{2,4}L_{3,1}(-1)L_{2,1}(2)A = U.$$

Analogamente,

$$T_{2,4}L_{3,1}(-1) = L_{3,1}(-1)T_{2,4} \quad \text{e} \quad T_{2,4}L_{2,1}(2) = L_{4,1}(2)T_{2,4},$$

nos dão

$$T_{4,5}L_{4,3}(-2)L_{2,1}(-3)L_{3,1}(-1)L_{4,1}(2)T_{2,4}A = U.$$

Resta-nos passar $T_{4,5}$ para à direita. Como ilustrado no procedimento anterior, isto equivale a trocar todos os 5 por 4 e todos os 4 por 5 nas matrizes elementares entre $T_{4,5}$ e A ; deixando as demais matrizes inalteradas. Fazendo isto, obtemos

$$L_{5,3}(-2)L_{2,1}(-3)L_{3,1}(-1)L_{5,1}(2)T_{4,5}T_{2,4}A = U.$$

Finalmente, multiplicando esta equação à esquerda pelas inversas das matrizes elementares, chegamos a

$$T_{4,5}T_{2,4}A = L_{5,1}(-2)L_{3,1}(1)L_{2,1}(3)L_{5,3}(2)U.$$

Temos, assim, $PA = LU$, em que

$$U = \begin{bmatrix} 1 & 2 & 3 & 2 & 1 \\ 0 & 7 & 5 & 0 & 3 \\ 0 & 0 & 3 & 8 & 4 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix}, \quad P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix},$$

ao passo que

$$L = L_{5,1}(-2)L_{3,1}(1)L_{2,1}(3)L_{5,3}(2) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -2 & 0 & 2 & 0 & 1 \end{bmatrix}.$$

5. Pivoteamento

É hora de admitir que os exemplos que fizemos até agora não ilustram o que realmente acontece quando calculamos a decomposição PLU de uma matriz construída a partir de um problema físico. A razão é que todos os nossos sistemas só envolviam números racionais, com os quais podíamos calcular de maneira exata. Mas quantidades obtidas a partir de medidas reais são conhecidas apenas aproximadamente, o que nos obriga a efetuar os cálculos usando números em ponto flutuante; que por sua vez, introduzem uma nova fonte de erros. Portanto, antes de utilizar esta maneira de resolver sistemas lineares, precisamos investigar como se comporta quando os cálculos são feitos, não de maneira exata, mas sim usando números em ponto flutuante.

Começaremos nossa análise com um exemplo aparentemente inofensivo. Suponhamos que queremos calcular a decomposição PLU da matriz

$$A = \begin{bmatrix} 10^{-40} & 1 \\ 1 & 1 \end{bmatrix}.$$

Multiplicando a primeira linha por -10^{40} e somando à segunda, obtemos a matriz

$$U = \begin{bmatrix} 10^{-40} & 1 \\ 0 & 1 - 10^{40} \end{bmatrix}.$$

Até aqui, nenhuma novidade, porque o cálculo foi exato. Mas, o que teria acontecido se tivéssemos feito a mesma conta usando ponto flutuante com arredondamento em um computador cuja mantissa tem 20 algarismos? Neste caso, seria necessário arredondar

$$1 - 10^{40} = \underbrace{99 \cdots 999}_{40 \text{ vezes}} = 0. \underbrace{99 \cdots 999}_{40 \text{ vezes}} \cdot 10^{40},$$

obtendo, como resultado, $0.1 \cdot 10^{41}$. Portanto, a decomposição PLU de A calculada num tal computador produziria as matrizes

$$L = \begin{bmatrix} 1 & 0 \\ 10^{40} & 1 \end{bmatrix} \quad \text{e} \quad U = \begin{bmatrix} 10^{-40} & 1 \\ 0 & 10^{40} \end{bmatrix}.$$

O problema é que estas matrizes têm como produto

$$\begin{bmatrix} 10^{-40} & 1 \\ 1 & 0 \end{bmatrix},$$

de modo que o erro cometido na entrada 2, 2, quando substituimos A por LU é igual a 1 e, portanto, muito grande. Felizmente, este é um obstáculo fácil de contornar. Para isto basta transpor as duas linhas da matriz. Tomando

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

calculamos, então, a decomposição LU de

$$PA = \begin{bmatrix} 1 & 1 \\ 10^{-40} & 1 \end{bmatrix},$$

obtendo

$$L = \begin{bmatrix} 1 & 0 \\ 10^{-40} & 1 \end{bmatrix} \quad \text{e} \quad U = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

em que a entrada 2, 2 já foi devidamente arredondada, a partir do valor exato

$$1 - 10^{-40} = 1 - 0. \underbrace{00 \cdots 00}_{39 \text{ vezes}} 1 = 0. \underbrace{99 \cdots 99}_{40 \text{ vezes}}.$$

Contudo, desta vez o produto LU dá

$$\begin{bmatrix} 1 & 1 \\ 10^{-40} & 1 \end{bmatrix}$$

em que, mais uma vez, tivemos que arredondar a entrada 2, 2, cujo valor exato também era igual a $1 - 10^{-40}$. Só que, desta vez, os dois arredondamentos acabaram nos dando de volta exatamente a matriz PA com a qual começamos.

O segredo por trás do que fizemos consistiu em substituir um pivô pequeno, em nosso exemplo 10^{-40} , por um que é comparativamente grande—igual a 1 em nosso exemplo. Em geral, para melhor controlar o erro inerente às aproximações de ponto flutuante, faremos o que é conhecido como *pivoteamento parcial*, que consiste em trocar linhas de posição de modo a substituir o pivô de cada etapa da eliminação pela maior entrada, em valor absoluto, que está abaixo do pivô, mas na mesma coluna que ele. Para formular isto de uma maneira mais precisa, digamos que estamos por aplicar eliminação a partir da posição k, k da matriz

$$(68) \quad \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,k-1} & a_{1,k} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,k-1} & a_{2,k} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & a_{k,k} & \cdots & a_{k,n} \\ 0 & 0 & \cdots & 0 & a_{k+1,k} & \cdots & a_{k+1,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & a_{n,k} & \cdots & a_{n,n} \end{bmatrix}.$$

Se estivermos usando *pivoteamento parcial* então, antes de eliminar as posições abaixo de $a_{k,k}$, determinamos $k \leq \ell \leq n$ tal que

$$|a_{\ell,k}| = \max\{|a_{j,k}| \mid j \geq k\}.$$

e multiplicamos a matriz acima por $T_{k,\ell}$, para trocar a linha ℓ de posição com a linha k . Só então efetuamos a eliminação das posições abaixo da diagonal na coluna k .

Um exemplo ajudará a tornar isto mais claro. Digamos que desejamos calcular a decomposição PLU da matriz

$$A = \begin{bmatrix} 1 & 3 & 2 \\ -2 & -4 & 1 \\ 2 & 2 & 2 \end{bmatrix},$$

que apareceu em um exemplo da primeira seção deste capítulo mas, desta vez, usando pivoteamento parcial. Como as posições abaixo do pivô têm ambas valor absoluto igual a $2 > 1$, podemos trocar a primeira linha pela segunda ou pela terceira, indiferentemente. Fazendo a troca com a segunda linha, obtemos

$$L_{3,1}(1)L_{2,1}(1/2) \cdot T_{1,2} \cdot A = \begin{bmatrix} -2 & -4 & 1 \\ 0 & 1 & \frac{5}{2} \\ 0 & -2 & 3 \end{bmatrix}$$

Desta vez o pivô está na posição 2, 2 e é igual a 1. Como a posição abaixo dele tem valor absoluto 2, trocaremos as duas últimas linhas de posição antes de fazer a

eliminação, o que nos dá

$$L_{3,2}(1/2) \cdot T_{2,3} \cdot L_{3,1}(1) L_{2,1}(1/2) \cdot T_{1,2} \cdot A = \begin{bmatrix} -2 & -4 & 1 \\ 0 & -2 & 3 \\ 0 & 0 & 4 \end{bmatrix}.$$

Denotando esta última matriz por U e passando as transposições de linhas para a direita,

$$L_{3,2}(1/2) L_{2,1}(1) L_{3,1}(1/2) \cdot T_{2,3} T_{1,2} \cdot A = U.$$

Finalmente, multiplicando os dois lados desta equação, à esquerda, pelas inversas das matrizes elementares,

$$\underbrace{T_{2,3} T_{1,2}}_P \cdot A = \underbrace{L_{3,1}(-1/2) L_{2,1}(-1) L_{3,2}(-1/2)}_L \cdot U.$$

Portanto, a decomposição PLU de A , com pivoteamento parcial, é dada pelas matrizes

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \quad L = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -\frac{1}{2} & -\frac{1}{2} & 1 \end{bmatrix} \quad \text{e} \quad U = \begin{bmatrix} -2 & -4 & 1 \\ 0 & -2 & 3 \\ 0 & 0 & 4 \end{bmatrix}.$$

A esta altura você pode estar se perguntando: se isto é pivoteamento *parcial*, existe um pivoteamento *total*? A resposta é sim, mas não vamos utilizá-lo. Para aplicar *pivoteamento total* à matriz (68) a partir da posição k, k , buscamos a entrada $a_{i,j}$ de maior valor absoluto com $k \leq i, j \leq n$. Se esta entrada for $a_{r,s}$, trocamos de lugar as linhas i e r e as colunas j e s . Para isto é necessário multiplicar a matriz à qual estamos aplicando eliminação à direita e à esquerda por transposições, o que complica bastante o algoritmo.

Uma análise mais detalhada do comportamento do erro na eliminação gaussiana nos levaria muito além dos limites de um livro elementar como este. Se o assunto lhe interessa, consulte [5, Chapter 9, p. 157].

CAPÍTULO 5

Ajuste de curvas

No capítulo 2, vimos como aproximar uma função, em um intervalo, usando seu polinômio de Taylor. Contudo isto requer que a função seja diferenciável. Entretanto, como veremos na seção 1, qualquer função contínua pode ser aproximada por um polinômio. Neste capítulo estudaremos maneiras de aproximar funções contínuas por polinômios que se aplicam mesmo que seja conhecido apenas um conjunto finito \mathcal{P} de pontos do seu gráfico. Na seção 2 exigiremos que o gráfico da função polinomial passe por cada um dos pontos de \mathcal{P} . Nas seções 3 e 4 consideraremos o problema, um pouco mais geral, de encontrar a curva polinomial que melhor se adapta aos pontos de \mathcal{P} . Nestas seções trataremos o mesmo problema sob dois enfoques diferentes. O enfoque analítico nos leva rapidamente à solução do problema; já o enfoque geométrico, mais elaborado, nos permite garantir que esta solução sempre existe e é única.

1. Introdução

No capítulo 3 nos deparamos com a necessidade de desenhar uma curva da qual conhecemos, aproximadamente, apenas alguns pontos. A solução que adotamos lá consistiu em ligar os pontos por segmentos de retas. Se a quantidade de pontos for suficientemente grande, nosso olho não será capaz de distinguir os segmentos individuais e teremos a impressão de que se trata de uma curva suave, como esperaríamos da solução de uma equação diferencial de segunda ordem. Afinal de contas, para ter chance de satisfazer a equação, a função teria que ser derivável, pelo menos, até ordem dois.

Entretanto, existem outras maneiras de achar a aproximação de uma curva a partir de alguns dos seus pontos. As duas que vamos estudar aproximam a curva usando polinômios. Talvez você esteja se perguntando sobre o porquê desta fixação em usar polinômios, que já foram utilizados para aproximar funções diferenciáveis no capítulo 2. Há duas razões principais para isto. A primeira é que as funções polinomiais são as únicas que sabemos calcular usando apenas as operações aritméticas básicas. A segunda, menos ingênua, mas não menos importante, é que podemos aproximar quaisquer funções contínuas usando polinômios, por causa do seguinte resultado, devido ao matemático alemão Karl Weierstrass.

TEOREMA DE APROXIMAÇÃO DE WEIERSTRASS. *Toda função contínua, definida em um intervalo fechado e limitado, pode ser aproximada por uma função polinomial, com um erro tão pequeno quanto desejado.*

Quando a função tem derivada de toda ordem, o teorema de Taylor nos permite calcular uma aproximação polinomial tão precisa quanto desejemos; mas, como proceder quando a função é apenas contínua? Veremos na seção 2 que, neste caso, basta que conheçamos $n + 1$ pontos sobre a curva para que possamos aproximá-la por um polinômio de grau n , conhecido como *polinômio interpolador*. Embora este resultado seja bastante útil, é necessário usá-lo com cautela, porque raramente conhecemos pontos que estejam exatamente sobre a curva que desejamos desenhar, o que pode afetar o resultado da aproximação.

A lei de Hooke nos permite dar uma ilustração bastante concreta deste problema. Suponhamos que um objeto é pendurado sob uma mola vertical cuja extremidade superior está fixa. O físico inglês Robert Hooke descobriu em 1676 que o peso do objeto é linearmente proporcional ao alongamento da mola. Denotando por m a massa do objeto e por g a aceleração da gravidade, a mola se distenderá de mg/k unidades de comprimento, em que k é uma constante que depende apenas da mola e representa sua *rigidez*. Portanto, para calcular a rigidez de uma mola, basta pendurar em sua extremidade livre um objeto de massa conhecida e medir o quanto a mola se distendeu quando parar de se mover. Na prática, repetindo o experimento com objetos de massas diferentes, podemos obter uma aproximação melhor para o valor da rigidez. Infelizmente isto normalmente produz vários valores diferentes para a rigidez, como ilustrado na tabela 1.

Massa (g)	0.01	0.02	0.03	0.04
Distensão (cm)	0.12	0.24	0.4	0.51
Rigidez	117.6	117.6	130.67	124.95

TABELA 1. Rigidez de uma mola calculada com diferentes pesos.

Como obter a rigidez da mola a partir destes dados? Uma saída possível seria calcular a média destes resultados. Outra possibilidade, consistiria em procurar a reta que melhor se adapta aos valores experimentais obtidos quando plotamos a massa versus a distensão. A interpolação em nada nos ajuda a achar esta reta, porque o polinômio interpolador produz uma curva, ilustrada na figura 1, que *passa através de cada ponto* e, por isso, não tem a menor chance de ser uma reta. Afinal, se os pontos estivessem alinhados, não teriam produzido valores diferentes para a rigidez da mola.

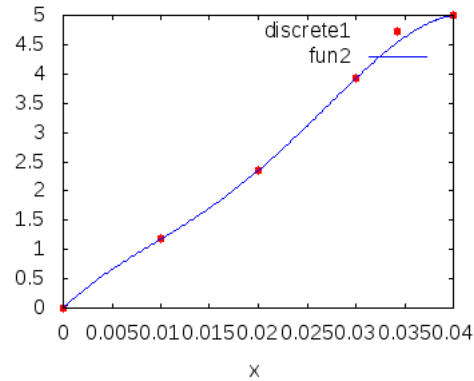


FIGURA 1. Interpolando os pontos da tabela.

Mas, o que significa *a reta que melhor se adapta* a um conjunto de pontos dados? Uma resposta possível a esta pergunta foi dada, no século XIX, por Legendre e Gauss que propuseram, independentemente, a solução hoje conhecida como *método dos mínimos quadrados*, que estudaremos em detalhe nas seções 3 e 4. Aplicando este método aos dados da tabela 1, obtemos a reta cuja equação é

$$127.40x - 0.059$$

e cujo gráfico é ilustrado na figura 2. Portanto, segundo o método dos mínimos quadrados, o coeficiente angular 127.40 desta reta é uma aproximação para a rigidez da mola utilizada no experimento.

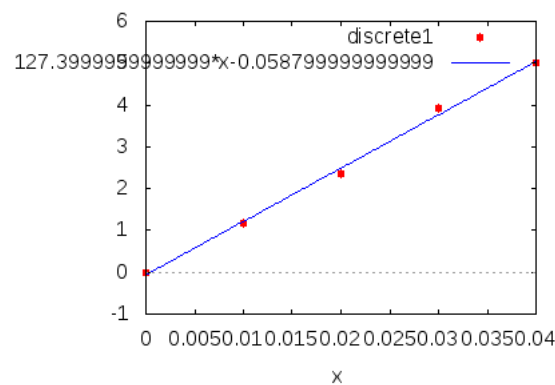



FIGURA 2. Reta que melhor se ajusta aos pontos da tabela.

2. Interpolação

Como mencionado na introdução, o objetivo da interpolação é o de encontrar um polinômio que defina uma função cujo gráfico passe por um conjunto (finito) de pontos cujas coordenadas são conhecidas. Digamos que as coordenadas destes pontos sejam

$$(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n).$$

 Como a remoção de pontos repetidos obviamente não afeta o resultado final, podemos sempre supor que estamos considerando $n + 1$ pontos distintos. Além disso, dois destes pontos não podem ter a mesma abscissa porque uma função associa, a cada ponto do seu domínio, um único valor na imagem. Por isso, suporemos sempre, de agora em diante, que os conjuntos tratados neste capítulo são todos formados por pontos distintos, cujas abscissas são também distintas.

Para que o gráfico de uma função $y = F(x)$ passe pelo ponto de coordenadas (x_0, y_0) é necessário que

$$(69) \quad F(x_0) = y_0;$$

Se a função for polinomial, então

$$F(x) = a_m x^m + \dots + a_1 x + a_0,$$

de modo que (69) pode ser escrita na forma

$$(70) \quad a_0 + a_1 x_0 + \dots + a_m x_0^m = y_0.$$

Como x_0 e y_0 são conhecidos, podemos considerar a equação acima como impondo uma restrição sobre quais são os valores que a_m, \dots, a_1, a_0 podem tomar. Isto sugere que uma estratégia possível consistiria em supor que os coeficientes de F são variáveis cujos valores têm que satisfazer equações equivalentes a (70) para cada um dos pontos de \mathcal{P} . Escrevendo todas estas equações juntas, obtemos o sistema linear

$$(71) \quad \begin{aligned} a_0 + a_1 x_0 + \dots + a_m x_0^m &= y_0 \\ a_0 + a_1 x_1 + \dots + a_m x_1^m &= y_1 \\ &\vdots \\ a_0 + a_1 x_n + \dots + a_m x_n^m &= y_n. \end{aligned}$$

O ideal seria que nosso sistema fosse determinado, mas para isto é necessário que o número de variáveis $m + 1$ seja igual ao número de equações $n + 1$; isto é, que

$m = n$. Supondo que isto ocorre, a matriz do sistema (71) será

$$V_n = \begin{bmatrix} 1 & x_0 & \cdots & x_0^n \\ 1 & x_1 & \cdots & x_1^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^n \end{bmatrix}$$

que é uma *matriz de Vandermonde*. Estas matrizes têm como determinante

$$\det(V_n) = \prod_{i < j} (x_j - x_i).$$

Como as abscissas dos pontos de \mathcal{P} são todas distintas, teremos que $\det(V_n) \neq 0$, de modo que o sistema (71) será sempre determinado. Logo, existe um único polinômio de grau n cujo gráfico passa por todos os pontos dados.

A curva da figura 3 foi construído aplicando interpolação polinomial à solução do problema de valor de contorno definido pela equação (50) da página 63. Como você pode constatar, a solução dada pela interpolação (em vermelho) praticamente coincide com a solução exata.

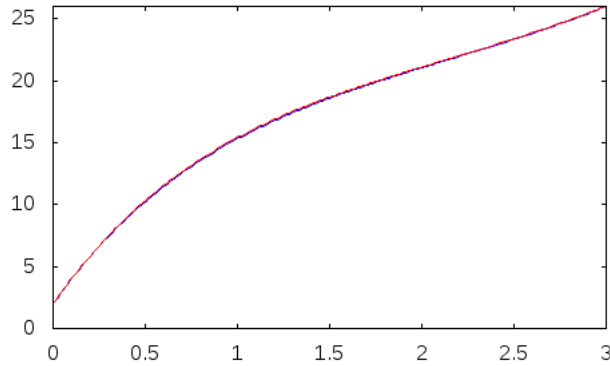


FIGURA 3. Interpolação aplicada ao problema de valor de contorno (50).

Infelizmente o resultado nem sempre é tão bom. Por exemplo, vejamos o que acontece quando aproximamos o gráfico da função $f(x) = 1/(1 + 25x^2)$, desenhado em azul 4, usando o polinômio que interpola os pontos

$$(-1 + i/3, f(-1 + i/3)) \quad \text{para} \quad i = 0, \dots, 6,$$

obtidos subdividindo o intervalo $[-1, 1]$ em seis partes iguais, cujo gráfico está desenhado em vermelho.

O resultado é bastante ruim, mas talvez você ache que isto ocorreu porque subdividimos o intervalo em um número relativamente pequeno de pontos. Dobrando

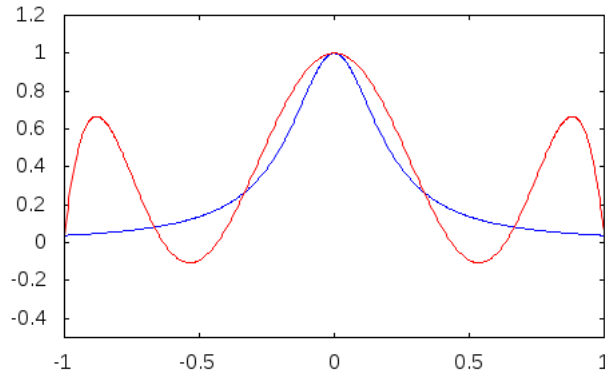


FIGURA 4. Interpolando $1/(1 + 25x^2)$ por um polinômio de grau 6.

a quantidade de segmentos na subdivisão, usaremos o polinômio interpolador com $n = 12$ e $h = 1/6$. O gráfico passa pelos pontos

$$(-1 + i/6, f(-1 + i/6)) \quad \text{para} \quad i = 0, \dots, 12,$$

para desenhar nossa próxima figura.

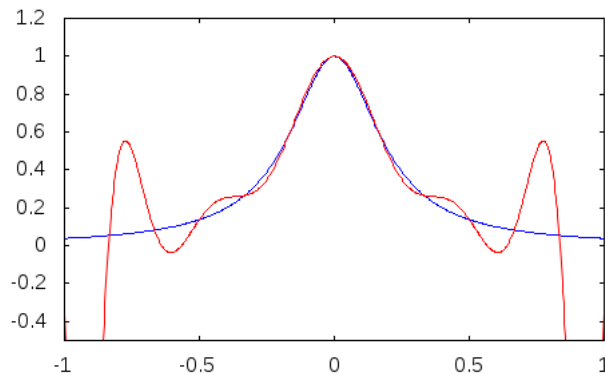


FIGURA 5. Interpolando $1/(1 + 25x^2)$ por um polinômio de grau 12.

Como o gráfico continua ruim, vamos dividir cada pequeno segmento por quatro, que equivale a tomar $n = 38$ e $h = 1/24$.

A esta altura você deve estar convencido de que aumentar a quantidade de pontos só está piorando o resultado. Na última figura o mínimo do gráfico nas extremidades direita e esquerda chegou a cair abaixo de -10^6 , ao passo que todos os pontos do gráfico de $y = 1/(1 + 25x^2)$ têm ordenadas positivas. Entretanto, como a escala do eixo vertical teve que ser drasticamente alterada para alcançar -10^6 , a parte do gráfico acima do eixo x ficou totalmente achatada. Para descobrir o que está

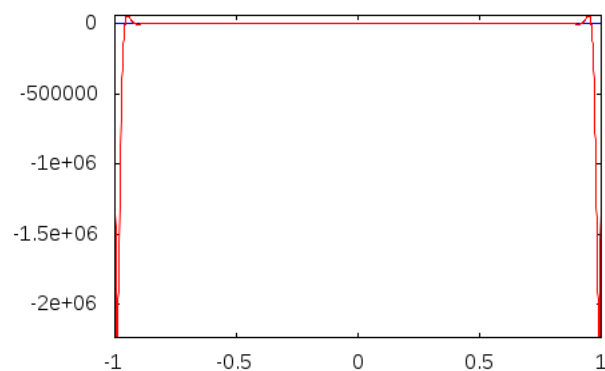


FIGURA 6. Interpolando $1/(1 + 25x^2)$ por um polinômio de grau 48.

acontecendo lá em cima, desenharemos a parte do mesmo gráfico que fica entre $-1/4$ e $1/4$. O resultado é ilustrado na figura 7, onde você pode constatar que, como seria de esperar, ao menos no centro da figura, o polinômio de grau 48 de fato produz uma aproximação excelente do gráfico de $y = 1/(1 + 25x^2)$.

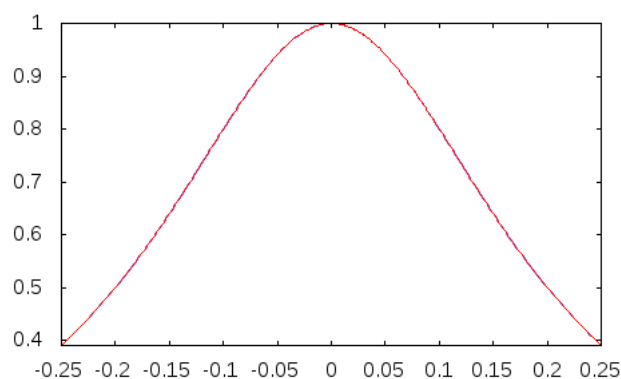


FIGURA 7. Interpolando $1/(1 + 25x^2)$ por um polinômio de grau 48.

Infelizmente, este é um pequeno consolo, diante do desastre que ocorre nas extremidades do gráfico. O problema ilustrado nestes gráficos é conhecido como *fenômeno de Runge* e foi descoberto pelo matemático alemão C. Runge em 1901, ao estudar o comportamento do erro na interpolação. Para evitar este tipo de problema, é necessário utilizar uma subdivisão do intervalo $[-1, 1]$ em subintervalos *desiguais*, cujos tamanhos precisam ser escolhidos com cuidado. Um estudo detalhado do fenômeno de Runge pode ser encontrado em [3].

Para encerrar a seção em um clima mais positivo, faremos um exemplo no qual a interpolação é usada com sucesso. A tabela 2 resume os dados de população dos censos do IBGE desde 1960.

Ano	1960	1970	1980	1991	2000	2010
População	70	93	121	146	168	189

TABELA 2. População brasileira em milhões de habitantes.

Como os censos só ocorrem a cada 10 anos, o que podemos fazer se, por alguma razão, precisarmos conhecer a população brasileira em um ano que cai entre dois censos? Uma saída é utilizar a interpolação. Por exemplo,

quantos eram os brasileiros em 1986, ano em que a IBM lançou o *PC Convertible*, que foi o primeiro laptop?

Usando os valores dos censos de 1960, 1970 e 1980 e contando o tempo em anos a partir de 1960, devemos achar o polinômio que interpola os pontos

$$\{(0, 70), (10, 93), (20, 121)\}.$$

As matrizes correspondentes a estes pontos são

$$V = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 10 & 100 \\ 1 & 20 & 400 \end{bmatrix} \quad \text{e} \quad \begin{bmatrix} 70 \\ 93 \\ 121 \end{bmatrix}$$

e a solução do sistema $V\mathbf{a} = \mathbf{b}$ é

$$\mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 70 \\ \frac{41}{20} \\ \frac{1}{40} \end{bmatrix}$$

de modo que o polinômio interpolador é

$$F_1(x) = x^2/40 + (41 * x)/20 + 70.$$

Como entre 1960 e 1986 passaram-se 26 anos, a população brasileira em 1986 era de, aproximadamente,

$$F_1(26) = 140.2$$

milhões de habitantes. Contudo, poderíamos, igualmente, ter usado os dados dos censos de 1970, 1980 e 1991 em nossa estimativa. Fazendo isto, e tomando o ano zero como sendo 1970, obtemos o polinômio

$$F_2(x) = -\frac{29x^2}{1155} + \frac{3524x}{1155} + 93.$$

Como entre 1970 e 1986 passaram-se 16 anos, a estimativa para a população em 1986, obtida a partir de $F_2(x)$ é de

$$F(16) = 135.4$$

milhões de habitantes. Finalmente, usando os dados referentes aos quatro anos em que o censo foi realizado entre 1960 e 1991, obtemos

$$F_3(x) = \frac{463x^3}{286440} + \frac{7017x^2}{95480} + \frac{247301x}{143220} + 70$$

para o qual a estimativa é de

$$F_3(26) = 136.2$$

milhões de habitantes.



Como você pode ver, escolhas de anos diferentes produzem estimativas diferentes. Além disso, o fenômeno de Runge mostra que, simplesmente aumentar a quantidade de dados a serem interpolados, pode produzir um efeito exatamente oposto ao desejado porque, para que a curva polinomial consiga passar por todos os pontos, ela pode vir a se contorcer a tal ponto que os resultados da interpolação tornam-se totalmente espúrios.

3. Mínimos quadrados: o enfoque analítico

Nesta seção veremos como formular e resolver um problema de minimização de uma função quadrática, cuja solução nos permite encontrar o polinômio cujo gráfico melhor se adapta a um dado conjunto \mathcal{P} de pontos do plano. A resposta não é necessariamente o polinômio interpolador, uma vez que não estamos exigindo que a curva passe por todos os pontos de \mathcal{P} ; o que queremos é a curva polinomial, de um dado grau, que melhor se adapta a estes pontos, e estas duas curvas geralmente não coincidem. Como no caso da interpolação, suporemos sempre que os pontos de \mathcal{P} são distintos e têm abscissas distintas.

A primeira coisa a ser decidida é o que deve ser entendido quando usamos a expressão *curva que melhor se adapta aos pontos de \mathcal{P}* . Na verdade, não há uma maneira única de responder a esta pergunta. A solução que adotaremos foi proposta no século XIX, independentemente, por A.-M. Legendre e C. F. Gauss.

Como a curva que estamos buscando é polinomial, podemos escrevê-la na forma $y = F(x)$, em que $F(x)$ é, por hipótese, um polinômio. O valor a ser escolhido para o grau de $F(x)$ vai depender do problema que estamos resolvendo. Segundo Gauss e Legendre, para que $y = F(x)$ melhor se adapte aos pontos

$$\mathcal{P} = \{(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)\},$$

devemos escolher seus coeficientes de modo a *minimizar* a soma dos *quadrados* das distâncias de cada ponto $P_i = (x_i, y_i) \in \mathcal{P}$ ao ponto da curva $y = F(x)$ cuja abscissa

é x_i ; veja Figura 8. Por isso o algoritmo inventado por Legendre e Gauss é conhecido como *método dos mínimos quadrados*.

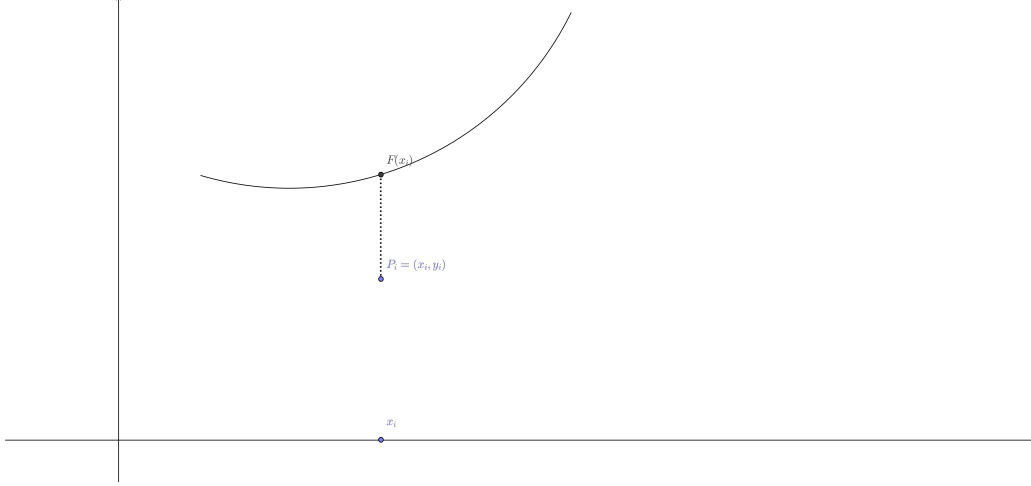


FIGURA 8. Distância entre P_i e $F(x_i)$

Para escrever isto de maneira mais precisa suporemos que $F(x)$ tem grau m , de modo que podemos escrevê-lo na forma

$$(72) \quad F(x) = a_m x^m + \cdots + a_1 x + a_0.$$

Como no caso da interpolação, os coeficientes a_0, a_1, \dots, a_m são indeterminados; ao encontrá-los, teremos achado $F(x)$ e, portanto, a curva desejada.

PROBLEMA DOS MÍNIMOS QUADRADOS. *Dados, um conjunto finito de pontos \mathcal{P} e um inteiro positivo m , determinar um polinômio F tal que*

$$(73) \quad (F(x_0) - y_0)^2 + \cdots + (F(x_n) - y_n)^2,$$

toma o menor valor possível.

É possível que você esteja imaginando porque alguém escolheria a soma dos *quadrados* das distâncias, em vez de simplesmente a soma das distâncias. A resposta é que, como veremos, a soma dos quadrados é muito mais fácil de minimizar.

Observe que desejamos minimizar (73) como função dos coeficientes a_0, a_1, \dots, a_m do polinômio F . Para tornar isto mais evidente, escreveremos

$$\delta(a_0, a_1, \dots, a_m) = (F(x_0) - y_0)^2 + \cdots + (F(x_n) - y_n)^2.$$

Como esta função está definida para todo valor de a_0, a_1, \dots, a_m , seu mínimo ocorre em seu ponto crítico. Em outras palavras, precisamos encontrar o ponto $p \in \mathbb{R}^n$ no

qual o gradiente

$$\nabla \delta = \left(\frac{\partial \delta}{\partial a_0}, \dots, \frac{\partial \delta}{\partial a_m} \right)$$

se anula. Contudo, pela regra da cadeia,

$$\frac{\partial \delta}{\partial a_j} = 2x_i^j \sum_{i=0}^n (F(x_i) - y_i),$$

para $0 \leq j \leq m$, porque

$$\frac{\partial F(x_i)}{\partial a_j} = \frac{\partial}{\partial a_j} \left(\sum_{j=0}^m a_j x_i^j \right) = x_i^j.$$

Assim, os valores de a_0, a_1, \dots, a_m nos quais o gradiente de δ se anula correspondem à solução do sistema

$$\sum_{i=0}^n (F(x_i) - y_i) x_i^j = 0 \quad \text{para todo} \quad 0 \leq i \leq n;$$

que podemos reescrever na forma

$$(74) \quad \sum_{i=0}^n x_i^j F(x_i) = \sum_{i=0}^n y_i x_i^j \quad \text{para todo} \quad 0 \leq i \leq n.$$

Como

$$(75) \quad F(x_i) = a_m x_i^m + \dots + a_1 x_i + a_0$$

é uma função linear dos coeficientes a_0, a_1, \dots, a_m , as equações (74) descrevem um sistema linear nestas variáveis. Para explicitar este sistema basta substituir (75) em (74). Contudo, as equações obtidas desta maneira têm uma forma bastante complicada e difícil de lembrar. Felizmente é possível reescrever todo o sistema de maneira mais compacta como uma equação matricial.

Nosso ponto de partida para obter a expressão matricial de (74) é a matriz de Vandermonde

$$V = \begin{bmatrix} 1 & x_0 & \cdots & x_0^m \\ 1 & x_1 & \cdots & x_1^m \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^m \end{bmatrix}.$$

Como m não é necessariamente igual a n , esta matriz pode não ser quadrada quadrada. Denotando por ℓ_i a i -ésima linha de V , temos que

$$\ell_i \mathbf{a} = \underbrace{\begin{bmatrix} 1 & x_0 & \cdots & x_0^m \end{bmatrix}}_{\ell_i} \underbrace{\begin{bmatrix} a_0 \\ \vdots \\ a_m \end{bmatrix}}_{\mathbf{a}} = a_0 + a_1 x_i + \cdots + a_m x_i^m = F(x_i),$$

em que \mathbf{a} é a matriz $m \times 1$ cujas entradas são os coeficientes de $F(x)$. Portanto, pela fórmula de multiplicação de matrizes,

$$V\mathbf{a} = \begin{bmatrix} \ell_0 \mathbf{a} \\ \vdots \\ \ell_m \mathbf{a} \end{bmatrix} = \begin{bmatrix} F(x_0) \\ \vdots \\ F(x_n) \end{bmatrix}.$$

Por outro lado, o lado esquerdo de (74) pode ser reescrito na forma

$$\sum_{i=0}^n x_i^j F(x_i) = \begin{bmatrix} x_0^j & \cdots & x_n^j \end{bmatrix} \begin{bmatrix} F(x_0) \\ \vdots \\ F(x_n) \end{bmatrix}.$$

Combinando estas duas últimas fórmulas, obtemos

$$\sum_{i=0}^n x_i^j F(x_i) = \begin{bmatrix} x_0^j & \cdots & x_n^j \end{bmatrix} V\mathbf{a}.$$

Como $\begin{bmatrix} x_0^j & \cdots & x_n^j \end{bmatrix}$ é a j -ésima coluna da matriz de Vandermonde V , podemos escrever todo o lado esquerdo do sistema (74) compactamente na forma $V^t V\mathbf{a}$. Entretanto,

$$\sum_{i=0}^n x_i^j y_i = \begin{bmatrix} x_0^j & \cdots & x_n^j \end{bmatrix} \underbrace{\begin{bmatrix} y_0 \\ \vdots \\ y_n \end{bmatrix}}_b.$$

De modo que (74) equivale à equação matricial

$$V^t V\mathbf{a} = V^t b,$$

que é conhecida como *equação normal*. A solução da equação normal nos dá uma matriz com os coeficientes de $F(x)$, resolvendo assim o problema dos mínimos quadrados. O algoritmo resultante está resumido abaixo.

MÉTODO DOS MÍNIMOS QUADRADOS. *Dados um inteiro positivo m e um conjunto*

$$\mathcal{P} = \{(x_0, y_0), \dots, (x_n, y_n)\}$$

de pontos do plano, o algoritmo retorna um polinômio $F(x)$, de grau m , cujo gráfico melhor se ajusta aos pontos de \mathcal{P} .

Etapa 1: *construa as matrizes*

$$V = \begin{bmatrix} 1 & x_0 & \cdots & x_0^m \\ 1 & x_1 & \cdots & x_1^m \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^m \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} a_0 \\ \vdots \\ a_m \end{bmatrix} \quad e \quad b = \begin{bmatrix} y_0 \\ \vdots \\ y_n \end{bmatrix}$$

em que as entradas de \mathbf{a} são variáveis;

Etapa 2: *resolva o sistema linear $V^t V \mathbf{a} = V^t b$;*

Etapa 3: *construa o polinômio $F(x) = a_0 + a_1 x + \cdots + a_m x^m$;*

Etapa 4: *retorne o gráfico de $y = F(x)$ no intervalo $[x_0, x_n]$.*

⚡ Esta maneira de deduzir a equação normal tem a virtude de ser bastante curta, mas não é, de forma alguma completa. Em primeiro lugar, não é claro porque o sistema linear definido pela equação normal deva ter sempre alguma solução. Em segundo lugar, não é claro que esta solução seja sempre um mínimo da função δ . Ainda que não seja difícil esclarecer estes dois pontos de maneira analítica, vamos deixá-los em aberto até chegarmos à próxima seção, onde a equação normal será deduzida de maneira puramente geométrica. A grande vantagem de fazer isto é que o tratamento geométrico leva a explicações mais intuitivas do que as que obteríamos se adotássemos o enfoque desta seção.

Encerraremos aplicando o que aprendemos a alguns exemplos. Para começar, digamos que queremos encontrar a reta que melhor se adapta ao conjunto dos pontos

$$\mathcal{P} = \{(1, 2), (2, 7), (3, 8)\}.$$

Neste caso, o polinômio tem grau um, de modo que

$$V = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} 2 \\ 7 \\ 8 \end{bmatrix} \quad \text{e} \quad \mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix}.$$

Assim,

$$V^t V = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} 17 \\ 40 \end{bmatrix}$$

Resolvendo o sistema $V^t V \mathbf{a} = V^t b$, obtemos

$$\mathbf{a} = \begin{bmatrix} -1/3 \\ 3 \end{bmatrix}.$$

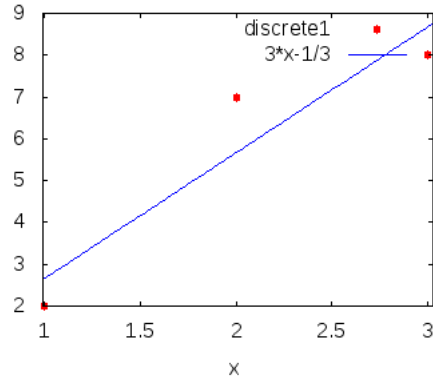
Logo, o polinômio desejado é

$$F(x) = 3x - \frac{1}{3}.$$

A figura abaixo ilustra os pontos de \mathcal{P} e a reta que melhor se ajusta a eles. Note que a reta *não* passa por nenhum dos pontos de \mathcal{P} .

Para nosso segundo exemplo, vamos procurar a curva de grau três que melhor se ajusta aos pontos

$$\mathcal{P} = \{(0, 0.2), (1, 1.3), (2, 8.1), (3, 27.9), (4, 16.8)\}.$$

FIGURA 9. Reta que melhor se ajusta aos pontos $\{(1, 2), (2, 7), (3, 8)\}$

Neste caso,

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \\ 1 & 4 & 16 & 64 \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} 0.2 \\ 1.3 \\ 8.1 \\ 27.9 \\ 16.8 \end{bmatrix},$$

que nos dão o sistema

$$\begin{bmatrix} 5 & 10 & 30 & 100 \\ 10 & 30 & 100 & 354 \\ 30 & 100 & 354 & 1300 \\ 100 & 354 & 1300 & 4890 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 54.3 \\ 168.4 \\ 553.6 \\ 1894.6 \end{bmatrix}$$

cuja solução aproximada é

$$a_0 = 0.9314, \quad a_1 = -16.9929 \quad a_2 = 17.4857 \quad a_3 = -3.05.$$

Portanto, o polinômio de grau três que melhor se ajusta aos pontos de \mathcal{P} é

$$-3.05x^3 + 17.4857x^2 - 16.9929x + 0.9314.$$

Na figura 10 a curva está desenhada em azul e os pontos de \mathcal{P} em vermelho.

Embora, o método dos mínimos quadrados, como exposto nesta seção, aplique-se apenas a funções polinomiais, há muitos problemas que envolvem funções que não são polinomiais, que podem ser reduzidos ao caso que sabemos resolver. Por exemplo, A. J. Lotka propôs, em um artigo [7] publicado em 1926, que a relação entre a quantidade x de artigos científicos publicados em um dado período e a porcentagem y de autores que publicaram x artigos neste período é dada por $y = cx^{-n}$. Os valores de c e n dependem da área de pesquisa que está sendo considerada. A tabela abaixo, cujos dados foram extraídos de [7], dá a relação entre x e y para artigos importantes de física até 1900:

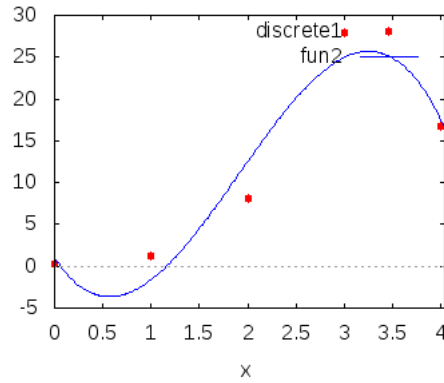


FIGURA 10. Curva de grau três que melhor se ajusta aos pontos $\{(0, 0.2), (1, 1.3), (2, 8.1), (3, 27.9), (4, 16.8)\}$

x	1	2	4	8
y	60.79	15.20	3.80	0.95

Embora a lei de Lotka seja definida por uma função exponencial, podemos usar mínimos quadrados para achar os valores de n e c . Para isto, é necessário reduzir a exponencial a uma função polinomial, o que faremos calculando o logaritmo de base 2 dos dois lados da equação $y = cx^{-n}$, que nos dá

$$\log_2(y) = \log_2(c) - n \log_2(x).$$

A escolha da base dois foi feita apenas porque isto simplifica os cálculos, já que os valores de x , na tabela acima, são todos potências de 2. Com isto $\log_2(y)$ é uma função linear de $\log_2(x)$, o que nos permite usar mínimos quadrados para achar $\log_2(c)$ e n . Para facilitar os cálculos, começaremos tabelando $\log_2(y)$ como função de $\log_2(x)$:

$\log_2(x)$	0	1	2	3
$\log_2(y)$	5.93	3.93	1.93	-0.07

As matrizes de Vandermonde e dos termos constantes, construídas a partir dos dados desta tabela, são

$$V = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} 5.93 \\ 3.93 \\ 1.93 \\ -0.07 \end{bmatrix}.$$

Escrevendo $\ell = \log_2(c)$ e levando em conta que

$$V^t V = \begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix} \quad \text{e} \quad V^t b = \begin{bmatrix} 11.8 \\ 7.59 \end{bmatrix},$$

precisamos apenas resolver o sistema

$$\begin{aligned} -6.0n + 4.0\ell &= 11.8 \\ -14.0n + 6.0\ell &= 7.59, \end{aligned}$$

cujas soluções são

$$n = 2.02 \quad \text{e} \quad \ell = 5.96.$$

Mas isto significa que $\log_2(c) = 5.96$, de modo que

$$c = 2^{5.96} = 62.3.$$

Logo, a fórmula da lei de Lotka neste caso é

$$(76) \quad y = 62.3x^{-2.02}.$$

Em particular, quando $x = 10$,

$$y = 62.3 \cdot 10^{-2.02} = 0.59.$$

Para falar a verdade, a tabela dada em [7] contém o valor exato de y quando $x = 10$, que é 0.61. Com isso, o erro relativo cometido quando usamos (76) para estimar $y(10)$ é 0.03. Você encontrará vários outros exemplos do mesmo tipo de técnicas nos exercícios deste capítulo.

4. Mínimos quadrados: o enfoque geométrico

Nesta seção veremos como traduzir o problema dos mínimos quadrados em um problema de geometria, a partir do qual deduziremos a equação normal. Esta maneira de obter a equação normal tem a grande vantagem de tornar mais fácil visualizar a maneira pela qual a função (73) é minimizada. Ainda que, à primeira vista, esta seção contenha apenas uma demonstração diferente da equação normal, na verdade ela completa o que foi feito anteriormente, porque nos permite deduzir que a solução da equação normal é sempre única e representa o mínimo da função δ .

O primeiro passo para proceder à tradução geométrica é lembrar que (73) corresponde ao quadrado da distância entre os pontos

$$\begin{bmatrix} F(x_0) \\ \vdots \\ F(x_n) \end{bmatrix} \quad \text{e} \quad \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

do espaço \mathbb{R}^{n+1} . Porém, como vimos na seção 2, se

$$F(x) = a_m x^m + \cdots + a_1 x + a_0,$$

então,

$$\begin{bmatrix} F(x_0) \\ \vdots \\ F(x_n) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & x_0 & \cdots & x_0^m \\ 1 & x_1 & \cdots & x_1^m \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^m \end{bmatrix}}_V \cdot \underbrace{\begin{bmatrix} a_0 \\ \vdots \\ a_m \end{bmatrix}}_{\mathbf{a}}$$

Contudo, desta vez, a matriz de Vandermonde V não é necessariamente quadrada, pois podemos ter $m \neq n$. Denotando por b a matriz

$$\begin{bmatrix} y_0 \\ \vdots \\ y_n \end{bmatrix},$$

das ordenadas dos pontos de \mathcal{P} e identificando vetores do \mathbb{R}^n com matrizes coluna, como é usual, podemos reformular o problema dos mínimos quadrados em termos de vetores.

PROBLEMA DOS MÍNIMOS QUADRADOS (versão 2). *Dados, um conjunto finito de pontos \mathcal{P} e um inteiro positivo m , determinar $\mathbf{a} \in \mathbb{R}^{m+1}$ de modo que $V\mathbf{a} - b$ tenha a menor norma possível.*

Embora esta versão do problema esteja escrita em termos de vetores, ela ainda nos diz muito pouco sobre sua natureza geométrica; para revelá-la, precisamos considerar o subconjunto

$$S_m(\mathcal{P}) = \{V(\mathcal{P})\mathbf{a} \mid \mathbf{a} \in \mathbb{R}^n\} \subset \mathbb{R}^{m+1}.$$

Se \mathbf{a}_1 e \mathbf{a}_2 são dois vetores quaisquer do \mathbb{R}^{n+1} , então

$$V\mathbf{a}_1 + V\mathbf{a}_2 = V(\mathbf{a}_1 + \mathbf{a}_2),$$

de modo que a soma de dois vetores de $S_m(\mathcal{P})$ também é um vetor de $S_m(\mathcal{P})$. Como o mesmo vale para o produto de um vetor de $S_m(\mathcal{P})$ por um escalar, podemos concluir que $S_m(\mathcal{P})$ tem a estrutura de um subespaço vetorial do \mathbb{R}^n . Com isto, estamos prontos para formular uma versão verdadeiramente geométrica do problema cuja solução estamos buscando.

PROBLEMA DOS MÍNIMOS QUADRADOS (versão 3). *Dados, um conjunto finito de pontos \mathcal{P} e um inteiro positivo m , determinar o ponto de $S_m(\mathcal{P})$ mais próximo de b .*

A bem da verdade, o vetor que é solução do problema acima não é exatamente o que queremos. Contudo, como ele pertence a $S_m(\mathcal{P})$, podemos escrevê-lo na forma Vw_0 para algum vetor $w_0 \in \mathbb{R}^{m+1}$ e este último é o vetor cujas entradas são os coeficientes de P .

Se \mathcal{P} contivesse apenas três pontos, $S_m(\mathcal{P})$ seria um subespaço vetorial do \mathbb{R}^3 . Consequentemente teria que ser um plano, se m fosse igual a um, ou o \mathbb{R}^3 inteiro, se

m fosse igual a dois. Digamos, para fixar as ideias, que $m = 1$. Neste caso saberíamos resolver a versão 3 do problema dos mínimos quadrados, porque o ponto do plano $S_1(\mathcal{P})$ mais próximo de b é o pé da perpendicular a $S_1(\mathcal{P})$ por b ; isto é, é o ponto em que a reta perpendicular a $S_1(\mathcal{P})$ por b intersecta o próprio plano $S_1(\mathcal{P})$. Por exemplo, se os pontos forem

$$\mathcal{P} = \{(1, 2), (2, 7), (3, 8)\},$$

então

$$V = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$$

e o plano $S_2(\mathcal{P})$ é o conjunto dos vetores da forma

$$V\mathbf{a} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = a_0 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + a_1 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

O produto vetorial de $v_1 = (1, 1, 1)$ por $v_2 = (1, 2, 3)$ é o vetor $v_1 \times v_2 = (1, -2, 1)$, que é perpendicular a v_1 e a v_2 . Logo, neste exemplo, a reta por b , ortogonal ao plano $S_1(\mathcal{P})$ é aquela que tem equação paramétrica $(2, 7, 8) + t(1, -2, 1)$. Por outro lado, como $S_1(\mathcal{P})$ é um plano pela origem cujo vetor normal é $(1, -2, 1)$, sua equação cartesiana é $x - 2y + z = 0$. Substituindo nesta última equação as coordenadas dos pontos da reta escritas em função de t , obtemos

$$0 = (2 + t) - 2(7 - 2t) + (8 + t) = 6t - 4,$$

donde $t = 2/3$. Portanto, o ponto de $S_1(\mathcal{P})$ mais próximo de $b = (2, 7, 8)$ é

$$(2, 7, 8) + \frac{2}{3}(1, -2, 1) = \frac{1}{3}(8, 17, 26).$$

Contudo, as entradas deste último vetor não são os coeficientes do polinômio F . Para achar estes coeficientes precisamos resolver o sistema

$$(77) \quad \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 8 \\ 17 \\ 26 \end{bmatrix}$$

Fazendo isto, obtemos

$$a_0 = -\frac{1}{3} \quad \text{e} \quad a_1 = 3,$$

de forma que o polinômio desejado é

$$F(x) = -\frac{1}{3} + 3x,$$

como já havíamos visto na seção 3.

Um sistema linear que tem mais equações do que variáveis, como é o caso de (77), pode ser impossível. Contudo, isto não pode acontecer com os sistemas que precisaremos resolver para achar o polinômio $F(x)$, porque qualquer vetor que é solução do problema dos mínimos quadrados é da forma Vw , para algum vetor $w \in \mathbb{R}^{m+1}$.

O exemplo parece sugerir que, para resolver o problema dos mínimos quadrados geral, bastaria escolher o ponto em $S_m(\mathcal{P})$ que pertence à reta perpendicular a este subespaço que passa por (y_0, \dots, y_n) . Mas será que isto vale em espaços de dimensão maior que três? A pergunta não é sem razão, porque nossa intuição geométrica frequentemente nos deixa na mão quando é aplicada a estes espaços. Por exemplo, dois planos do \mathbb{R}^3 ou são paralelos ou se intersectam em uma reta, mas isto não vale no \mathbb{R}^4 , no qual dois planos podem se intersectar em um único ponto; veja exercício 5. A saída é dar uma demonstração do resultado de que precisamos baseada apenas nas propriedades básicas do \mathbb{R}^n .

TEOREMA 1. *Se S é um subespaço vetorial do \mathbb{R}^n e $p \in \mathbb{R}^n$, então o ponto de S mais próximo de p é o pé da reta por p que é perpendicular a S .*

DEMONSTRAÇÃO. Seja s_0 um ponto de S tal que o vetor $s_0 - p$ é ortogonal a *todos* os pontos de S e seja s_1 um outro vetor qualquer de S . Teremos provado o teorema se formos capazes de mostrar que $\|s_1 - p\| \geq \|s_0 - p\|$. Como a norma de um vetor é um número real positivo, esta desigualdade é equivalente a $\|s_1 - p\|^2 \geq \|s_0 - p\|^2$. A vantagem de trabalhar com o quadrado da norma é que podemos expressá-lo em termos do produto interno entre vetores; assim,

$$\|s_1 - p\|^2 = \langle s_1 - p | s_1 - p \rangle.$$

Mas, se $e = s_1 - s_0$, então

$$s_1 - p = (s_1 - s_0) + (s_0 - p).$$

Aplicando a

$$\langle s_1 - p | s_1 - p \rangle = \langle s_0 - p + e | s_0 - p + e \rangle$$

as propriedades do produto interno, obtemos

$$(78) \quad \langle s_1 - p | s_1 - p \rangle = \langle s_0 - p | s_0 - p \rangle - 2\langle s_0 - p | e \rangle + \langle e | e \rangle.$$

Como s_0 e s_1 pertencem ao subespaço S , sua diferença e também pertence a S . Mas, $s_0 - p$ é, por hipótese, ortogonal a todos os vetores de S , de modo que

$$\langle s_0 - p | e \rangle = 0.$$

Substituindo isto em (78), temos que

$$\langle s_1 - p | s_1 - p \rangle = \langle s_0 - p | s_0 - p \rangle + \langle e | e \rangle,$$

donde,

$$\|s_1 - p\|^2 = \|s_0 - p\|^2 + \|e\|^2 \geq \|s_0 - p\|^2$$

pois $\|e\|^2 \geq 0$; que é o que queríamos provar. \square

Para referência futura, enunciamos abaixo a forma na qual usaremos o resultado do teorema que acabamos de provar.

TEOREMA 2. *Dados, um conjunto finito de pontos \mathcal{P} e um inteiro positivo m , o ponto de $S_m(\mathcal{P})$ mais próximo de b é o pé da perpendicular a $S_m(\mathcal{P})$ por b .*

Para reaver a equação normal a partir do teorema 2 precisamos trabalhar um pouco mais. Como na seção anterior,

$$\mathcal{P} = \{(x_0, y_0), \dots, (x_n, y_n)\}$$

é o conjunto dos pontos aos quais queremos ajustar o gráfico de um polinômio

$$F(x) = a_m x^m + \dots + a_1 x + a_0,$$

de grau m . De acordo com o método dos mínimos quadrados, isto é feito escolhendo

$$\mathbf{a} = \begin{bmatrix} a_0 \\ \vdots \\ a_m \end{bmatrix} \in \mathbb{R}^{m+1}$$

de modo que minimize a distância entre $V\mathbf{a}$ e b , em que

$$V = \begin{bmatrix} 1 & x_0 & \cdots & x_0^m \\ 1 & x_1 & \cdots & x_1^m \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^m \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} y_0 \\ \vdots \\ y_n \end{bmatrix}.$$

Contudo, pelo teorema 2, o valor de \mathbf{a} é aquele para o qual $V\mathbf{a} - b$ é ortogonal a todos os vetores do subespaço

$$S_m(\mathcal{P}) = \{V\mathbf{a} \mid \mathbf{a} \in \mathbb{R}^n\}.$$

Mas isto equivale a dizer que

$$\langle Vw \mid V\mathbf{a} - b \rangle = 0, \quad \text{para todo } w \in \mathbb{R}^n,$$

que pode ser reescrito, em forma matricial, como

$$w^t V^t (V\mathbf{a} - b) = 0, \quad \text{para todo } w \in \mathbb{R}^n.$$

Contudo, este produto de matrizes só pode ser nulo *para todo* $w \in \mathbb{R}^n$ se a matriz $V^t(V\mathbf{a} - b)$ for igual a zero; isto é, se

$$V^t V\mathbf{a} = V^t b.$$

Temos, assim, uma dedução puramente geométrica da equação normal, que já havíamos obtido na seção 3 usando métodos analíticos. Uma consequência importante do enfoque geométrico é que ele explica porque a equação normal sempre tem uma única solução. Isto ocorre porque, pelo teorema 2, a solução desta equação corresponde ao único ponto que é o pé da perpendicular a $S_m(\mathcal{P})$ por b . O mesmo teorema mostra que a solução da equação normal minimiza a função δ .

Exercícios

1. A tabela abaixo contém alguns valores da função $f : [-2, 2] \rightarrow \mathbb{R}$. Use interpolação para achar uma aproximação para a raiz de f no intervalo $[-2, 2]$.

x	-1.2	0.3	1.1
$f(x)$	-5.76	-5.61	-3.69

2. A tabela abaixo resulta da medida da densidade do ar ρ em várias altitudes h . Use interpolação para encontrar o polinômio quadrático que modela a variação de densidade com a altitude.

$h(\text{km})$	0	3	6
$\rho(\text{kg/m}^3)$	1.225	0.905	0.652

3. Use uma calculadora para determinar os valores do cosseno nos pontos 0.8, 0.9, 1.0, 1.1, .2 e 1.3. Use estes dados e interpolação para achar um polinômio de grau 3 que lhe permita calcular $\cos(1.07)$.
4. Sejam $v_1 = (1, 1, 2)$ e $v_2 = (2, -1, 1)$ vetores do \mathbb{R}^3 . Determine:
- um vetor perpendicular a v_1 e v_2 ;
 - a equação cartesiana do plano π que contém os vetores v_1 e v_2 ;
 - as equações paramétricas da reta que é perpendicular a π e passa pelo ponto $(1, 1, 1)$;
 - o ponto do plano que está mais próximo de $(1, 1, 1)$.
5. Mostre que os planos do \mathbb{R}^4 definidos pelas equações $x = y = 0$ e $y = z = 0$ se intersectam apenas na origem.
6. Considere os dados da tabela abaixo:

x	1	3	5	8
y	2	3	6	7

- Determine a reta que melhor se adapta a estes dados.
- Esboce a reta e os pontos dados.
- Calcule os erros e_i cometidos ao aproximar y_i pelo ponto correspondente da reta que você obteve em (a).
- Calcule a soma dos quadrados dos erros encontrados em (c).

7. Calcule o polinômio quadrático cujo gráfico melhor se ajusta aos pontos $(1, -2)$, $(0, -1)$, $(1, 0)$ e $(2, 4)$.
8. Determine o polinômio linear e o polinômio quadrático cujos gráficos melhor se adaptam aos pontos da tabela abaixo.

x	-3	-1	0	1	3
y	3	2	1	-1	-4

9. Mostre que a reta que melhor se ajusta aos pontos $(x_0, y_0), \dots, (x_n, y_n)$ passa pelo ponto (\bar{x}, \bar{y}) , em que \bar{x} é a média das abscissas e \bar{y} a média das ordenadas dos pontos dados.
10. Considere os dados da seguinte tabela

x	2.5	3.5	5	6	7.5	10	12.5	15	17.5	20
y	13	11	8.5	8.2	7	6.2	5.2	4.8	4.6	4.3

- (a) Ache a curva da forma $y = ax^b$ que melhor se ajusta a estes dados.
- (b) Use a curva obtida em (a) para calcular uma aproximação de y em $x = 9$.

Sugestão: Use ajuste polinomial para encontrar os valores de $\ln(a)$ e b que melhor se ajustam a $\ln(y) = \ln(a) + b \ln(x)$. Note que para isso é necessário calcular a tabela que relaciona $\ln(y)$ a $\ln(x)$.

11. Os dados abaixo foram obtidos monitorando a concentração da bactéria *E. coli* em uma certa praia, depois de uma tempestade:

t	4	8	12	16	20	24
c	1590	1320	1000	900	650	560

Nesta tabela, t corresponde ao tempo em horas transcorrido a partir do final da tempestade e c à concentração de bactérias em UFC/1000 ml (UFC = unidade de formação de colônias).

- (a) Calcule a reta e a curva da forma $y = at^b$ que melhor aproximam estes dados e mostre que produzem resultados impossíveis.
- (b) Determine a exponencial $y = a \exp(bt)$ que melhor se ajusta a estes dados.
- (c) Use a curva obtida em (b) para estimar a concentração de bactérias no início da tempestade.
- (d) Use a curva obtida em (b) para estimar depois de quantas horas a concentração atingirá 200 UFC/1000 ml.

Parte 2

Métodos iterativos e problemas de valor inicial

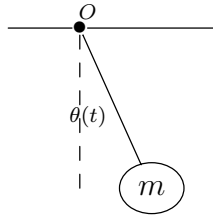
CAPÍTULO 6

O pêndulo simples

Iniciaremos nosso estudo analisando um sistema mecânica típico, o pêndulo. Depois de deduzir a equação diferencial que descreve o movimento do pêndulo, estudaremos o comportamento de suas soluções. Embora este seja um tema tradicional, o foco nos cursos elementares costuma recair sobre o caso de pequenas oscilações, porque leva a uma equação solúvel analiticamente. Contudo, nosso objetivo é tratar precisamente do caso em que a equação não tem solução analítica em termos de funções elementares, porque este caso somos obrigados a recorrer a métodos numéricos.

1. O pêndulo

O pêndulo que vamos analisar consiste de uma haste rígida de comprimento ℓ , que pode girar livremente em torno de uma de suas extremidades e que tem uma bola de massa m presa à outra extremidade. Denotaremos por $\theta(t)$ o ângulo que a haste forma com a vertical num instante de tempo t , como ilustrado na figura abaixo.



Suporemos que este pêndulo não está sujeito a atrito ou resistência do ar, de modo que oscilará indefinidamente. Há várias maneiras de chegar à equação diferencial que descreve o movimento desse pêndulo. A mais comum analisa as forças às quais o pêndulo está sujeito, mas a que vamos utilizar é baseada no princípio de conservação da energia.

Se medirmos o potencial a partir da linha horizontal pelo ponto mais baixo em que a bola do pêndulo passa, teremos que, no tempo t , a distância da bola a esta linha será $\ell - \ell \cos(\theta(t))$. Como a haste tem comprimento constante, a velocidade da bola em t será igual a $\ell \dot{\theta}(t)$, em que usamos um ponto acima da variável para denotar

a derivada relativamente ao tempo. Portanto, a energia total do pêndulo será igual a

$$(79) \quad \frac{1}{2}m(\ell\dot{\theta}(t))^2 + mg\ell(1 - \cos(\theta(t))),$$

em que m é o peso da bola e g é a aceleração da gravidade. A fórmula acima pressupõe que a haste do pêndulo é muito mais leve que a bola, a ponto de nos permitir ignorá-la. Como também estamos supondo que não há atrito, nem resistência do ar, a energia total do pêndulo é constante. Denotando-a por E , podemos resolver a equação (79) relativamente a $\dot{\theta}(t)$, obtendo

$$\dot{\theta}(t) = \sqrt{\frac{2E}{\ell^2 m} - \frac{2g}{\ell}(1 - \cos(\theta(t)))}.$$

É possível simplificar um pouco mais esta equação se tomarmos m como sendo a unidade de massa, ℓ como sendo a unidade de comprimento e $\sqrt{\ell/g}$ como sendo a unidade de tempo. Fazendo isto, a equação acima torna-se

$$(80) \quad \dot{\theta}(t) = \sqrt{2(E - 1 + \cos(\theta(t)))} \quad \text{em que} \quad \theta(0) = 0,$$

porque estamos supondo que o pêndulo está partindo da vertical com a bola para baixo. Antes de dar por terminado nosso modelo, devemos decidir o que fez com que nosso pêndulo se movesse. Digamos que isso ocorreu por causa de um peteleco que imprimiu à bola a velocidade v_0 quando o pêndulo estava parado, com a haste ao longo da vertical. Como neste ponto a energia potencial é nula, a energia total será igual à energia cinética; isto é,

$$E = \frac{1}{2}v_0^2.$$

Quando E é pequeno, o ângulo θ fica próximo de zero. Neste caso, obtemos da fórmula de Taylor que

$$(81) \quad \cos(\theta) \approx 1 - \frac{\theta^2}{2}.$$

Substituindo isto em (80),

$$\dot{\theta}(t) = \sqrt{E - \theta(t)^2}, \quad \text{em que} \quad \theta(0) = 0.$$

Fazendo

$$\theta(t) = \sqrt{E}\alpha(t),$$

nesta última equação, obtemos

$$\dot{\alpha}(t) = \sqrt{1 - \alpha(t)^2}, \quad \text{em que} \quad \alpha(0) = 0;$$

que equivale a

$$\frac{\dot{\alpha}(t)}{\sqrt{1 - \alpha(t)^2}} = \frac{1}{E}.$$

Integrando os dois lados relativamente a t e levando em conta que

$$\int \frac{\dot{\alpha}(t)}{\sqrt{1 - \alpha(t)^2}} dt = \arccos(\alpha(t)),$$

obtemos

$$\arccos(\alpha(t)) = \frac{t}{E} + c;$$

donde

$$\alpha(t) = \cos\left(\frac{t}{E} + c\right).$$

Levando em conta que $\theta(t) = \sqrt{E}\alpha(t)$ e que $\theta(0) = 0$, obtemos

$$\theta(t) = \frac{1}{\sqrt{E}}\left(\cos\left(\frac{t}{E} + \frac{\pi}{2}\right)\right).$$

Quando v_0 não é pequeno, não podemos usar a aproximação (81). Como a equação (80) não pode ser integrada em termos das funções elementares do cálculo, resta-nos descobrir como integrá-la numericamente. Contudo, como veremos na seção 4, é possível integrar (80) se juntarmos as funções elípticas às funções elementares.

2. Problemas de valor inicial e o método de Euler

O problema do pêndulo, como formulado em (80), é um exemplo do que os matemáticos chamam de problema de valor inicial. No caso em que a equação diferencial $\dot{y} = f(t, y)$, é de primeira ordem, basta que conheçamos o valor α que a solução $y = y(t)$ toma no momento inicial. Supondo que começamos a contar o tempo a partir de um valor t_0 , podemos formular o *problema de valor inicial* neste caso por

$$(82) \quad \dot{y} = f(t, y) \quad \text{e} \quad y(t_0) = y_0.$$

Nesta seção descreveremos o mais antigo e mais simples dos métodos numéricos para resolver problemas de valor inicial, o método de Euler.

Digamos que queremos achar a solução do problema (82) no intervalo $[a, b]$. Começamos por dividir este intervalo em n partes iguais. O valor de n a ser escolhido depende, naturalmente, da precisão desejada. Sejam

$$h = \frac{b - a}{n}, \quad t_k = a + kh \quad \text{e} \quad y_k = y(t_k), \quad \text{em que} \quad 0 \leq k \leq n.$$

Como no caso dos problemas de valor de contorno, vamos aproximar $y'(t_k)$ por uma diferença finita que, neste caso, será

$$(83) \quad \frac{y(t_{k+1}) - y(t_k)}{h} = \frac{y_{k+1} - y_k}{h}.$$

Sabemos, da equação $\dot{y} = f(t, y)$, que

$$\dot{y}(t_k) = f(t_k, y(t_k)) = f(t_k, y_k).$$

Substituindo a aproximação dada pela diferença finita (83) nesta última equação, obtemos

$$\frac{y_{k+1} - y_k}{h} = f(t_k, y_k),$$

que pode ser reescrita como a recorrência,

$$(84) \quad y_{k+1} = y_k + hf(t_k, y_k) \quad \text{e} \quad y_0 = y(0) = \alpha.$$

Por exemplo, no problema de valor inicial

$$\dot{y} = y \quad \text{e} \quad y(0) = 1,$$

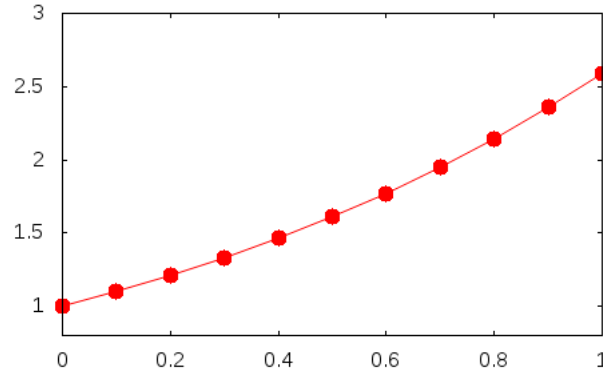
a função $f(t, y)$ é igual a y , de modo que a recorrência (84) será, neste caso,

$$(85) \quad y_{k+1} = y_k + hy_k = y_k(1 + h) \quad \text{e} \quad y_0 = 1.$$

Aplicando esta recorrência para $t \in [0, 1]$ e $n = 10$, obtemos

k	t_k	y_k
0	0	1
1	0.1	$1 + 0.1 = 1.1$
2	0.2	$1.1 \cdot 1.1 = 1.21$
3	0.3	$1.21 \cdot 1.1 = 1.33$
4	0.4	$1.33 \cdot 1.1 = 1.46$
5	0.5	$1.46 \cdot 1.1 = 1.61$
6	0.6	$1.61 \cdot 1.1 = 1.77$
7	0.7	$1.77 \cdot 1.1 = 1.95$
8	0.8	$1.95 \cdot 1.1 = 2.14$
9	0.9	$2.14 \cdot 1.1 = 2.36$
10	1.0	$2.36 \cdot 1.1 = 2.59.$

A figura resultante quando plotamos os resultados em um gráfico, com os pontos ligados, é a seguinte.

FIGURA 1. Solução numérica de $\dot{y} = y$ com $y(0) = 1$.

Para ser honesto, a recorrência (85) é tão simples que podemos achar uma fórmula fechada simplesmente iterando o processo, como abaixo:

$$\begin{aligned}
 y_n &= y_{n-1}(1+h) \\
 &= \underbrace{(y_{n-2}(1+h))}_{y_{n-1}}(1+h) = y_{n-2}(1+h)^2 \\
 &= \underbrace{(y_{n-3}(1+h))}_{y_{n-2}}(1+h)^2 = y_{n-3}(1+h)^3 \\
 &\vdots \\
 &= \underbrace{(y_0(1+h))}_{y_1}(1+h)^{n-1} = y_0(1+h)^n.
 \end{aligned}$$

Levando em conta a condição inicial, obtemos

$$y_n = (1+h)^n.$$

Supondo, como acima, que $[0, 1]$ é o intervalo em que desejamos resolver o problema, então $h = 1/n$, e a fórmula para y_n pode ser reescrita na forma

$$y_n = \left(1 + \frac{1}{n}\right)^n.$$

Note que $\lim_{n \rightarrow \infty} y_n = e$, como seria de esperar.

Na prática, raro é o caso em que, como no exemplo acima, a recorrência produz uma fórmula fechada. Neste sentido, um exemplo mais típico seria

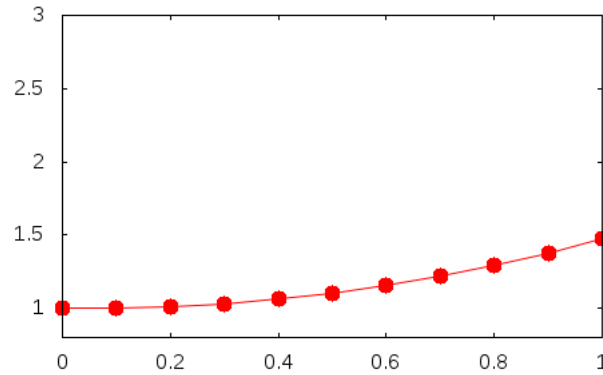
$$\dot{y} = \sin(yt) \quad \text{com} \quad y(0) = 1.$$

Desta vez a tabela é

k	0	1	2	3	4	5	6	7	8	9	10
t_k	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
y_k	1.0	1.0	1.01	1.03	1.06	1.10	1.15	1.21	1.29	1.37	1.47

TABELA 1. Iteração de Euler aplicada a $\dot{y} = \sin(ty)$ com passo $h = 0.1$.

que corresponde à figura abaixo.

FIGURA 2. Solução numérica de $\dot{y} = \sin(ty)$ com $y(0) = 1$.

O método de Euler admite uma interpretação geométrica que tem a vantagem de dar uma ideia intuitiva do porquê o método deveria funcionar. A ideia por trás desta interpretação remonta diretamente a Newton. Na Proposição II da seção II do Livro I dos *Princípios Matemáticos da Filosofia Natural*, Newton demonstra que qualquer corpo que gire em torno de um centro de força respeita a *Segunda Lei de Kepler*, segundo a qual o segmento de reta que une o corpo ao centro de força percorre áreas iguais em tempos iguais, não importando em que ponto da órbita o corpo esteja. A estratégia de Newton consiste em imaginar que, ao invés de agir continuamente sobre o corpo em órbita, a força atua apenas em momentos discretos, como se fosse ligada apenas por um instante infinitesimal, a intervalos de tempo iguais. Quando não houver nenhuma força atuando sobre o corpo, a lei da inércia garante que ele vai se mover ao longo de uma reta; quando a força agir, o corpo terá um brusco desvio de direção. O resultado é uma órbita poligonal, como ilustrado na figura 3, extraída de uma das primeiras edições do *Principia*.

A ideia do Newton é que, se o impulso ocorre a intervalos cada vez menores, a poligonal resultante se aproxima cada vez mais da curva realmente descrita pelo corpo que está sujeito à força central.

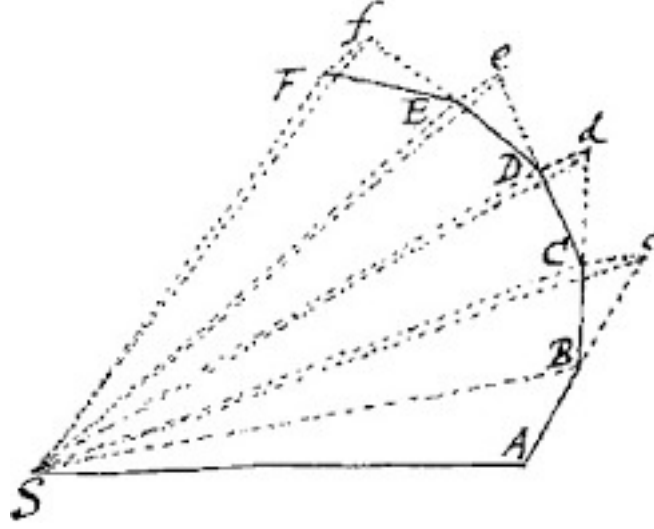


FIGURA 3. Principia: Livro I, Seção II, Proposição II

O *método de Euler* é baseado exatamente na mesma ideia. Digamos que desejamos resolver o problema de valor inicial definido pela equação diferencial

$$\dot{y} = f(t, y)$$

com condição inicial $y(t_0) = \alpha$, no intervalo $[a, b]$. Se supusermos que a condição dada pela equação diferencial só é aplicada nos pontos $t_j = a + jh$, em que

$$h = \frac{b - a}{n},$$

então a órbita será aproximada por uma série de segmentos de retas cujas inclinações correspondem a

$$\dot{y}(t_j) = f(t_j, y(t_j)) \approx f(t_j, y_j).$$

As retas que servem de suporte a estes segmentos têm por equações

$$y - y_j = f(t_j, y_j)(t - t_j).$$

Como o $(j + 1)$ -ésimo segmento começa em (t_j, y_j) e acaba em (t_{j+1}, y_{j+1}) , teremos que

$$y_{j+1} = y_j + f(t_j, y_j)(t_{j+1} - t_j) = y_j + f(t_j, y_j)h,$$

que é a iteração do método de Euler. A figura abaixo ilustra duas etapas consecutivas da aplicação do método de Euler. A curva solução da equação foi desenhada em azul e as tangentes em vermelho. A aproximação da curva entre 0 e 2 (com passo igual a 1) corresponde à linha poligonal $P_1P_2P_3$. O passo foi escolhido intencionalmente grande para facilitar a interpretação da figura.

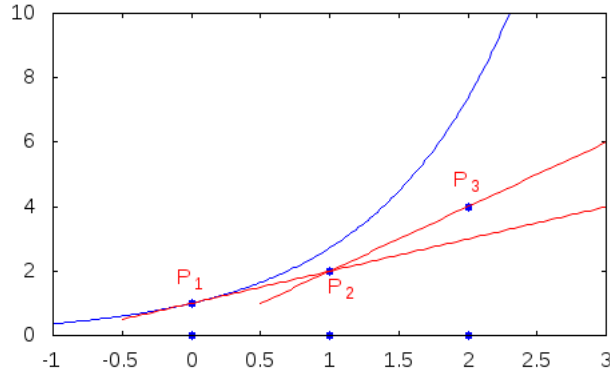


FIGURA 4. O método de Euler

3. Aplicando o método de Euler ao pêndulo

Vejamos como aplicar o método de Euler ao problema de valor inicial (80). Para fixar as ideias escolheremos $E = 3$, de modo que o problema passa a ser

$$(86) \quad \dot{\theta} = \sqrt{4 + 2 \cos(\theta)}, \quad \text{em que} \quad \theta(0) = 0.$$

Adaptando este problema à equação (84) e escolhendo $h = 0.1$, obtemos

$$\theta_{k+1} = \theta_k + 0.1 \cdot \sqrt{4 + 2 \cos(\theta_k)}, \quad \text{e} \quad \theta_0 = 0.$$

Segue disto que,

$$\theta_1 = 0.1 \cdot \sqrt{4 + 2 \cos(0)} = 0.24494897427832,$$

ao passo que

$$\theta_2 = 0.24494897427832 + 0.1 \cdot \sqrt{4 + 2 \cos(0.24494897427832)} \approx 0.48867626861858.$$

Portanto, os três primeiros pontos que estão aproximadamente sobre o gráfico da solução $y = \theta(t)$ de (86) são

$$(0, 0), \quad (0.1, 0.24494897427832) \quad \text{e} \quad (0.2, 0.48867626861858).$$

Ligando os 30 primeiros pontos, obtemos o gráfico ilustrado na figura 5 da página 121.

Se você esperava uma solução que oscilasse, para cima e para baixo, em torno do eixo das abscissas, provavelmente se surpreendeu com a figura 5. Para entender melhor o que está acontecendo vamos considerar o caso em que $E = 2$, que corresponde ao problema de valor inicial

$$(87) \quad \dot{\theta} = \sqrt{2 + 2 \cos(\theta)}, \quad \text{em que} \quad \theta(0) = 0.$$

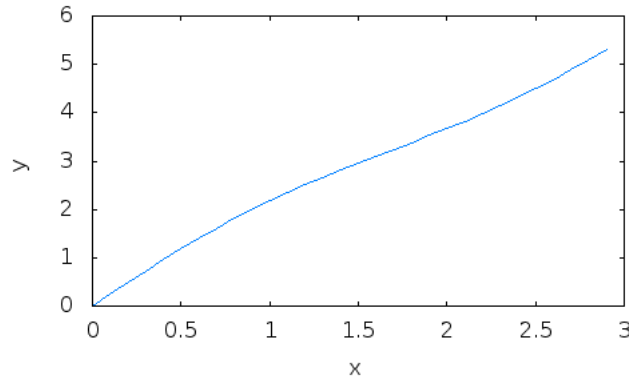


FIGURA 5. Aplicando o método de Euler ao problema (86)

Desta vez, a figura tem o aspecto da figura 6.

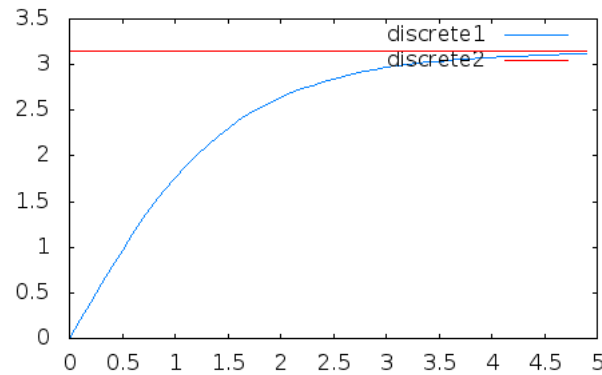


FIGURA 6. Aplicando o método de Euler ao problema (87)

Os pontos da reta horizontal, aos quais a solução é assíntota, têm altura igual a π . Em outras palavras, o pêndulo vai parando quando $\theta(t)$ tende a π . Para entender porque isto está acontecendo, lembre-se que nosso pêndulo tem haste rígida. Portanto, uma pessoa que tenha a mão muito firme conseguiria equilibrá-lo com a haste vertical e a bola *acima* do ponto em torno do qual a haste gira. Assim, quando $E = 2$ o pêndulo tem a energia exata para atingir este ponto de equilíbrio, mas não para levá-lo além dele; fazendo com que pare neste ponto. Entretanto, este ponto é um ponto de equilíbrio instável: embora seja possível fazer com que o pêndulo pare neste ponto, qualquer pequeno desvio da vertical fará com que volte a descer. Em particular, sempre que $E > 2$ o pêndulo terá energia suficiente para passar deste ponto e fazer um giro completo em torno do pivô. Com estamos supondo que não há perda de energia, o pêndulo vai continuar girando em torno do pivô, fazendo com que o ângulo $\theta(t)$ aumente continuamente, como no gráfico da figura 5.

Para fazer o pêndulo oscilar no movimento de vai-e-vem que a palavra pêndulo suscita em nossa mente, precisamos escolher $E < 2$. Por exemplo,

$$(88) \quad \dot{\theta} = \sqrt{2 \cos(\theta) - 1/2}, \quad \text{em que} \quad \theta(0) = 0$$

corresponde a tomar $E = 1/2$ no problema de valor inicial (86). Contudo, iterando

$$\theta_{k+1} = \theta_k + h\sqrt{2 \cos(\theta) - 1/2}$$

16 vezes, obtemos

$$\theta_{16} = 5.172310839363217 \cdot 10^{-4}i + 1.047212996863676)$$

que não é um número real! Isto está acontecendo porque, como

$$\cos(\theta_{15}) = \cos(1.047212996863676) = 0.49998662360029 < 1/2$$

fazendo com que o argumento da raiz quadrada neste momento da iteração seja menor que 0. Portanto, como $\cos(\theta) = 1/2$ corresponde a $\theta = \pi/3$, este é o ângulo máximo que o pêndulo pode atingir quando $v_0 = 1/2$. No momento em que a haste forma um ângulo de $\pi/3$ com vertical, a velocidade do pêndulo se anula, fazendo com que a oscilação se inverta. O método de Euler não está conseguindo detectar o anulamento da velocidade porque $\pi/3$ não aparece como um dos valores de θ ao longo da iteração.

Infelizmente, mesmo que escolhamos h de modo que $\theta_k = \pi/3$, para algum valor de k , o problema não terá sido resolvido. Para entender porque, supohamos que o pêndulo é solto do repouso quando a haste forma um ângulo de $\pi/3$ radianos com a vertical. Neste caso, pela equação (79), a energia total do pêndulo será

$$E = mg\ell(1 - \cos(\theta(t_0))) = mg\ell \left(1 - \cos\left(\frac{\pi}{3}\right)\right) = \frac{1}{2},$$

por causa de nossa escolha de unidades. Mas, substituindo isto em (80), obtemos a equação diferencial em (88). Logo, do ponto de vista desta equação, tanto faz considerar o pêndulo como atingindo $\theta = \pi/3$ na extremidade direita de uma oscilação, iniciada da vertical com velocidade inicial $v_0 = 1$, quanto como se fosse solto do repouso a partir de $\theta(0) = \pi/3$, porque em ambos os casos $E = 1/2$. Mas, aplicando o método de Euler ao problema de valor inicial

$$\dot{\theta} = \sqrt{2 \cos(\theta) - 1} \quad \text{em que} \quad \theta(0) = \pi/3,$$

obtemos a iteração

$$\theta_{k+1} = \theta_k + 0.1 \cdot \sqrt{2 \cos(\theta_k) - 1}, \quad \text{e} \quad \theta_0 = \pi/3.$$

Contudo, como

$$2 \cos\left(\frac{\pi}{3}\right) - 1 = 0,$$

temos que se $\theta_k = \pi/3$, para algum $k \geq 0$, então $\theta_{k+1} = \pi/3$; de modo que o pêndulo não sai do lugar! Curiosamente, isto não teria transparecido se você tivesse estudado a solução desta equação como costuma ser apresentada em um livro de mecânica

clássica. Discutiremos esta solução na próxima seção, ao final da qual estaremos em condições de decifrar completamente porque não fomos capazes de encontrar a solução desejada para (88) usando o método de Euler.

Contudo, nada disto nos impede de usar o método de Euler para gerar os pontos do arco que o pêndulo percorre entre 0 e $\pi/3$. Por exemplo, dividindo o intervalo $[0, \pi/3]$ em 100 partes iguais e começando de $\theta(0) = 0$, obtemos o arco da figura 7 que corresponde a metade daquele que é descrito pelo pêndulo ao longo de suas iterações.

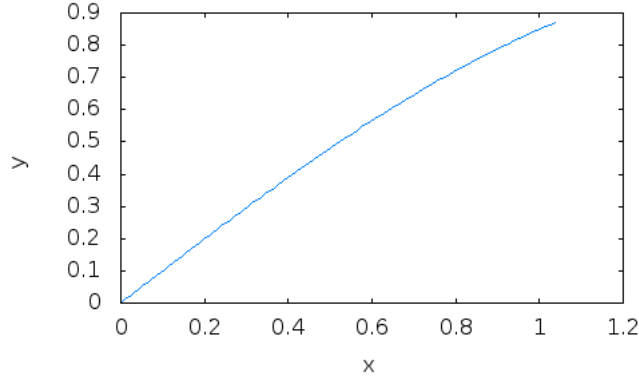


FIGURA 7. Arco da solução do problema de valor inicial (88)

4. A solução geral da equação do pêndulo

Como já observamos anteriormente, é comum encontrar uma solução analítica para a equação do pêndulo em livros de mecânica clássica; por exemplo, [2, pp. 50-54] e [14, pp. 71-74]. Para tornar a abordagem mais concreta, consideraremos apenas o problema de valor inicial

$$(89) \quad \dot{\theta} = \sqrt{2 \cos(\theta) - 1} \quad \text{em que} \quad \theta(0) = \pi/3,$$

com o qual já nos deparamos na seção 3, onde vimos que uma aplicação ingênua do método de Euler a este problema pode levar a soluções com valores imaginários.

O enfoque que você encontrará nos livros de mecânica clássica mencionados acima começa por transformar a equação diferencial em outra cuja solução é conhecida. Lembrando que

$$\cos(\theta) = 1 - 2 \sin^2 \left(\frac{\theta}{2} \right),$$

podemos escrever a equação diferencial em (89) na forma

$$(90) \quad \dot{\theta} = \sqrt{1 - 4 \sin^2 \left(\frac{\theta}{2} \right)}.$$

Mas, tomando

$$(91) \quad y = 2 \operatorname{sen} \left(\frac{\theta}{2} \right)$$

e lembrando que θ , e portanto y , são funções de t , temos que

$$\dot{y} = \cos \left(\frac{\theta}{2} \right) \dot{\theta}$$

donde,

$$\dot{y} = \left(\sqrt{1 - \operatorname{sen}^2 \left(\frac{\theta}{2} \right)} \right) \dot{\theta} = \left(\sqrt{1 - \frac{y^2}{4}} \right) \dot{\theta}.$$

Logo,

$$(92) \quad \dot{\theta} = \frac{\dot{y}}{\sqrt{1 - \frac{y^2}{4}}}.$$

Substituindo (91) e (92) em (90), obtemos

$$\frac{\dot{y}}{\sqrt{1 - \frac{y^2}{4}}} = \sqrt{1 - y^2};$$

que equivale a

$$\dot{y} = \sqrt{(1 - y^2)(1 - y^2/4)}.$$

Mas, como estamos supondo que o pêndulo partiu da vertical,

$$y(0) = \operatorname{sen}(0) = 0.$$

Logo, o problema de valor inicial (89) é equivalente a

$$(93) \quad \dot{y} = \frac{1}{2} \sqrt{(1 - y^2)(1 - y^2/4)} \quad \text{e} \quad y(0) = 0.$$

Para resolver este problema, dividimos os dois lados da equação diferencial por $\sqrt{(1 - y^2)(1 - y^2/4)}$ e integramos a equação resultante, obtendo

$$(94) \quad \int_0^y \frac{dz}{\sqrt{(1 - z^2)(1 - z^2/4)}} = t.$$

Por analogia com o fato de que, se

$$\int_0^y \frac{dz}{\sqrt{1 - z^2}} = w,$$

então $y = \operatorname{sen}(w)$, dizemos que, se

$$\int_0^y \frac{dz}{\sqrt{(1 - k^2 z^2)(1 - z^2)}} = u,$$

então $y = \operatorname{sn}(u, k)$ é o *seno elíptico* de u . Usando esta terminologia, (94) toma a forma

$$y = \operatorname{sn}(t, 1/2),$$

que é a solução do problema de valor inicial (93). O gráfico desta solução, desenhado com a ajuda do *Maxima*, é ilustrado na figura 8 da página 125.

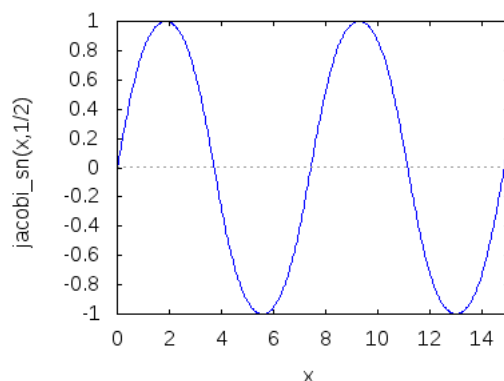


FIGURA 8. Solução do problema de valor inicial (93)

Como você pode ver, desta vez o gráfico oscila em torno do eixo das abscissas, como seria natural de esperar do movimento de um pêndulo. Isto nos trás de volta ao que aconteceu quando aplicamos o método de Euler ao problema de valor inicial (88). Por que, em vez da solução periódica ilustrada na figura 8, obtivemos como solução um valor constante? A resposta requer que generalizemos um pouco o problema.

O argumento desta seção mostra que a equação diferencial

$$(95) \quad \dot{y} = \sqrt{(1 - y^2)(1 - y^2/4)}$$

tem como solução geral

$$y_c(t, 1/2) = \operatorname{sn}(t + c, 2)$$

em que o valor de c deve ser determinado a partir da condição inicial dada. No entanto, (95) também admite as soluções constantes

$$y = 2 \quad \text{e} \quad y = 1,$$

que não podem ser obtidas a partir da solução geral e que, por isso, são conhecidas como *soluções singulares*. Isto não é de forma alguma atípico; muitas equações diferenciais simples têm soluções singulares. Por exemplo,

$$\dot{y} = y^2 - 1$$

pode ser facilmente integrada. Para isto basta reescrevê-la na forma

$$(96) \quad \frac{\dot{y}}{y^2 - 1} = 1.$$

Mas, usando frações parciais,

$$\int \frac{\dot{y}}{y^2 - 1} = \frac{1}{2} \ln \left(\frac{y-1}{y+1} \right).$$

Logo, integrando os dois lados de (96), obtemos

$$\frac{1}{2} \ln \left(\frac{y-1}{y+1} \right) = 2t + e,$$

em que e é a constante de integração. Tomando $c = \exp(e)$, esta última equação pode ser reescrita na forma

$$\frac{y-1}{y+1} = c \exp(2t).$$

Resolvendo relativamente a y , encontramos

$$y_c(t) = \frac{1 + c \exp(2t)}{1 - c \exp(2t)},$$

que é a solução geral da equação $\dot{y} = y^2 - 1$. Por outro lado, como $y = \pm 1$ anulam o lado direito de $\dot{y} = y^2 - 1$, podemos concluir que $y = 1$ e $y = -1$ também são soluções. Contudo, embora $y = 1$ corresponda ao caso particular da solução geral em que $c = 0$, não existe nenhum valor de c para o qual $y_c(t) = -1$. De fato,

$$\frac{1 + c \exp(2t)}{1 - c \exp(2t)} = -1$$

implica que

$$1 + c \exp(2t) = -(1 - c \exp(2t)),$$

donde $1 = -1$, que é uma contradição. Portanto, $y = -1$ é uma solução singular de $\dot{y} = y^2 - 1$. Como no caso do pêndulo, se aplicarmos o método de Euler ao problema de valor inicial

$$y' = y^2 - 1 \quad \text{com} \quad y(0) = -1$$

obteremos a solução singular $y(t) = -1$ e não a solução geral

$$y_{-1}(t) = \frac{1 - \exp(2t)}{1 + \exp(2t)}.$$

5. Funções e integrais elípticas

Na seção anterior utilizamos funções elípticas para resolver a equação diferencial do pêndulo. Mas, de onde vieram estas funções e por que são chamadas de elípticas? A resposta às duas perguntas é que surgiram originalmente no cálculo do comprimento de um arco de elipse, que John Wallis foi um dos primeiros a tentar em 1655. Por exemplo, a parte da elipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

acima do eixo das abscissas é definida pela função

$$(97) \quad y = b\sqrt{1 - \frac{x^2}{a^2}}.$$

Pela fórmula para comprimento de uma curva, o arco da elipse entre os pontos $(0, b)$ e $(a, 0)$ é igual a

$$\int_0^a \sqrt{1 + \left(\frac{dy}{dx}\right)^2} dx.$$

Mas, de (97),

$$\frac{dy}{dx} = -\frac{bx}{a\sqrt{a^2 - x^2}}.$$

Substituindo isto na fórmula para comprimento de arco, obtemos

$$\int_0^a \sqrt{\frac{(-b^2 + a^2) x^2 - a^4}{a^2 x^2 - a^4}} dx.$$

Tomando $x = au$, esta integral se transforma em

$$a \int_0^1 \sqrt{\frac{1 - k^2 u^2}{1 - u^2}} du,$$

em que

$$k = \sqrt{\frac{a^2 - b^2}{a^2}}.$$

Tradicionalmente

$$E(x, k) = \int_0^x \sqrt{\frac{1 - k^2 u^2}{1 - u^2}} du,$$

é conhecida como uma *integral elíptica de segundo tipo*. Já as integrais da forma

$$F(x, k) = \int_0^x \frac{dz}{\sqrt{(1 - k^2 z^2)(1 - z^2)}},$$

que apareceram em nosso estudo do pêndulo, são chamadas de *integrais elípticas de primeiro tipo*. Essas integrais, e outras semelhantes, surgem frequentemente na solução de muitos problemas de física e engenharia.

Cedo ficou claro aos matemáticos do século XVIII que estas integrais não poderiam ser calculadas em termos das *funções elementares*, que são, basicamente, combinações de polinômios, senos, cossenos, exponenciais e logaritmos através das operações aritméticas básicas, além da radiciação e da composição de funções. Entretanto, uma demonstração correta de que isto é realmente impossível só se tornou viável com o trabalho de Joseph Liouville no século XIX. Portanto, precisamos utilizar métodos numéricos, se quisermos calcular integrais elípticas. A maneira mais ingênua de fazer isto consiste em aproximar estas integrais por somas de retângulos, como fazemos ao definir integrais definidas.

Seja $f : [a, b] \rightarrow \mathbb{R}$ uma função contínua. Usando a terminologia tradicional, a área abaixo do gráfico $y = f(x)$, entre a e b , é definida como o limite quando n tende a infinito da soma

$$\sum_{i=0}^n f(\alpha_i) \frac{b-a}{n} \quad \text{em que} \quad \alpha_i \in \left[a + i \frac{b-a}{n}, a + (i+1) \frac{b-a}{n} \right].$$

Em outras palavras,

$$\int_a^b f(t) dt = \lim_{n \rightarrow \infty} \sum_{i=0}^n f(\alpha_i) \frac{b-a}{n}.$$

Utilizando uma estratégia análoga à que usamos para aproximar derivadas, temos que, para um valor de

$$h = \frac{b-a}{n}$$

suficientemente pequeno,

$$\int_a^b f(t) dt \approx \sum_{i=0}^n f(\alpha_i) h.$$

Naturalmente, quão pequeno h tem que ser, depende do erro máximo que estamos preparados para admitir no cálculo da integral. Para delimitar este erro, podemos usar duas somas, uma para retângulos inscritos, outra para retângulos circunscritos, à curva $y = f(x)$ em cada intervalo em que $[a, b]$ foi subdividido. Escrevendo

$$x_i = a + \frac{ih}{n}$$

e denotando por M_i e m_i o máximo e o mínimo de $f(x)$ no intervalo $[x_i, x_{i+1}]$, teremos que

$$\sum_{i=0}^n m_i h \leq \int_a^b f(t) dt \leq \sum_{i=0}^n M_i h,$$

mas também que

$$\sum_{i=0}^n m_i h \leq \sum_{i=0}^n f(\alpha_i) h \leq \sum_{i=0}^n M_i h,$$

quaisquer que sejam os valores escolhidos para $\alpha_i \in [x_i, x_{i+1}]$. Portanto,

$$\left| \int_a^b f(t) dt - \sum_{i=0}^n f(\alpha_i) h \right| \leq \sum_{i=0}^n (M_i - m_i) h.$$

É claro que, na prática, determinar os valores de M_i e m_i para cada intervalo em que subdividimos $[a, b]$ pode não ser tão simples quanto o argumento acima parece sugerir. Esta maneira de calcular integrais é conhecida como *regra do retângulo*.

Vamos aplicar esta estratégia à integral elíptica de segunda espécie

$$F(1, 1/2) = \int_0^{1/2} \frac{dz}{\sqrt{(1-4z^2)(1-z^2)}}$$

cujo valor corresponde a um quarto do período do pêndulo descrito pela equação (89). A primeira coisa a observar é que a função sob o sinal de integração explode a infinito quando $z = 1/2$ mas, apesar disto, a integral tem um valor finito. Se não fosse este o caso, o pêndulo teria período infinito; mas, evidentemente, isto não constitui uma prova matemática de que a integral é finita. Mais detalhes podem ser encontrados em ????. Ao calcular $F(1, 1/2)$ pela regra do retângulo podemos facilmente evitar este problema definindo a altura de cada retângulo a partir de sua extremidade esquerda. Em outras palavras, se

$$f(z) = \frac{1}{\sqrt{(1-4z^2)(1-z^2)}},$$

aproximamos $F(1, 1/2)$ pela soma

$$\sum_{i=0}^{n-1} f(ih)h.$$

Por exemplo,

$$\sum_{i=0}^{10^7-1} f\left(i \frac{1}{10^7}\right) \frac{1}{10^7} \approx 1.685373342900064$$

Esta não é a melhor maneira de calcular estas integrais. O procedimento usado na prática é descrito na próxima seção, que é independente do resto do livro.

Na seção anterior expressamos a solução da equação do pêndulo em termos da função inversa da integral elíptica de primeira espécie. Assim, $y = \operatorname{sn}(u, k)$ quando

$$\int_0^y \frac{dz}{\sqrt{(1-k^2z^2)(1-z^2)}} = u.$$

Como vimos no início da seção, as integrais elípticas começaram a ser estudadas muito cedo, mas a teoria só tomou a forma atual quando Abel e Jacobi passaram a tomar as funções inversas, e não as integrais, como fundamentais. Essas funções podem ser calculadas usando-se um procedimento recursivo semelhante ao utilizado na próxima seção para calcular as integrais elípticas de primeira espécie.

6. Integrais elípticas de primeira espécie

Nesta seção estudaremos um procedimento recursivo para calcular a integral elíptica

$$(98) \quad F(z, k) = \int_0^1 \frac{dz}{\sqrt{(1 - k^2 z^2)(1 - z^2)}}$$

que usamos na seção 5 para calcular o período do pêndulo definido pelo problema de valor inicial

$$\dot{\theta} = \sqrt{2 \cos(\theta) - 1} \quad \text{em que} \quad \theta(0) = 0.$$

Nosso primeiro passo consiste em aplicar duas transformações à integral (98). A primeira consiste em substituir z por $\sin(t)$, obtendo

$$F(\sin(t), k) = \int_0^{\pi/2} \frac{\cos(t) dt}{\sqrt{(1 - k^2 \sin^2(t))(1 - \sin^2(t))}},$$

que denotaremos por $\hat{F}(t, k)$. Como $\sqrt{1 - \sin^2(t)} = \cos(t)$,

$$(99) \quad \hat{F}(t, k) = \int_0^{\pi/2} \frac{dt}{\sqrt{1 - k^2 \sin^2(t)}}.$$

Para a segunda transformação, tomaremos

$$(100) \quad t = \arctan\left(\frac{\sin(2\omega)}{k + \cos(2\omega)}\right);$$

donde podemos concluir que

$$(101) \quad \frac{\sin(t)}{\cos(t)} = \frac{\sin(2\omega)}{k + \cos(2\omega)}.$$

Elevando os dois lados ao quadrado e levando em conta que $\cos^2(t) = 1 - \sin^2(t)$, obtemos

$$\frac{\sin^2(t)}{1 - \sin^2(t)} = \frac{\sin^2(2\omega)}{(k + \cos(2\omega))^2};$$

donde

$$\sin^2(t) = \frac{\sin^2(2\omega)}{k^2 + 1 + 2k \cos(2\omega)}.$$

Isto nos permite escrever

$$(102) \quad 1 - k^2 \sin^2(t) = \frac{(1 + k \cos(2\omega))^2}{k^2 + 1 + 2k \cos(2\omega)}.$$

Mas, de (100), temos também que

$$(103) \quad dt = 2 \frac{k \cos(2\omega) + 1}{k^2 + 2k \cos(2\omega) + 1} d\omega.$$

Substituindo (102) e (103) em (99) e simplificando o resultado

$$\hat{F}(t, k) = \int_0^{\pi/2} \frac{2d\omega}{\sqrt{k^2 + 2k \cos(2\omega) + 1}} = \int_0^{\pi/2} \frac{2d\omega}{\sqrt{(k+1)^2 - 2k(1 - \cos(2\omega))}}.$$

Mas, levando em conta que

$$\cos(2\omega) = 1 - 2 \sin^2(\omega),$$

podemos escrever

$$\hat{F}(t, k) = \int_0^{\pi/2} \frac{2d\omega}{\sqrt{(k+1)^2 \left(1 - \frac{4k}{(k+1)^2} \sin^2(\omega)\right)}} = \frac{2}{k+1} \int_0^{\pi/2} \frac{2d\omega}{\sqrt{1 - \frac{4k}{(k+1)^2} \sin^2(\omega)}};$$

que equivale à igualdade

$$\hat{F}(t, k) = \frac{2}{k+1} \hat{F}\left(\omega, \frac{2\sqrt{k}}{k+1}\right).$$

Como, por (101),

$$k \sin(t) = \sin(2\omega) \cos(t) - \sin(t) \cos(2\omega) = \sin(2\omega - t),$$

podemos concluir que ω está relacionado a t por

$$(104) \quad \omega = \frac{\arcsen(k \sin(t)) + t}{2}.$$

Levando em conta que, se $k < 1$ então $1 < k+1 < 2$ e $k < \sqrt{k} < 1$, temos que

$$k < \frac{2\sqrt{k}}{k+1}.$$

Por outro lado,

$$0 < (k-1)^2 = k^2 - 2k + 1$$

implica que

$$4k < k^2 + 2k + 1 = (k+1)^2;$$

donde

$$\frac{4k}{(k+1)^2} < 1.$$

Logo,

$$\frac{2\sqrt{k}}{k+1} < 1.$$

Segue do que acabamos de mostrar que a recorrência

$$k_{n+1} = \frac{2\sqrt{k_n}}{k_n + 1}$$

define uma sequência crescente cujos elementos são todos menores que 1. Usando (104) como inspiração, definiremos uma segunda sequência pela regra

$$\phi_{n+1} = \frac{\arcsen(k_n \sen(\phi_n)) + \phi_n}{2},$$

de modo que

$$\hat{F}(\phi_n, k_n) = \frac{2}{k_n + 1} \hat{F}(\phi_{n+1}, k_{n+1}).$$

PROPOSIÇÃO 1. *A sequência k_n tem 1 como limite quando n tende a infinito.*

DEMONSTRAÇÃO. A ser escrita. □

Tomando k_0 e ϕ_0 como ponto de partida das duas recorrências,

$$\hat{F}(\phi_0, k_0) = \left(\frac{2}{k_0 + 1}\right) \left(\frac{2}{k_2 + 1}\right) \cdots \left(\frac{2}{k_{n-1} + 1}\right) \hat{F}(\phi_n, k_n)$$

Escolhendo n suficientemente grande para que $k_n \approx 1$, obtemos

$$\hat{F}(\phi_0, k_0) \approx \left(\frac{2}{k_1 + 1}\right) \left(\frac{2}{k_2 + 1}\right) \cdots \left(\frac{2}{k_{n-1} + 1}\right) \cdot \hat{F}(\phi_n, 1).$$

Contudo,

$$\hat{F}(u, 1) = \int_0^u \frac{dt}{\sqrt{1 - \sen^2(t)}} = \log \tan \left(\frac{\pi}{4} + \frac{u}{2} \right).$$

de modo que

$$\hat{F}(\phi_0, k_0) \approx \left(\frac{2}{k_1 + 1}\right) \left(\frac{2}{k_2 + 1}\right) \cdots \left(\frac{2}{k_{n-1} + 1}\right) \cdot \log \tan \left(\frac{\pi}{4} + \frac{\phi_n}{2} \right).$$

Vejamos como este procedimento pode ser usado para calcular uma aproximação para a integral

$$F(1, 1/2) = \hat{F}(\pi/2, 1/2)$$

para a qual encontramos uma aproximação na seção anterior usando a regra do retângulo. Tabelando os valores de k_n , ϕ_n e

$$F_n = \left(\frac{2}{1 + k_{n-1}}\right) F_{n-1},$$

com 10 casas decimais, a partir de

$$k_0 = \frac{1}{2}, \quad \phi_0 = \frac{\pi}{2} \quad \text{e} \quad F_0 = 1,$$

obtemos os seguintes resultados.

n	k_n	ϕ_n	F_n
0	1/2		
1	0.9428090415	1.0471975511	1.3333333333
2	0.9995666302	1.0012570846	1.3725830020
3	0.9999999765	1.0009188616	1.3728804844
4	1.0	1.0009188433	1.3728805006

Note que, em apenas quatro iterações, obtivemos $|k_n - 1| < 10^{-10}$. Segue da tabela que

$$F_4 = \left(\frac{2}{k_0 + 1}\right) \left(\frac{2}{k_2 + 1}\right) \left(\frac{2}{k_3 + 1}\right) = 1.3728805006$$

e como

$$\log \tan \left(\frac{\pi}{4} + \frac{1.0009188433}{2} \right) \approx 1.2433203640,$$

obtemos

$$\hat{F}(\pi/2, 1/2) \approx F_4 \log \tan \left(\frac{\pi}{4} + \frac{1.0009188433}{2} \right) \approx 1.7069302837.$$

Aumentando para 10 o número de iterações,

$$\hat{F}(\pi/2, 1/2) \approx 1.6857503548.$$

Esquemas iterativos semelhantes podem ser usados para calcular as demais integrais elípticas, assim como o seno elíptico; mais detalhes podem ser encontrados em [10] e [6].

CAPÍTULO 7

Sistemas dinâmicos

Um sistema dinâmico é um modelo matemático de processos naturais que evoluem ao longo do tempo. Um tal sistema pode ser contínuo ou discreto, dependendo se o fenômeno modelado varia continuamente, como um carro em movimento, ou ocorre somente a intervalos de tempo fixos, como as variações das populações de pássaros que só se reproduzem na primavera. Além de modelarem sistemas naturais, os sistemas dinâmicos podem ser usados para resolver, de maneira eficiente, problemas matemáticos importantes em aplicações, como o de encontrar as soluções de uma dada equação.

No capítulo anterior vimos vários exemplos de sistemas dinâmicos. O pêndulo é um sistema dinâmico contínuo, já o método de Euler usado para resolvê-lo numericamente é um exemplo de sistema dinâmico discreto. No primeiro caso o sistema é um modelo de processo natural, no segundo o sistema é usado para resolver um problema puramente matemático: achar uma aproximação numérica para a solução de um problema de valor inicial.

1. Iterações e pontos fixos

Na seção 3 do capítulo 6, aplicamos o método de Euler ao problema de valor inicial

$$\dot{\theta} = \sqrt{2 + 2 \cos(\theta)}, \quad \text{em que} \quad \theta(0) = 0,$$

com $h = 0.1$, obtendo a iteração

$$\theta_{k+1} = \theta_k + 0.1 \sqrt{2 + 2 \cos(\theta_k)}, \quad \text{com} \quad \theta_0 = 0,$$

A figura 1 da página 136 ilustra os 50 primeiros pontos desta iteração. Como você pode constatar, os pontos convergem para a reta horizontal $\theta = \pi$, que desenhamos em vermelho acima dos pontos. Por outro lado, tomando $\theta_0 = \pi$ como ponto de partida para a mesma iteração, obtemos $\theta_k = \pi$, porque

$$\sqrt{2 + 2 \cos(\theta_k)} = \sqrt{2 + 2 \cos(\pi)} = 0.$$

Portanto, o ponto π não apenas fica fixo pela iteração, como atrai para si a iteração que parte do 0.

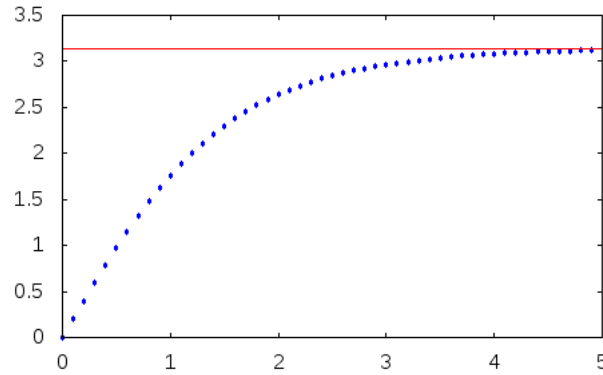


FIGURA 1. Pontos da iteração $\theta_{k+1} = \theta_k + h\sqrt{2 + 2\cos(\theta_k)}$ com $\theta_0 = 0$.

A existência de pontos fixos não é privilégio da iteração de Euler do pêndulo, nem sequer da iteração de Euler, mas ocorre em qualquer iteração definida por uma função contínua. Embora boa parte do que faremos nesta seção, e nas seguintes, possa ser aplicado a iterações em qualquer dimensão, vamos nos restringir ao caso unidimensional porque é mais simples de analisar.

Seja $g : [a, b] \rightarrow \mathbb{R}$ uma função e considere a iteração definida por

$$x_0 \in [a, b] \quad \text{e} \quad x_{k+1} = g(x_k).$$

Um *ponto fixo* desta iteração é um número $x_* \in [a, b]$ tal que $g(x_*) = x_*$. Note que estamos supondo que g *depende apenas de* x_k . No caso da iteração de Euler de uma equação diferencial de primeira ordem isto corresponde a dizer que a equação é *autônoma*; isto é, que é da forma $\dot{y} = f(y)$, em que a função f não depende diretamente de t , mas apenas de y ; como é o caso da equação do pêndulo.

Começaremos analisando um exemplo. Considere a iteração correspondente à função $g : [0, 2.2] \rightarrow \mathbb{R}$, definida pela regra

$$g(x) = \frac{-2}{(3x - 7)}.$$

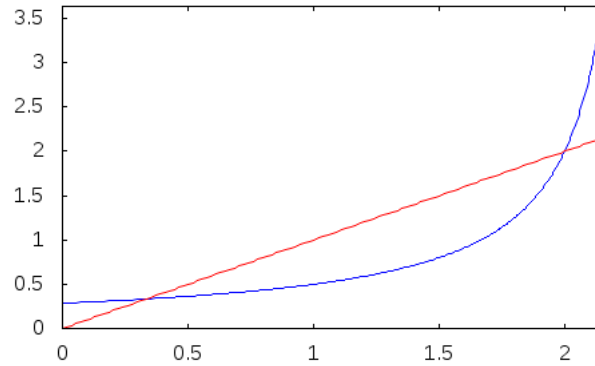
Os pontos fixos desta iteração são as soluções da equação $x = g(x)$; isto é, são as raízes $1/3$ e 2 de $3x^2 - 7x + 2 = 0$. Os dados numéricos relativos aos primeiros sete passos da iteração $x_{k+1} = g(x_k)$, para diferentes valores diferentes de x_0 , são apresentados na tabela 1.

Observe que, mesmo quando x_0 é um valor próximo do ponto fixo $x = 2$, a iteração converge para $x_* = 1/3 \approx 0.333$ e não para $x = 2$.

$x_0 \backslash k$	1	2	3	4	5	6	7
1.0	0.5	0.364	0.338	0.334	0.333	0.333	0.333
1.5	0.8	0.435	0.351	0.336	0.334	0.333	0.333
1.9	1.54	0.842	0.447	0.353	0.337	0.334	0.333
2.1	2.86	- 1.27	0.185	0.31	0.33	0.333	0.333

TABELA 1. Sete passos da iteração partindo de pontos iniciais distintos.

Para entender melhor o que está acontecendo interpretaremos a iteração geometricamente. Para começar, a equação $x = g(x)$ define os pontos de interseção da reta $y = x$ com a curva $y = g(x)$.

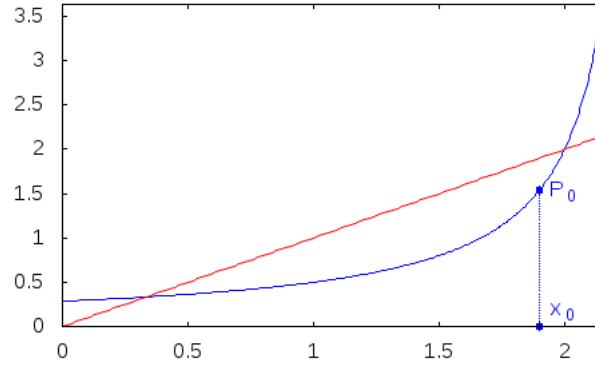
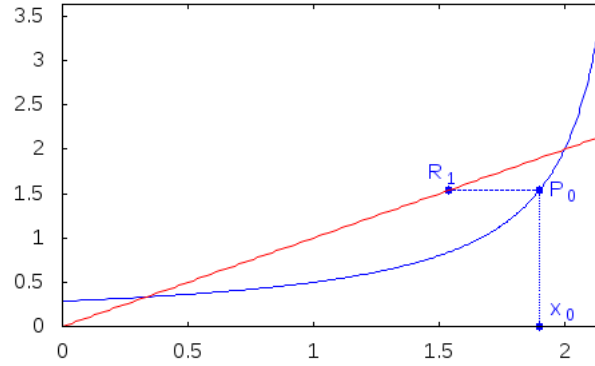
FIGURA 2. Interseção de $y = x$ e $y = g(x)$.

Digamos que nossa iteração comece em $x_0 = 1.9$. A perpendicular, traçada sobre o ponto $(x_0, 0)$ encontra a curva no ponto $P_0 = (x_0, g(x_0))$.

Construímos, então, a paralela ao eixo das abscissas que liga P_0 ao ponto R_1 na reta $y = x$.

Mas P_0 e R_1 têm a mesma ordenada $g(x_0)$, ao passo que ambas as coordenadas de R_1 são iguais; logo $R_1 = (g(x_0), g(x_0))$. Assim, a perpendicular, pelo ponto R_1 , ao eixo das abscissas encontra este eixo no ponto $x_1 = g(x_0)$.

Para simplificar a figura vamos apagar as linhas que usamos na construção, exceto pelos segmentos que ligam P_0 a R_1 e R_1 a P_1 , porque são bastante sugestivos.

FIGURA 3. Construindo o ponto $P_0 = (x_0, g(x_0))$.FIGURA 4. Construindo o ponto $R_1 = (g(x_0), g(x_0))$.

Repetindo a iteração mais duas vezes, encontramos uma escada, que vai descendo, de degrau em degrau, na direção do ponto (x_*, x_*) .

Portanto, quando um ponto fixo é um atrator, como é o caso de $x_* = 1/3$, podemos determiná-lo com qualquer precisão desejada, simplesmente repetindo a iteração uma quantidade suficientemente grande de vezes. Quando escolhemos $x_0 > 2$, a escada comporta-se de maneira extremamente espetacular, pulando de um ramo ao outro da hipérbole $y = -2/(3x - 7)$, como mostra a figura 15.

Como ocorre com $x = 3$, no exemplo que acabamos de analisar, uma iteração pode ter um ponto fixo que não é atrator. Ao contrário dos atratores, tais pontos só podem ser determinados resolvendo a equação $x = g(x)$. Sua primeira impressão pode ser que, dependendo de quão complicada seja a função g , isto pode ser bastante difícil de fazer. Por sorte, em um caso surpreendente de justiça poética, existe uma maneira de

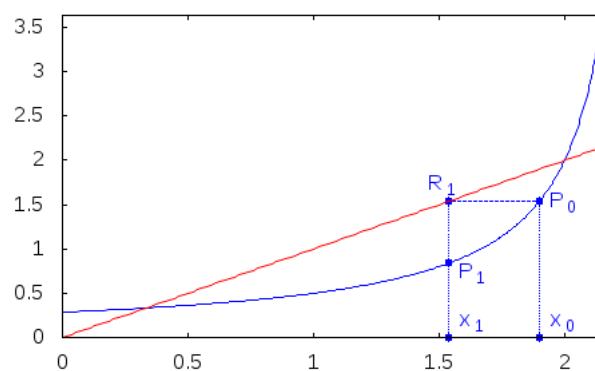
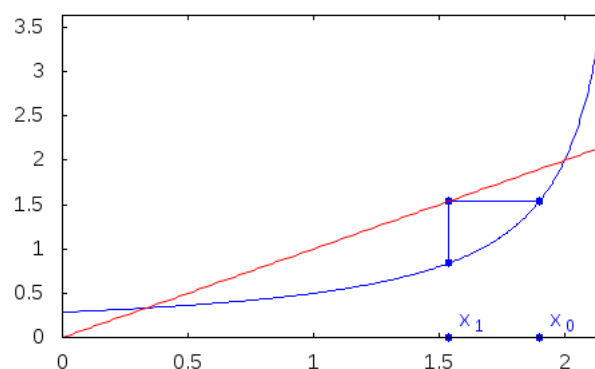
FIGURA 5. Construindo o ponto $P_1 = (g(x_0), 0)$.

FIGURA 6. O primeiro degrau da escada.

construir uma iteração, diferente de $x_{k+1} = g(x_k)$, da qual uma determinada solução de $x = g(x)$ é um ponto fixo. Esta iteração foi descoberta por Newton e será o tema da seção 3. Antes disto, porém, precisamos determinar sob que condições

1. uma iteração tem um ponto fixo;
2. um ponto fixo de uma iteração é um atrator.

Estes serão os temas de nossa próxima seção. Finalmente, convém observar que nem sempre a escada correspondente a uma iteração desce diretamente até o ponto fixo. Algumas vezes ela tem a forma de uma espiral, como no caso em que

$$g(x) = -\frac{1}{6}x^2 + \frac{1}{6}x + 4 \quad \text{e} \quad x_0 = \frac{1}{2},$$

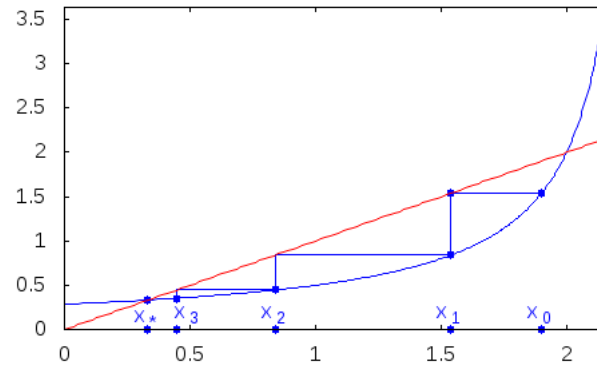
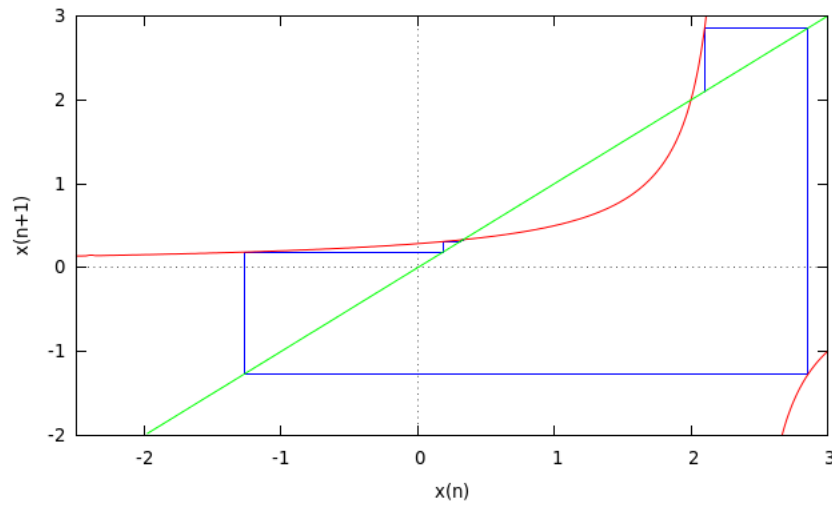


FIGURA 7. Os primeiros três degraus da escada.

FIGURA 8. A escada que começa em $x_0 = 2.2$.

como mostra a figura 16. Neste caso o ponto fixo para o qual a iteração converge é $x_* = 3$.

2. Existência de pontos fixos e de atratores

Suponhamos, como na seção anterior, que $g : [a, b] \rightarrow \mathbb{R}$ seja uma função e que estamos interessados em estudar a iteração definida por

$$(105) \quad x_0 \in [a, b] \quad \text{e} \quad x_{k+1} = g(x_k).$$

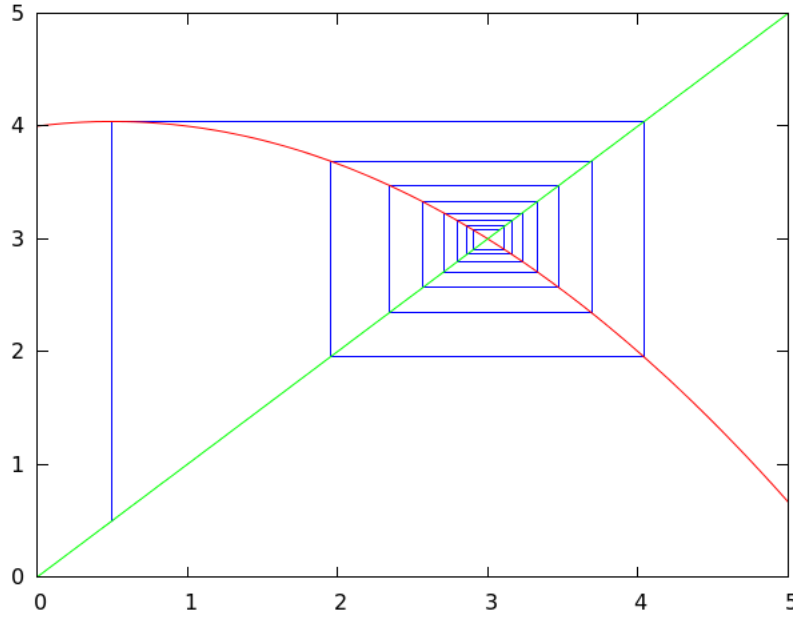


FIGURA 9. Uma escada em espiral.

Começaremos mostrando que basta que g seja contínua para que a iteração admita pontos fixos.

TEOREMA 3. *Toda iteração definida por uma função contínua em um intervalo fechado e limitado admite um ponto fixo.*

Se x_* for um ponto fixo de (105), então $g(x_*) - x_* = 0$. Portanto, para mostrar que (105) tem um ponto fixo no intervalo $[a, b]$, basta provar que a função

$$f(x) = g(x) - x$$

tem um zero neste mesmo intervalo. Como a imagem de g está contida em $[a, b]$, temos que $a \leq g(x) \leq b$, para todo $x \in [a, b]$. Em particular, $a \leq g(a)$ e $g(b) \leq b$, donde

$$f(a) = g(a) - a \geq 0 \quad \text{e} \quad f(b) = g(b) - b \leq 0.$$

Como a é um ponto fixo quando $f(a) = 0$ e b quando $f(b) = 0$, podemos supor que ambas as desigualdades são estritas. Contudo, como estamos supondo que g é uma função contínua, isto também valerá para f . Mas, pelo Teorema do Valor Intermediário, uma função contínua que passa de um valor positivo $f(a) > 0$ para um valor negativo $f(b) < 0$ tem que ter um zero no intervalo $[a, b]$; este zero é o ponto fixo desejado.

Tendo respondido à primeira das perguntas formuladas ao final da seção anterior, resta-nos considerar a segunda: sob que condições um ponto fixo é um atrator? Para poder respondê-la de maneira satisfatória, precisaremos supor que a primeira derivada de g existe e é contínua. Como uma função diferenciável é necessariamente contínua, o Teorema 3 nos garante que g tem um ponto fixo x_* no intervalo $[a, b]$. A hipótese que fizemos sobre g nos permite aplicar a fórmula de Taylor com resto de ordem um a g , de modo que

$$g(x) = g(x_*) + E_n(x) \quad \text{com} \quad |E_n(x)| \leq M(x - x_*),$$

em que M é o máximo de $g'(x)$ em $[a, b]$. Levando em conta que x_* é ponto fixo de g , resulta da equação anterior que

$$g(x) = x_* + E_n(x).$$

Assim,

$$|g(x) - x_*| \leq M|x - x_*|.$$

Se $M < 1$, então, da equação anterior,

$$|g(x) - x_*| \leq M \cdot |x - x_*|;$$

donde

$$|x_k - x_*| = |g(x_{k-1}) - x_*| \leq M \cdot |x_{k-1} - x_*|.$$

Iterando esta última desigualdade, obtemos

$$\begin{aligned} |x_k - x_*| &\leq M \cdot |x_{k-1} - x_*| \\ &\leq M^2 \cdot |x_{k-2} - x_*| \\ &\leq M^3 \cdot |x_{k-3} - x_*| \\ &\vdots \\ &\leq M^k \cdot |x_0 - x_*|. \end{aligned}$$

Porém, como $|x_0 - x_*|$ independe de k ,

$$0 \leq \lim_{k \rightarrow \infty} |x_k - x_*| \leq |x_0 - x_*| \lim_{k \rightarrow \infty} M^k = 0,$$

pois $0 < M < 1$. Mas isto é equivalente a dizer que,

$$\lim_{k \rightarrow \infty} x_k = x_*.$$

Logo, as hipóteses que impusemos a g garantem que o ponto fixo x_* da iteração (105) é um atrator. Em particular, este é o *único* ponto fixo possível no intervalo $[a, b]$, porque o argumento mostra que a iteração converge para x_* qualquer que seja o valor inicial $x_0 \in [a, b]$. Resumindo, provamos o seguinte teorema.

TEOREMA 4. *Seja $g : [a, b] \rightarrow \mathbb{R}$ uma função cuja primeira derivada existe e é contínua. Se houver um número real positivo $M < 1$ tal que $|g'(x)| < M$ para todo $x \in [a, b]$, então $x_{k+1} = g(x_k)$ tem um único ponto fixo $x_* \in [a, b]$ em $[a, b]$. Além disso, esta iteração converge para x_* , para todo $x_0 \in [a, b]$.*

⚡ Note que a hipótese de que “existe $0 < M < 1$ tal que $|g'(x)| < M$ ” não é equivalente a dizer simplesmente que $|g'(x)| < 1$. Na verdade, a demonstração do Teorema 4 não vai funcionar sob esta última hipótese, porque precisamos produzir uma sequência maior que $|x_k - x_*|$ e que necessariamente tenda a zero; contudo, uma sequência de números menores que 1 pode perfeitamente ter 1 como limite, como é o caso de $1 - 1/k$.

Vejamos o que este teorema nos diz sobre os dois exemplos apresentados na seção anterior. Como a derivada de

$$g(x) = \frac{-2}{(3x - 7)},$$

é igual a

$$g'(x) = \frac{6}{(3x - 7)^2},$$

temos que $|g'(x)| < 1$ quando

$$\left| \frac{(3x - 7)^2}{6} \right| = \frac{(3x - 7)^2}{6} > 1;$$

que equivale a

$$x < \frac{-\sqrt{6} + 7}{3} \approx 1.51 \quad \text{ou} \quad x > \frac{\sqrt{6} + 7}{3} \approx 3.14$$

Portanto, embora $|g'(x)|$ seja menor que 1 próximo de $x_* = 1/3$, temos

$$g'(1.9) \approx 3.55 > 1.$$

Isto mostra que, apesar da condição sobre o módulo da derivada nas hipóteses do teorema 4 ser *suficiente* para garantir que a iteração tende a $x_* = 1/3$, ela *não é necessária*. De fato, $x_0 = 1.9$ não pertence ao intervalo no qual $|g'(x)| < 1$ mas, mesmo assim, a iteração converge para x_* a partir 1.9.

A segunda iteração mencionada na seção anterior corresponde a tomar

$$g(x) = -\frac{1}{6}x^2 + \frac{1}{6}x + 4,$$

cujas derivada é

$$g'(x) = -\frac{1}{3}x + \frac{1}{6}.$$

Logo, para esta função, $|g'(x)| < 1$ se, e somente se,

$$-\frac{5}{2} < x < \frac{7}{2}.$$

Assim, $x_* = 3$ pertence a um intervalo adequado à aplicação do Teorema 4. Por exemplo, $|g'(x)| \leq 0.9$ para todo $x \in [-2.2, 3.2]$, o que nos permite tomar $L = 0.9$. Mesmo assim, a iteração converge para $x_* = 3$ fora deste intervalo, bastando que $x_0 \in [-8, 9]$. Contudo, se o valor inicial for escolhido fora de $[-8, 9]$, a iteração vai para $-\infty$.

O que vimos até agora mostra que, quando as condições do Teorema 4 são satisfeitas, o ponto fixo pode ser encontrado usando a iteração $x_{k+1} = g(x_k)$ e tendo como ponto de partida qualquer ponto no intervalo $[a, b]$ em que a função g está definida. Como esta iteração apenas converge para o ponto fixo x_* , precisamos de um critério para saber quando parar de iterar. Normalmente deveríamos iterar até que o erro correspondente a uma dada etapa da iteração estivesse abaixo de uma tolerância τ , que depende da precisão com a qual se deseja obter o ponto fixo. O problema é que x_* não é conhecido, ou não estaríamos tentando achar uma aproximação para ele. Como, então, determinar que a tolerância desejada foi alcançada? Como $g(x_*) - x_* = 0$, uma resposta possível consiste em parar quando

$$|g(x_k) - x_k| = |x_{k+1} - x_k| < \tau$$

ou quando

$$\frac{|x_{k+1} - x_k|}{|x_{k+1}|} \leq \tau.$$

Encerraremos esta seção com um exemplo em que usamos o primeiro destes critérios para encontrar o ponto fixo da iteração

$$g(x) = 1.982e^{-x/4},$$

no intervalo $[0, 2]$, com tolerância $\tau = 0.001$. Como g é diferenciável, começaremos verificando se o Teorema 4 pode ser aplicado. Calculando a derivada de g , obtemos

$$g'(x) = -\frac{1.982}{4}e^{-x/4}.$$

Como $e^{x/4}$ é uma função crescente, sua inversa $e^{-x/4} = 1/e^{x/4}$ é decrescente. Logo,

$$|g'(x)| = \frac{1.982}{4e^{x/4}} < \frac{1.982}{4} \leq 0.4995 < 1,$$

para todo $x \geq 0$. Portanto, g admite um único ponto fixo em $[0, 2]$; como este ponto é um atrator, podemos encontrá-lo usando a iteração $x_{k+1} = g(x_k)$ a partir de qualquer ponto no intervalo $[0, 2]$. Escolhendo $x_0 = 1.3$, os valores de x_k e das duas formas de calcular o erro nas primeiras seis iterações são dados na tabela 2. Estes dados mostram que a aproximação desejada é $x_4 = 1.396$.

k	x_k	$ x_{k+1} - x_k $	$ x_{k+1} - x_k / x_{k+1} $
1	1.4321	0.132	0.09221
2	1.3855	0.0465	0.033567
3	1.4018	0.0162	0.011566
4	1.3961	0.0057	0.0040659
5	1.3981	0.002	0.0014189
6	1.3974	0.0007	0.00049685

TABELA 2. Valores de x_k e dos erros na primeiras seis iterações

3. Zeros de funções

Como vimos em seções anteriores, os pontos fixos de uma iteração $x_{k+1} = g(x_k)$ correspondem aos zeros de $g(x) - x = 0$. Mas nada nos impede de inverter os termos desta relação e usar a iteração $x_{k+1} = g(x_k)$ para calcular os zeros de $g(x) - x = 0$, especialmente se estes zeros corresponderem a atratores da iteração.

Suponhamos, por exemplo, que desejamos achar os zeros de $f(x) = x \cos(x) - x^2 - 8x - 1$ no intervalo $[-1, 0]$ com erro inferior a 10^{-7} . Reescrevendo $x \cos(x) - x^2 - 8x - 1 = 0$ na forma,

$$x = \frac{x \cos(x) - x^2 - 1}{8},$$

podemos tomar

$$(106) \quad g(x) = \frac{x \cos(x) - x^2 - 1}{8}$$

e considerar a iteração $x_{k+1} = g(x_k)$. Para que esta estratégia seja bem sucedida é preciso que o ponto fixo em questão seja um atrator no intervalo $[-1, 0]$. Mas,

$$g'(x) = \frac{-x \sin(x) + \cos(x) - 2x}{8}$$

e como $|x|$, $|\cos(x)|$ e $|\sin(x)|$ são todos menores ou iguais a 1, a desigualdade triangular nos dá

$$|g'(x)| \leq \frac{|x| \cdot |\sin(x)| + |\cos(x)| + 2|x|}{8} \leq \frac{4}{8} = \frac{1}{2},$$

para todo $x \in [-1, 0]$. Portanto, pelo Teorema 4 a iteração $x_{k+1} = g(x_k)$ tem, realmente, um atrator em $[-1, 0]$. Em particular, $x \cos(x) - x^2 - 8x - 1$ tem apenas um zero no intervalo $[-1, 0]$, e para achá-lo podemos partir de qualquer $x_0 \in [-1, 0]$. Como a iteração é usada para calcular o zero de $f(x)$, podemos considerar $|f(x_k)|$

como uma terceira maneira de estimar o erro, juntamente com as outras duas que havíamos discutido na seção anterior. Escolhendo $x_0 = 0$ e calculando 10 iterações temos os seguintes valores para x_k e para os erros $|x_{k+1} - x_k|$ e $|f(x_k)|$.

k	x_k	$ x_k - x_{k-1} $	$ f(x_k) $
1	-0.125	0.125	1.0
2	-0.14245621355	0.01745621355	0.139649708404
3	-0.145163367774	0.002707154224	0.0216572337908
4	-0.145588623064	0.00042525529	0.00340204232043
5	-0.145655555252	0.000066932188	0.000535457502565
6	-0.145666093132	0.00001053788	0.0000843030371698
7	-0.145667752307	0.00000165918	0.0000132734043063
8	-0.145668013544	0.26123710^{-6}	0.00000208989627026
9	-0.145668054676	0.4113210^{-7}	0.32905518310110^{-6}
10	-0.145668061152	0.647610^{-8}	0.51809251999710^{-7}

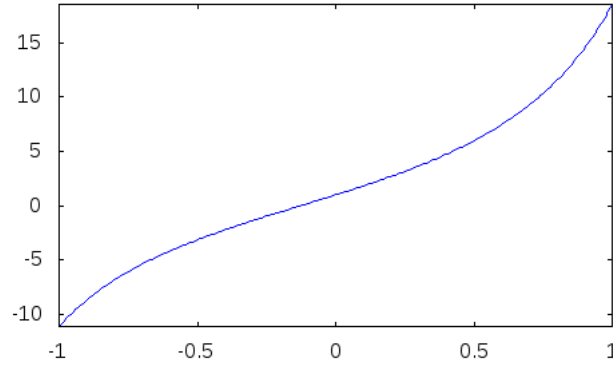
TABELA 3. Zero de $x \cos(x) - x^2 - 8x - 1$ calculado usando método iterativo

Logo, -0.145668061152 é uma aproximação para $f(x)$ no intervalo $[-1, 0]$ com erro inferior a 10^{-7} . O problema desta estratégia é que nem sempre é fácil identificar qual é a função que deve ser usada na iteração. Para ilustrar isto, digamos que em vez de (106), tivéssemos escolhido

$$g(x) = \frac{x^2 + 8x + 1}{\cos(x)}.$$

Como mostra a figura 17, esta função é crescente no intervalo $[-1, 1]$, de modo que a sequência de valores gerada por $x_{k+1} = g(x_k)$ cresce a cada iteração e, portanto, não converge para nenhum número neste intervalo.

Felizmente a saída para este problema já havia sido inventada antes mesmo do problema surgir. Inspirando-se em um exemplo publicado por Newton, Joseph Raphson (c. 1648 – c. 1715) propôs uma estratégia iterativa que se revelou extremamente eficiente para achar zeros de funções. A ideia por trás do método de *Newton-Raphson* é a seguinte. Suponhamos que uma função diferenciável $f(x)$ tem um zero x_* e que x_0 é um número próximo de x_* . Usando a fórmula de Taylor de ordem um, podemos

FIGURA 10. Gráfico de $y = g(x)$.

escrever

$$f(x_*) \approx f(x_0) + f'(x_0)(x_* - x_0),$$

pois o erro, que depende de $(x_* - x_0)^2$, será muito pequeno quando x_0 estiver muito próximo de x_* . Mas, x_* é, por hipótese, um zero de $f(x)$, de modo que

$$0 \approx f(x_0) + f'(x_0)(x_* - x_0).$$

Isolando o valor de x_* desta equação, obtemos

$$x_* \approx x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Portanto, desde que x_0 tenha sido escolhido próximo de x_* ,

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

pode ser considerado como uma aproximação de x_* . Naturalmente, podemos considerar x_1 como ponto de partida para obter uma aproximação ainda melhor

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)},$$

e assim por diante. Em outras palavras, o *método de Newton-Raphson* consiste em usar a iteração $x_{k+1} = g(x_k)$ na qual

$$(107) \quad g(x) = x - \frac{f(x)}{f'(x)},$$

para achar um zero de uma função diferenciável $f(x)$.

Do ponto de vista geométrico, este método corresponde, essencialmente, a considerar a reta tangente à curva em um dado ponto como uma aproximação da curva e a usar isto para calcular o zero. Supondo, como acima, que a função cujo zero

queremos calcular seja $f(x)$, a inclinação da tangente a $y = f(x)$ no ponto x_0 será $f'(x_0)$. Com isso, a reta tangente a $y = f(x)$ em $(x_0, f(x_0))$ terá por equação

$$y - f(x_0) = f'(x_0)(x - x_0).$$

Resolvendo esta equação quando $y = 0$, verificamos que esta reta intersecta o eixo das abscissas no ponto cuja primeira coordenada é

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)},$$

que corresponde à primeira iteração do método de Newton-Raphson. O processo é então repetido a partir de x_1 . A figura abaixo ilustra duas iterações do método de Newton-Raphson quando aplicado a $x^3 + x^2 - 10$ a partir de $x_0 = 2$.

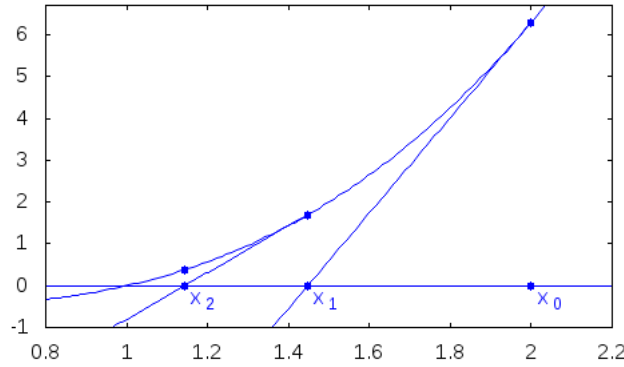


FIGURA 11. Duas iterações do método de Newton-Raphson.

Agora que já temos a iteração do método de Newton-Raphson, resta-nos mostrar que o zero de $f(x)$ é, de fato, um atrator desta iteração. Para isto usaremos o Teorema 4. Mas,

$$g'(x) = \frac{f(x)f''(x)}{f'(x)^2}$$

de modo que, se $f'(x_*) \neq 0$,

$$(108) \quad g'(x_*) = \frac{f(x_*)f''(x_*)}{f'(x_*)^2} = 0,$$

pois x_* é um zero de $f(x)$. Assim, se supusermos que a primeira e a segunda derivadas de f são contínuas em seu intervalo de definição e que $f'(x_*) \neq 0$, teremos que $g'(x_*) = 0$ e que $g'(x)$ é contínua neste mesmo intervalo. Logo, escolhendo $e > 0$ suficientemente pequeno, teremos

$$|g'(x)| < 1 \quad \text{para todo} \quad x \in (x_* - e, x_* + e).$$

Portanto, a intuição de Newton e Raphson estava correta: x_* funciona como um atrator da iteração definida pela função (107), desde que o ponto de partida esteja suficientemente próximo de x_* .

A tabela a seguir mostra o que acontece quando aplicamos o método de Newton-Raphson para calcular um zero de $f(x) = x \cos(x) - x^2 - 8x - 1$ no intervalo $[-1, 0]$, começando de $x_0 = 0$ e com tolerância $\tau = 10^{-8}$.

k	x_k	$ x_k - x_{k-1} $	$ f(x_k) $
1	-0.1428571429	0.1428571429	0.0189529148
2	-0.1456671421	0.0028099993	$0.62027496778 \cdot 10^{-5}$
3	-0.1456680624	0.920210^{-6}	$0.66280314570 \cdot 10^{-12}$
4	-0.1456680624	$0.9 \cdot 10^{-13}$	$0.27755575616 \cdot 10^{-16}$

TABELA 4. Zero de $x \cos(x) - x^2 - 8x - 1$ calculado usando Newton-Raphson

Note que, neste caso, atingimos em apenas 4 iterações uma precisão melhor, do que a iteração definida pela função (106) foi capaz de nos dar em 10 iterações.

A razão pela qual o método de Newton-Raphson tem uma performance tão melhor que os métodos iterativos que utilizam funções construídas de maneira *ad-hoc* é que tem convergência quadrática, ao passo que a convergência das últimas é, quase sempre, linear. Para entender porque isto acontece, suponhamos que $g : [a, b] \rightarrow [a, b]$ é uma função cujas primeiras duas derivadas existem e são contínuas em (a, b) . Digamos que x_* seja um ponto fixo atrator da iteração $x_{k+1} = g(x_k)$. Calculando a fórmula de Taylor de ordem um de g na vizinhança de x_* , obtemos

$$g(x) - g(x_*) = g'(x_*)(x - x_*) + E_2, \quad \text{com} \quad |E_2| \leq \frac{M(x - x_*)^2}{2},$$

em que M é o máximo de $g''(x)$ entre x e x_* . Portanto, na k -ésima etapa, o erro cometido no cálculo de x_* vai satisfazer

$$|x_k - x_*| = |g'(x_*)(x_{k-1} - x_*) + E_2| \leq \left(|g'(x_*)| + \frac{M|x - x_*|}{2} \right) |x_{k-1} - x_*|.$$

Logo, *a não ser que $g'(x_*) = 0$* , o erro será diretamente proporcional a $|(x - x_*)|$ e, portanto, linear. Entretanto, a situação é bem diferente quando g é obtida através da aplicação do método de Newton-Raphson a alguma função f . Supondo que as duas primeiras derivadas de f são contínuas e que $f(x_*) = 0$, mas $f'(x_*) \neq 0$, a equação (108) nos garante que $g'(x_*) = 0$, de modo que, neste caso o erro é sempre

diretamente proporcional a $|(x - x_*)|^2$. Grosso modo, isto significa que o número de casas decimais corretas dobra a cada iteração, como ocorre, por exemplo, com o erro na terceira coluna da tabela 4. Note, porém, que o argumento acima falhará se $f'(x_*) = 0$. Quando f é um polinômio, isto corresponde a dizer que x_* é uma raiz dupla de f . Mais detalhes de como tratar este caso podem ser encontrados na lista de exercícios.

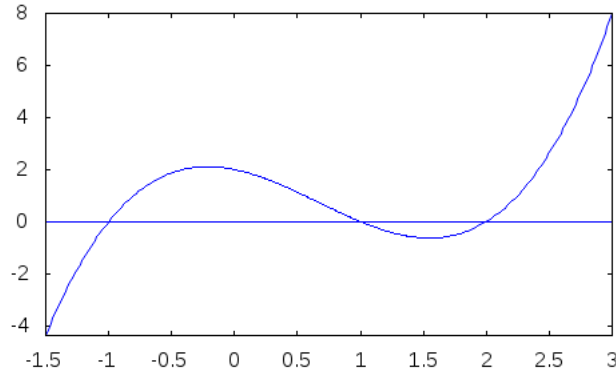
Ainda há um problema que precisamos considerar, antes de dar por encerrado nosso estudo do método de Newton-Raphson. Como o argumento que usamos para introduzir este método sugere, só podemos ter a garantia de que vai convergir para o zero desejado se começamos a iterar de um ponto “suficientemente próximo” deste zero. Por exemplo, a cúbica $f(x) = x^3 - 2x^2 - x + 2$ cujo gráfico é ilustrado na figura 19 tem zeros nos pontos $x = -1$, $x = 1$ e $x = 2$. A tabela 5 lista, para alguns pontos de partida x_0 , o zero x_* que foi encontrado e a quantidade k de iterações necessárias para determiná-lo com erro inferior a 10^{-8} .

x_0	-2.0	-0.3	-0.2	0.0	0.1	0.2	0.5	1.1	1.4	1.6	2.5	3.0
x_*	-1.0	-1.0	2.0	2.0	2.0	1.0	1.0	1.0	1.0	2.0	2.0	2.0
k	6	10	14	2	13	7.0	2	5	6	7	7	7

TABELA 5. Zeros de $f(x) = x^3 - 2x^2 - x + 2$.

Os dados da tabela mostram que, como já havíamos visto, um zero funciona como atrator dos pontos que lhe são próximos, mas, também, que o comportamento dos pontos distantes pode ser menos predizível. Assim, por exemplo, as iterações que têm como ponto de partida -0.2 , 0.0 e 0.1 convergem para 2.0 , e não para 1.0 , que é o zero mais próximo. Portanto, precisamos mesmo estar bastante próximos para que possamos ter certeza que o método de Newton-Raphson vá convergir para o zero cuja aproximação desejamos calcular.

A estratégia que usaremos para achar pontos próximos a um zero é apenas um outro método, não tão eficiente, para achar zeros de uma função. Trata-se de um procedimento semelhante ao que usamos para encontrar palavras em um dicionário: começamos abrindo o dicionário mais ou menos ao meio e verificando em qual das duas metades encontra-se a palavra que procuramos; repetimos, então, o mesmo procedimento com aquela metade, continuando assim até que a página onde a palavra se encontra seja achada. Este método funciona por duas razões. A primeira é que as palavras de um dicionário estão ordenadas em ordem alfabética; a segunda é que a primeira e a última palavras de cada página estão escritas em seu cabeçalho, o

FIGURA 12. Gráfico da cúbica $f(x) = x^3 - 2x^2 - x + 2$.

que torna fácil verificar em qual das duas metades está a palavra procurada. Para ser bem sucedida, nossa aplicação do método do dicionário, ou *método de bisseção* como é conhecido em análise numérica, precisa satisfazer as duas condições acima. A primeira é evidentemente satisfeita, porque os números reais têm uma ordenação natural; a segunda é consequência do Teorema do Valor Intermediário:

se uma função contínua $f : [a, b] \rightarrow \mathbb{R}$ satisfaz $f(a)f(b) < 0$, então f tem um zero entre a e b .

Antes de sistematizar o método de bisseção, analisaremos como se comporta quando aplicado para achar os zeros da cúbica $f(x) = x^3 - 2x^2 - x + 2$, que analisamos anteriormente. Digamos, para começar, que estamos interessados nos zeros negativos da cúbica. Como

$$f(0) = 2 \quad \text{e} \quad \lim_{x \rightarrow -\infty} f(x) = -\infty,$$

esta cúbica tem que ter um zero negativo. Na verdade, como $f(-3) = -40$, tem que haver um zero entre -3 e 0 . Subdividindo o intervalo $[-3, 0]$ ao meio, calculamos $f(-3/2) = -35/8 < 0$. Apelando ao Teorema do Valor Intermediário, isto nos garante que há um zero entre $[-3/2, 0]$. Subdividindo este último intervalo e levando em conta que $f(-3/4) = 77/64 > 0$, o mesmo argumento nos permite concluir que o zero encontra-se em $[-3/2, -3/4]$. Continuando desta maneira, obteremos ao cabo de n etapas um intervalo de comprimento $3/n$ no qual o zero desejado está localizado. Se tivéssemos começado procurando os zeros positivos de $f(x)$ teríamos nos deparado com uma situação bem menos satisfatória, porque,

$$f(0) = 2 \quad \text{e} \quad \lim_{x \rightarrow +\infty} f(x) = +\infty,$$

não nos permitem usar o Teorema do Valor Intermediário para assegurar que $f(x)$ tenha um zero positivo. Contudo, sabemos que estes zeros existem. Portanto, *uma*

função pode ter zeros em um intervalo mesmo que as condições do Teorema do Valor Intermediário não sejam satisfeitas. Na prática, geralmente estamos à procura de um zero em um intervalo relativamente restrito, que podemos subdividir até ter uma noção mais clara da posição dos zeros.

Voltando ao caso geral, digamos que uma função contínua $f : [a, b] \rightarrow \mathbb{R}$, que satisfaz $f(a)f(b) < 0$, tem um único zero em $[a, b]$. Para achar este zero pelo *método de bisseção* começamos definindo

$$\alpha_0 = a \quad \text{e} \quad \beta_0 = b$$

e calculando $\mu_0 = (\alpha_0 + \beta_0)/2$. Se $f(\alpha_0)f(\mu_0) < 0$ então, pelo Teorema do Valor Intermediário, a raiz de f pertence ao intervalo $[\alpha_0, \mu_0]$. Neste caso, tomamos

$$\alpha_1 = \alpha_0 \quad \text{e} \quad \beta_1 = \mu_0.$$

Caso contrário, o zero pertence a $[\mu_0, \beta_0]$ e tomamos

$$\alpha_1 = \mu_0 \quad \text{e} \quad \beta_1 = \beta_0.$$

Este procedimento é, então, repetido até que a tolerância desejada seja atingida. Chamando de τ esta tolerância, isto vai ocorrer quando

$$(109) \quad |\beta_n - \alpha_n| < \tau.$$

Contudo, como a cada iteração o intervalo é dividido em dois, temos que

$$(110) \quad |\beta_n - \alpha_n| = \frac{|\beta_{n-1} - \alpha_{n-1}|}{2},$$

donde podemos concluir que (109) equivale a

$$|\beta_n - \alpha_n| = \frac{|\beta_0 - \alpha_0|}{2^n} = \frac{b - a}{2^n} < \tau.$$

Portanto, (109) vale sempre que

$$n > \log_2 \left(\frac{b - a}{\tau} \right).$$

Se você está se perguntando porque simplesmente não usamos este método para achar a raiz com a aproximação desejada, a resposta está na equação (110), segundo a qual o decaimento do erro no método de bisseção é linear, e não quadrático como no método de Newton-Raphson.

4. Métodos iterativos para sistemas lineares

Seja A uma matriz real de tamanho $n \times n$ e b um vetor do \mathbb{R}^n . Queremos achar a solução v^* do sistema $AX = b$ usando um método iterativo. Isto é, queremos inventar uma sequência de vetores $v^{(m)}$ que tende a $A^{-1}b$ quando m tende a infinito. Note que, para evitar confusão entre as coordenadas de um vetor e sua posição na

seqüência de iterações, denotaremos esta última por um número entre parêntesis sobrescrito ao vetor. Assim, a seqüência de vetores gerada por uma dada iteração será $v^{(0)}, v^{(1)}, v^{(2)}, \dots$ e as coordenadas do k -ésimo vetor desta seqüência serão

$$v^{(k)} = (x_1^{(k)}, \dots, x_n^{(k)}).$$

Suponha que possamos decompor a matriz A , do sistema $AX = b$, na forma

$$A = M - K,$$

em que M é uma matriz inversível, temos que

$$Mv^* - Kv^* = b,$$

pode ser reescrito na forma

$$v^* = M^{-1}Kv^* + M^{-1}b.$$

Assim, quando $R = M^{-1}K$ e $c = M^{-1}b$, a solução v^* do sistema $AX = b$ é ponto fixo da iteração

$$(111) \quad v^{(m+1)} = Rv^{(m)} + c.$$

Para podermos resolver o sistema desta maneira, precisamos provar que a iteração (111) converge para o ponto fixo. Suponhamos, para começar, que R seja uma matriz, diagonal:

$$R = \text{diag}(\lambda_1, \dots, \lambda_n) = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}.$$

Neste caso, se

$$c = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix},$$

então a iteração (111) toma a forma

$$\begin{bmatrix} x_1^{(m+1)} \\ \vdots \\ x_m^{(m+1)} \end{bmatrix} = \begin{bmatrix} \lambda_1 x_1^{(m)} + c_1 \\ \vdots \\ \lambda_n x_n^{(m)} + c_n \end{bmatrix}$$

Note que cada linha corresponde a uma iteração da forma

$$x_j^{(m+1)} = \lambda_j x_j^{(m)} + c_j \quad (j = 1, \dots, n),$$

que é independente do que acontece nas outras entradas do vetor. Mas estas são iterações lineares e basta que todos os λ s tenham módulo menor que 1 para que seja convergente. Provamos, portanto, o seguinte resultado:

se $R = \text{diag}(\lambda_1, \dots, \lambda_n)$ e $|\lambda_j| < 1$, para todo $1 \leq j \leq n$, então a iteração $v^{(m+1)} = Rv^{(m)} + c$ converge a partir de qualquer $v^{(0)}$.

Consideramos, em seguida, o caso, um pouco mais geral, em que R é diagonalizável. Isto é, estamos supondo que existe uma matriz inversível Q , a matriz de mudança de base, e uma matriz diagonal D , tais que

$$R = Q^{-1}DQ.$$

Este caso pode ser facilmente reduzido ao anterior, porque, multiplicando

$$v^{(m+1)} = Rv^{(m)} + c = Q^{-1}DQv^{(m)} + c$$

à esquerda por Q , obtemos

$$Qv^{(m+1)} = DQv^{(m)} + Qc.$$

Mas, escrevendo $w^{(m)} = Qv^{(m)}$, esta última iteração toma a forma

$$(112) \quad w^{(m+1)} = Dw^{(m)} + Qc.$$

Como D é diagonal, já sabemos que basta que as entradas da sua diagonal tenham módulo menor que 1 para que (112) seja convergente. Levando em conta que Q é inversível e não depende de m , temos que

$$\lim_{m \rightarrow \infty} v^{(m)} = \lim_{m \rightarrow \infty} Qw^{(m)} = Q(\lim_{m \rightarrow \infty} w^{(m)}).$$

Como este último limite existe, então $\lim_{m \rightarrow \infty} v^{(m)}$ também existe.

Provamos, assim, um resultado semelhante ao anterior no caso em que R é diagonalizável. Só que desta vez a hipótese que garante a convergência é bem mais difícil de verificar, porque diz respeito à forma diagonalizada de R . Note, porém que, se $D = \text{diag}(\lambda_1, \dots, \lambda_n)$, então

$$R - \lambda_1 \cdot I = Q^{-1}DQ - \lambda_1 \cdot Q^{-1}Q = Q^{-1}(D - \lambda_1 \cdot I)Q,$$

em que a primeira igualdade segue do fato de $Q^{-1}Q$ ser igual à matriz identidade I . Calculando o determinante da equação anterior, obtemos

$$\det(R - \lambda_1 \cdot I) = \det(Q^{-1}) \det(D - \lambda_1 \cdot I) \det(Q) = \det(D - \lambda_1 \cdot I),$$

pois $\det(AB) = \det(A) \det(B)$, quaisquer que sejam as matrizes A e B de tamanho $n \times n$. Como a primeira linha da matriz $D - \lambda_1 \cdot I$ é nula, seu determinante é zero. Portanto,

$$\det(R - \lambda_1 \cdot I) = \det(D - \lambda_1 \cdot I) = 0,$$

e um argumento semelhante mostra que o mesmo vale para todos os outros λ s.

Temos, assim, uma maneira de determinar os elementos não nulos na forma diagonal de R , sem a necessidade de encontrar a matriz Q . Basta, para isto, achar as raízes da equação polinomial $\det(R - t \cdot I) = 0$, conhecidas como *autovalores* de

R . Definindo o *raio espectral* $\rho(R)$ de R , como o máximo dos módulos dos seus autovalores, podemos enunciar o resultado que provamos acima da seguinte maneira:

se R é diagonalizável e seu raio espectral é menor que um, então a iteração $v^{(m+1)} = Rv^{(m)} + c$ converge a partir de qualquer $v^{(0)}$.

Note que tanto a equação $\det(R - t \cdot I) = 0$ como a noção de raio espectral continuam fazendo sentido mesmo que R não seja diagonalizável. Na verdade, o resultado que provamos não foi agraciado com o título de teorema justamente porque, também ele continua válido mesmo que R não seja diagonalizável. Contudo a demonstração do caso geral requer um cuidado maior do que desejamos dispender, por isso não vamos apresentá-la aqui; você pode encontrá-la, por exemplo, em [4, p. 511, Theorem 10.1.1]. O resultado geral é o seguinte.

TEOREMA 5. *Se $\rho(R) < 1$, a iteração dada pela fórmula (111) converge qualquer que seja o vetor inicial escolhido.*

Resumindo, dado um sistema linear $AX = b$, em que A é uma matriz quadrada que pode ser decomposta na forma $A = M - K$ com

- M uma matriz inversível e
- $R = M^{-1}K$ de raio espectral menor que um

então a iteração dada por

$$v^{(m+1)} = Rv^{(m)} + c,$$

em que $c = M^{-1}b$ converge para a solução de $AX = b$, qualquer que seja a escolha do vetor inicial $v^{(0)} \in \mathbb{R}^n$.

Para que esta seja uma maneira prática de resolver sistemas lineares, é necessário que seja possível inverter M facilmente, para que possamos achar R ; caso contrário, seria preferível inverter A diretamente e calcular $A^{-1}b$, que é a solução de $AX = b$. Uma maneira extremamente simples de fazer isto é decompor A na forma

$$A = L + D + U,$$

em que L é uma matriz triangular inferior e U uma matriz triangular superior, *ambas com diagonal nula* e D é uma matriz diagonal. Supondo que todas as entradas da diagonal de D são diferentes de zero, podemos tomar

$$M = D, \quad K = -L - U \quad \text{e} \quad R = -D^{-1}(L + U);$$

que são fáceis de obter porque, para inverter D , basta inverter as entradas de sua diagonal. Neste caso obtemos a iteração

$$(113) \quad v^{(m+1)} = -D^{-1}(L + D)v^{(m)} + D^{-1}b.$$

Escrevendo

$$(114) \quad v^{(m)} = (x_1^{(m)}, \dots, x_n^{(m)})$$

e supondo que

$$(115) \quad A = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{bmatrix},$$

as coordenadas de $v^{(m+1)}$ em (113) têm a forma

$$(116) \quad x_i^{(m+1)} = \frac{1}{a_{i,i}} \left(b_i - \sum_{k \neq i} a_{i,k} x_k^{(m)} \right).$$

O algoritmo para resolver sistemas lineares usando esta iteração é conhecido como *método de Gauss-Jacobi*. De acordo com o que vimos anteriormente, para que o método de Gauss-Jacobi seja convergente basta que o raio espectral da matriz $R = -D^{-1}(L + D)$ seja menor que 1. Infelizmente, este não é um critério que possamos aplicar na prática, porque calcular os autovalores de uma matriz grande tem um custo bastante alto. O ideal seria um critério de convergência que dependesse, de maneira simples, apenas dos coeficientes da matriz R . No caso do método de Gauss-Jacobi este critério existe e é bastante fácil de aplicar. Antes de estabelecer o critério, precisamos de uma definição e de um resultado auxiliar.

Diremos que uma matriz real $n \times n$

$$(117) \quad B = \begin{bmatrix} b_{1,1} & b_{1,2} & \dots & b_{1,n} \\ b_{2,1} & b_{2,2} & \dots & b_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n,1} & b_{n,2} & \dots & b_{n,n} \end{bmatrix}$$

tem *diagonal dominante* se, para todo $1 \leq i \leq n$,

$$|b_{i,i}| > \sum_{j \neq i} |b_{i,j}|;$$

isto é, para cada linha, a soma dos módulos das entradas fora da diagonal é menor que a entrada na diagonal. Assim, no caso em que $n = 3$, a matriz B terá diagonal dominante quando as três desigualdades seguintes forem satisfeitas:

$$\begin{aligned} |b_{1,1}| &> |b_{1,2}| + |b_{1,3}| \\ |b_{2,2}| &> |b_{2,1}| + |b_{2,3}| \\ |b_{3,3}| &> |b_{3,1}| + |b_{3,2}|. \end{aligned}$$

Por exemplo, a matriz

$$(118) \quad B = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 3 & 1 \\ 2 & 1 & 5 \end{bmatrix}$$

tem diagonal dominante. O resultado auxiliar que usaremos relaciona os autovalores de uma matriz à soma das entradas de uma linha.

LEMA 1. *O módulo de um autovalor de uma matriz $n \times n$ é sempre menor ou igual ao máximo das somas dos módulos das entradas de uma linha.*

Antes de provar este resultado, precisamos de uma definição. A *norma infinito* de um vetor $u = (x_1, \dots, x_n) \in \mathbb{R}^n$ é

$$\|u\|_\infty = \max\{|x_1|, \dots, |x_n|\}.$$

Por exemplo, $\|(1, 2, 9, 2)\|_\infty = 9$. Como no caso da norma euclidiana usual, vale que $\|\lambda u\|_\infty = |\lambda| \|u\|_\infty$, quaisquer que sejam o escalar λ e o vetor u . Com isso estamos prontos para provar o teorema.

DEMONSTRAÇÃO. Seja λ um autovalor da matriz B da equação (117), precisamos mostrar que

$$(119) \quad \lambda \leq \max\left\{\sum_{j=1}^n |b_{i,j}| : 1 \leq i \leq n\right\}.$$

Mas se λ é um autovalor de B , então existe um vetor não nulo

$$v = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

tal que $Bv = \lambda v$. Expandindo o lado esquerdo, verificamos que

$$\|Bv\|_\infty = \max\{|b_{i,1}x_1 + \dots + b_{i,n}x_n| : 1 \leq i \leq n\}.$$

Aplicando a desigualdade triangular,

$$|b_{i,1}x_1 + \dots + b_{i,n}x_n| \leq |b_{i,1}||x_1| + \dots + |b_{i,n}||x_n|,$$

para todo $1 \leq i \leq n$, de modo que

$$(120) \quad \|Bv\|_\infty \leq \max\{|b_{i,1}||x_1| + \dots + |b_{i,n}||x_n| : 1 \leq i \leq n\}.$$

Como,

$$|x_j| \leq \|v\|_\infty,$$

para todo $1 \leq j \leq n$, obtemos de (120) a desigualdade

$$\|Bv\|_\infty \leq \max\{|b_{i,1}| + \dots + |b_{i,n}| : 1 \leq i \leq n\} \cdot \|v\|_\infty.$$

Por outro lado, $Bv = \lambda v$ nos dá

$$\|Bv\|_\infty = |\lambda| \|v\|_\infty;$$

donde

$$|\lambda| \|v\|_\infty \leq \max\{|b_{i,1}| + \cdots + |b_{i,n}| : 1 \leq i \leq n\} \cdot \|v\|_\infty.$$

Observe que esta desigualdade difere de (119) apenas pelo fato de termos $\|v\|_\infty$ multiplicado dos lados esquerdo e direito. Portanto, teremos a desigualdade desejada se pudermos provar que v tem norma infinito positiva; mas para isto basta que $v \neq 0$, que é verdadeiro por hipótese. \square

Com isto estamos prontos para provar a convergência do método de Gauss-Jacobi.

TEOREMA 6. *Se A é uma matriz real inversível $n \times n$ que tem diagonal dominante, então o método de Gauss-Jacobi aplicado a A e $b \in \mathbb{R}^n$ converge para uma solução do sistema linear $AX = b$, qualquer que seja $v^{(0)} \in \mathbb{R}^n$.*

DEMONSTRAÇÃO. Como a iteração do método de Gauss-Jacobi é dada por

$$v^{(m+1)} = Rv^{(m)} + D^{-1}b.$$

em que $R = -D^{-1}(L+D)$, o resultado desejado segue do teorema 5 se formos capazes de mostrar que o raio espectral de R é menor que 1. Contudo, $R = -D^{-1}(L+U)$ é igual a

$$\begin{bmatrix} 0 & a_{1,2}/a_{1,1} & a_{1,3}/a_{1,1} & \cdots & a_{1,n-1}/a_{1,1} & a_{1,n}/a_{1,1} \\ a_{2,1}/a_{2,2} & 0 & a_{2,3}/a_{2,2} & \cdots & a_{2,n-1}/a_{2,2} & a_{2,n}/a_{2,2} \\ a_{3,1}/a_{3,3} & a_{3,2}/a_{3,3} & 0 & \cdots & a_{3,n-1}/a_{3,3} & a_{3,n}/a_{3,3} \\ \vdots & & \vdots & \ddots & \vdots & \\ a_{n-1,1}/a_{n-1,n-1} & a_{n-1,2}/a_{n-1,n-1} & \cdots & & 0 & a_{n-1,n}/a_{n-1,n-1} \\ a_{n,1}/a_{n,n} & 0 & a_{n,3}/a_{n,n} & \cdots & a_{n,n-1}/a_{n,n} & 0 \end{bmatrix},$$

de modo que, pelo lema 1, se λ for um autovalor de R , então seu módulo é menor que o número

$$M = \max \left\{ \sum_{j=1}^n \frac{a_{i,j}}{a_{i,i}} : 1 \leq i \leq n \right\}.$$

Finalmente,

$$\sum_{j=1}^n \frac{a_{i,j}}{a_{i,i}} = \frac{1}{a_{i,i}} \left(\sum_{j=1}^n a_{i,j} \right) < 1$$

pois A tem diagonal dominante. Note que não precisamos excluir a posição diagonal dos somatórios porque a diagonal de R é nula. Logo, $\rho(R) \leq M < 1$ e o resultado desejado segue do teorema 5, como afirmamos anteriormente. \square

Vejam os um exemplo. Digamos que queremos resolver o sistema $BX = b$, em que A é a matriz dada em (118) e $b = [1, 4, 7]^t$. Tomando como ponto de partida o vetor $v^{(0)} = [1, 1, 1]^t$, as duas primeiras iterações nos dão, em ponto flutuante com três dígitos significativos

$$v^{(1)} = \begin{bmatrix} -3.1666 \cdot 10^{-1} \\ 1.2333 \cdot 10^0 \\ 1.4667 \cdot 10^0 \end{bmatrix}, v^{(2)} = \begin{bmatrix} -7.9166 \cdot 10^{-1} \\ 9.5 \cdot 10^{-1} \\ 1.28 \cdot 10^0 \end{bmatrix} \quad \text{e} \quad v^{(3)} = \begin{bmatrix} -6.275 \cdot 10^{-1} \\ 1.1706 \cdot 10^0 \\ 1.5267 \cdot 10^0 \end{bmatrix}.$$

Para obter um erro inferior a 10^{-2} é necessário efetuar 13 iterações, ao final das quais obtemos

$$v^{(1)} = \begin{bmatrix} -7.6577 \cdot 10^{-1} \\ 1.0946 \cdot 10^0 \\ 1.4898 \cdot 10^0 \end{bmatrix}.$$

Outra maneira de decompor A , que leva a uma iteração ainda mais eficiente, é escolher

$$M = L + D, \quad K = -U \quad \text{e} \quad R = -(L + D)^{-1}U,$$

de modo que a iteração se torna

$$v^{(m+1)} = -(L + D)^{-1}Uv^{(m)} + (L + D)^{-1}b.$$

Na prática é preferível escrever a versão matricial do método de Gauss-Seidel na forma

$$(121) \quad Dv^{(m+1)} = Uv^{(m)} - Lv^{(m+1)} + b.$$

Isso porque, usando a notação de (114) e (115) podemos escrever as coordenadas de $v^{(m+1)}$ em (121) na forma

$$(122) \quad x_i^{(m+1)} = \frac{1}{a_{i,i}} \left(b_i - \sum_{k < i} a_{i,k} x_k^{(m+1)} - \sum_{k > i} a_{i,k} x_k^{(m)} \right).$$

O algoritmo de solução de sistemas lineares decorrente desta iteração é conhecido como *método de Gauss-Seidel*. Sugerindo a um amigo que usasse este método, Gauss escreveu:

Recomendo que imite este método. Você dificilmente precisará eliminar diretamente outra vez, pelo menos não quando houver mais do que duas equações. O procedimento indireto [o que hoje chamamos de método de Gauss-Seidel] pode ser executado mesmo quando se está meio dormindo, ou pensando em outra coisa.

A diferença entre os métodos de Gauss-Jacobi e Gauss-Seidel fica mais evidente quando comparamos (116) com (122). No método de Gauss-Jacobi, o valor de cada

$x_i^{(m+1)}$ depende diretamente dos $x_k^{(m)}$ da iteração anterior. Já no método de Gauss-Seidel, o valor de $x_i^{(m+1)}$ depende dos valores das coordenadas de $v^{(m+1)}$ calculados antes de $x_{n+1,i}$ (isto é, dos $x_k^{(m+1)}$ com $k < i$) e das coordenadas $v^{(m)}$ indexadas por números maiores que i (isto é, dos $x_k^{(m)}$ com $k > i$). Assim como o método de Gauss-Jacobi, o método de Gauss-Seidel também converge quando a matriz do sistema tem diagonal dominante.

Como a matriz A dada em (118) tem diagonal dominante, o método de Gauss-Seidel também vai convergir quando aplicado a A . Digamos que, como no exemplo do método de Gauss-Jacobi, queremos resolver o sistema $AX = b$, com $b = [1, 4, 7]^t$ usando o método de Gauss-Seidel a partir do valor inicial $v^{(0)} = [1, 1, 1]^t$. as três primeiras iterações nos dão, em ponto flutuante com três dígitos significativos

$$v^{(1)} = \begin{bmatrix} -7.25 \cdot 10^{-1} \\ 1.1195 \cdot 10^0 \\ 1.4661 \cdot 10^0 \end{bmatrix}, v^{(2)} = \begin{bmatrix} -7.6291 \cdot 10^{-1} \\ 1.0989 \cdot 10^0 \\ 1.4854 \cdot 10^0 \end{bmatrix} \quad \text{e} \quad v^{(3)} = \begin{bmatrix} -7.6742 \cdot 10^{-1} \\ 1.094 \cdot 10^0 \\ 1.4882 \cdot 10^0 \end{bmatrix}.$$

Neste exemplo, $v^{(3)}$ já é uma solução de $AX = b$ com erro inferior a 10^{-2} . Observe que obtivemos a solução do sistema com a tolerância desejada em apenas 3 iterações, contra as 13 que foram necessárias quando aplicamos a este mesmo sistema o método de Gauss-Jacobi.

CAPÍTULO 8

Integração

Ao contrário da impressão que seu curso de cálculo pode lhe ter dado, a maioria das funções elementares não admite primitiva, o que torna impossível calcular suas integrais usando o teorema fundamental do cálculo. A saída é determinar aproximações numéricas para estas integrais. Embora este seja o principal tema deste capítulo, começaremos estudando, um pouco mais a fundo, o problema da interpolação, porque os métodos de integração que analisaremos partem de uma aproximação do integrando por um polinômio interpolador. Em particular, para poder controlar o erro na integração numérica, precisamos saber como o erro se comporta na interpolação.

1. Interpolação pelo método de Lagrange

Seja

$$\mathcal{P} = \{(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)\},$$

um conjunto de pontos e seja $P(x)$ o polinômio de grau n cujo gráfico passa por cada um destes pontos.

Começaremos construindo um polinômio $L_0(x)$, de grau n , que satisfaça

$$L_0(x_i) = \begin{cases} 1 & \text{se } i = 0 \\ 0 & \text{se } i \neq 0. \end{cases}$$

Como um tal polinômio tem x_1, \dots, x_n como raízes, ele precisa ser da forma

$$L_0(x) = a(x - x_1) \cdots (x - x_n),$$

para algum número real não-nulo a . Por outro lado, $L_0(x_0) = 1$ implica que

$$a(x_0 - x_1) \cdots (x_0 - x_n) = 1;$$

donde

$$a = \frac{1}{(x_0 - x_1) \cdots (x_0 - x_n)},$$

o que nos permite concluir que

$$L_0(x) = \frac{(x - x_1) \cdots (x - x_n)}{(x_0 - x_1) \cdots (x_0 - x_n)}.$$

Argumentando de maneira semelhante, verificamos que se um polinômio $L_j(x)$, de grau n , satisfaz

$$(123) \quad L_j(x_i) = \begin{cases} 1 & \text{se } i = j \\ 0 & \text{se } i \neq j, \end{cases}$$

então

$$L_j(x) = \frac{(x - x_1) \cdots (x - x_{j-1})(x - x_{j+1}) \cdots (x - x_n)}{(x_j - x_1) \cdots (x_j - x_{j-1})(x_j - x_{j+1}) \cdots (x_j - x_n)}.$$

Note que no numerador e no denominador excluímos o fator $x - x_j$; isto é,

$$L_j(x) = \prod_{i \neq j} (x - x_i) / \prod_{i \neq j} (x_j - x_i).$$

Isto nos permite escrever o polinômio interpolador $P(x)$ na forma

$$(124) \quad P(x) = \sum_{j=0}^n y_j L_j(x),$$

porque, por (123),

$$P(x_i) = \sum_{j=0}^n y_j L_j(x_i) = y_i L_i(x_i) = y_i,$$

para todo $0 \leq i \leq n$, como a definição de interpolação requer. A fórmula (124) é conhecida como a *forma de Lagrange* do polinômio interpolador. Por exemplo, o polinômio que interpola os pontos

$$(1, 2), (2, 5) \text{ e } (7, 9)$$

é dado pela fórmula

$$P(x) = 2 \frac{(x-2)(x-7)}{(1-2)(1-7)} + 5 \frac{(x-1)(x-7)}{(2-1)(2-7)} + 9 \frac{(x-1)(x-2)}{(7-1)(7-2)}.$$

É importante observar que, na prática, não vale à pena multiplicar estes fatores para obter os coeficientes de $P(x)$. Por exemplo, para determinar $P(3)$, substituímos $x = 3$ na fórmula acima, obtendo

$$2 \frac{(3-2)(3-7)}{(1-2)(1-7)} + 5 \frac{(3-1)(3-7)}{(2-1)(2-7)} + 9 \frac{(3-1)(3-2)}{(7-1)(7-2)}$$

e só então efetuamos os cálculos para achar $P(3) = 109/15$.

Não é improvável que sua reação seja “por que você não me disse isso antes?” Afinal é muito mais fácil calcular o polinômio interpolador usando (124) do que resolvendo o sistema linear dado pela matriz de Vandermonde. A resposta é que, no capítulo 5, usamos a interpolação, basicamente, como motivação para o ajuste de curvas e como uma maneira natural de introduzir a matriz de Vandermonde.

A bem da verdade, o fato de termos introduzido dois procedimentos diferentes para calcular o polinômio interpolador põe imediatamente a pergunta: os polinômios calculados destas duas maneiras coincidem? Se você leu o início deste parágrafo com atenção, já deve ter desconfiado que a resposta é sim; do contrário não teríamos usado o artigo definido quando nos referimos ao polinômio interpolador. Para mostrar que existe um único polinômio interpolador, suporemos que $P(x)$ e $Q(x)$ são polinômios interpoladores para os pontos do conjunto \mathcal{P} do início desta seção. Por definição, isto significa que $P(x)$ e $Q(x)$ são ambos polinômios de grau n que satisfazem

$$P(x_i) = Q(x_i) = y_i \quad \text{para} \quad i = 0, \dots, n.$$

Mas isto implica que $P(x) - Q(x)$ é um polinômio de grau, no máximo, n que satisfaz

$$P(x_i) - Q(x_i) = y_i - y_i = 0 \quad \text{para} \quad i = 0, \dots, n.$$

Logo, $P(x) - Q(x)$ tem $n + 1$ raízes. Como um polinômio não-nulo de grau menor ou igual a n não pode ter mais de n raízes, concluímos que $P(x) = Q(x)$; isto é, que o polinômio interpolador é, de fato, único.

2. Interpolação pelo método de Newton e erros

Mesmo antes de Lagrange propor seu método de interpolação, Newton havia introduzido uma outra maneira de calcular o polinômio interpolador. Seja

$$\mathcal{P} = \{(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)\},$$

um conjunto de pontos e seja $P(x)$ o polinômio de grau n cujo gráfico passa por cada um destes pontos. No método de Lagrange, escrevemos

$$P(x) = y_0 L_0(x) + \dots + y_n L_n(x);$$

no método de Newton, $P(x)$ é escrito na forma

$$P(x) = c_0 \omega_0(x) + \dots + c_n \omega_n(x),$$

em que

$$\omega_j(x) = \begin{cases} 1 & \text{quando } j = 0 \\ (x - x_0) \cdots (x - x_j) & \text{quando } j > 0. \end{cases}$$

Para achar os valores dos coeficientes c_0, \dots, c_n , note que, se $j > i \geq 0$, então $x - x_i$ aparece como um dos fatores que multiplicamos para calcular $\omega_j(x)$. Logo,

$$\omega_j(x_i) = 0 \quad \text{sempre que } j > i \geq 0.$$

Portanto, pela definição do polinômio interpolador,

$$y_i = P(x_i) = \sum_{j=0}^{i-1} c_j \omega_j(x_i);$$

que nos dá um sistema triangular inferior cuja solução são os valores de c_0, \dots, c_{n-1} . Por exemplo, para obter o polinômio que interpola os pontos

$$(1, 2), (2, 5) \text{ e } (7, 9)$$

pelo método de Newton, construímos primeiro os polinômios

$$\omega_0(x) = 1, \quad \omega_1(x) = (x - 1) \quad \text{e} \quad \omega_2(x) = (x - 1)(x - 2).$$

Substituindo x por 1, 2 e 7 em

$$P_2(x) = c_0\omega_0(x) + c_1\omega_1(x) + c_2\omega_2(x),$$

obtemos as equações lineares

$$2 = P_2(1) = c_0$$

$$5 = P_2(2) = c_0 + c_1$$

$$9 = P_2(7) = c_0 + 6c_1 + 30c_2.$$

Resolvendo este sistema por substituição direta, obtemos

$$c_0 = 2, \quad c_1 = 3 \quad \text{e} \quad c_2 = -\frac{11}{30};$$

donde podemos concluir que

$$P_2(x) = 2 + 3\omega_1 - \frac{11}{30}\omega_2(x)$$

é o polinômio interpolador desejado.

Como já sabemos que o polinômio interpolador é único, obtivemos, com isto, apenas uma terceira maneira de achá-lo. Maneira esta, você deve estar pensando, que é mais complicada de executar, ao menos com papel e lápis, que a proposta por Lagrange. Embora isto seja, até certo ponto, verdadeiro, o procedimento de Newton tem algumas vantagens.

Para ilustrar uma destas vantagens, imagine que, após ter calculado o polinômio $P_n(x)$, que interpola os pontos

$$(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$$

chegamos à conclusão de que a curva polinomial que desejamos também deve passar pelo ponto (x_{n+1}, y_{n+1}) . Para achar $P_{n+1}(x)$ pelo método de Newton devemos determinar números reais c_0, \dots, c_n tais que

$$(125) \quad P_{n+1}(x) = c_0\omega_0(x) + \dots + c_n\omega_n(x) + c_{n+1}\omega_{n+1}(x).$$

Contudo, como o sistema que usamos para achar os c s é triangular inferior, podemos concluir que

$$c_0\omega_0(x_i) + \dots + c_n\omega_n(x_i) = 0$$

para todo $0 \leq i \leq n-1$. Assim, pela unicidade do polinômio interpolador de grau n ,

$$P_n(x) = c_0\omega_0(x) + \cdots + c_n\omega_n(x);$$

de modo que a equação (125) pode ser escrita na forma

$$(126) \quad P_{n+1}(x) = P_n(x) + c_{n+1}\omega_{n+1}(x).$$

Como estamos supondo que já conhecemos $P_n(x)$, temos que

$$(127) \quad c_n = \frac{P_{n+1}(x_{n+1}) - P_n(x_{n+1})}{\omega_{n+1}(x_{n+1})} = \frac{y_{n+1} - P_n(x_{n+1})}{\omega_{n+1}(x_{n+1})},$$

já que queremos que $P_{n+1}(x)$ também passe por (x_{n+1}, y_{n+1}) . Por exemplo, já sabemos que a curva polinomial que passa pelos pontos

$$(1, 2), (2, 5) \text{ e } (7, 9),$$

é definida por

$$P_2(x) = 2 + 3\omega_1 - \frac{11}{30}\omega_2(x).$$

Se quisermos uma curva que, além destes pontos, passe também por $(3, 1)$, basta calcular

$$c_3 = \frac{1 - P_2(3)}{(3-1)(3-2)(3-7)}.$$

Como

$$P_2(3) = \frac{109}{15},$$

obtemos

$$c_3 = \frac{47}{60};$$

donde

$$P_3(x) = 2 + 3\omega_1 - \frac{11}{30}\omega_2(x) + \frac{47}{60}\omega_3(x).$$

A propósito, neste exemplo, o ponto que foi acrescentado não está à direita do último ponto usado para calcular $P_2(x)$. Na verdade, o novo ponto pode estar em qualquer lugar: à esquerda, à direita ou entre os pontos usados para calcular o primeiro polinômio.

Introduzimos o método de Newton em grande parte porque nos permite chegar facilmente à fórmula para o erro absoluto cometido quando usamos o polinômio interpolador para aproximar o gráfico de uma função, à semelhança do que fizemos para o polinômio de Taylor. Digamos que seja dada uma função $f : [a, b] \rightarrow \mathbb{R}$ e números reais

$$x_0, \dots, x_n \in [a, b].$$

Podemos usar o polinômio $P_n(x)$ que interpola os pontos

$$(128) \quad (x_0, f(x_0)), \dots, (x_n, f(x_n))$$

como uma aproximação de $f(x)$. Como já sabemos que isto pode produzir sérios problemas, queremos ser capazes de estimar o erro absoluto que cometeríamos se usássemos $P_n(\alpha)$ para aproximar $f(\alpha)$. Isto é, queremos achar uma cota superior para o módulo do erro

$$E_n(\alpha) = f(\alpha) - P_n(\alpha).$$

Note que

$$E_n(x_i) = f(x_i) - P_n(x_i) = 0$$

para $i = 0, 1, \dots, n$. Neste ponto precisamos usar o teorema de Rolle generalizado, que enunciamos abaixo.

TEOREMA DE ROLLE GENERALIZADO. *Seja $h : [a, b] \rightarrow \mathbb{R}$ uma função cujas $n + 1$ primeiras derivadas existem e são contínuas. Se $h(x)$ tem $n + 1$ zeros no intervalo (a, b) , então sua $n + 1$ -ésima derivada se anula em algum ponto de (a, b) .*

Naturalmente, para que o teorema possa ser aplicado, precisamos supor que a função $f(x)$ que estamos considerando tem suas $n + 1$ primeiras derivadas contínuas. Entretanto, ao contrário do que você pode estar esperando, não aplicaremos o teorema a $E_n(x)$, mas sim a

$$(129) \quad E_{n+1}(x) = f(x) - P_{n+1}(x),$$

em que $P_{n+1}(x)$ interpola o conjunto obtido acrescentando-se $(\alpha, f(\alpha))$ aos pontos

$$(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n).$$

Como

$$E_{n+1}(x_0) = \dots = E_{n+1}(x_n) = E_{n+1}(\alpha) = 0,$$

o teorema de Rolle generalizado garante que existe $\xi \in (a, b)$ tal que

$$(130) \quad \frac{d^{n+1}E_{n+1}}{dx^{n+1}}(\xi) = 0.$$

Por outro lado, (129) nos dá

$$\frac{d^{n+1}E_{n+1}}{dx^{n+1}}(\xi) = \frac{d^{n+1}f}{dx^{n+1}}(\xi) - \frac{d^{n+1}P_{n+1}}{dx^{n+1}}(\xi),$$

o que nos permite deduzir de (130) que

$$f^{(n+1)}(\xi) = \frac{d^{n+1}f}{dx^{n+1}}(\xi) = \frac{d^{n+1}P_{n+1}}{dx^{n+1}}(\xi).$$

Mas, derivando

$$P_{n+1}(x) = P_n(x) + c_{n+1}\omega_{n+1}(x).$$

$n + 1$ vezes e levando em conta que $P_n(x)$ tem grau menor que n e que o coeficiente de x^n em $\omega_n(x)$ é igual a um, obtemos

$$(131) \quad \frac{d^{n+1}P_{n+1}}{dx^{n+1}}(x) = (n + 1)!c_{n+1}.$$

Assim,

$$f^{(n+1)}(\xi) = \frac{d^{n+1}P_{n+1}}{dx^{n+1}}(\xi) = (n+1)!c_{n+1}$$

donde

$$E_{n+1}(x) = f(x) - P_n(x) + c_{n+1}\omega_{n+1}(x) = E_n(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x).$$

Substituindo x por α nesta equação e lembrando que $E_{n+1}(\alpha) = 0$,

$$0 = E_n(\alpha) + \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(\alpha);$$

de modo que

$$E_n(\alpha) = -\frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(\alpha).$$

Note que o valor de ξ depende de α . Para chamar a atenção para isto escreveremos, frequentemente, $\xi(\alpha)$ em vez de ξ . Resumimos isto abaixo, para referência futura.

ERRO NA INTERPOLAÇÃO. *Seja $f : [a, b] \rightarrow \mathbb{R}$ uma função cujas $n+1$ primeiras derivadas existem e são contínuas e sejam $x_0, \dots, x_n \in (a, b)$. Se $P_n(x)$ é o polinômio que interpola estes $n+1$ pontos e $\alpha \in (a, b)$, então, para algum $\xi(\alpha) \in (a, b)$, cujo valor depende de α ,*

$$(132) \quad f(\alpha) - P_n(\alpha) = -\frac{f^{(n+1)}(\xi(\alpha))}{(n+1)!}\omega_{n+1}(\alpha),$$

em que $\omega_{n+1}(x) = (x - x_0) \cdots (x - x_n)$.

Como vimos no capítulo 3 uma das maneiras mais fáceis de traçar uma curva é aproximá-la por uma poligonal formada a partir de pontos bastante próximos que estejam sobre a curva. Em termos analíticos, estes segmentos correspondem a polinômios interpoladores lineares entre dois pontos da curva. Agora que temos uma fórmula para o erro na interpolação, podemos estimar o erro que é cometido quando usamos estes segmentos como aproximações para a curva. Para isso, suporemos que a curva é o gráfico de $y = f(x)$ e que $f : [a, b] \rightarrow \mathbb{R}$ tem segunda derivada contínua. Por (132), o erro cometido quando aproximamos o gráfico da função por um segmento de reta entre dois pontos $x_0 < x_1$ do intervalo $[a, b]$ é dado, para um ponto $x \in (x_0, x_1)$, por

$$E_1(x) = \frac{f''(\xi(x))}{2!}(x - x_0)(x - x_1),$$

em que $\xi(x) \in (x_0, x_1)$. Mas, por hipótese, a segunda derivada de f é contínua no intervalo $[x_0, x_1] \subseteq [a, b]$. Portanto, $|f''(x)|$ atinge um valor máximo $M > 0$ em $[x_0, x_1]$, o que nos permite escrever

$$|E_1(x)| \leq \frac{M|(x - x_0)(x - x_1)|}{2!}.$$

Por outro lado, a derivada da função $\omega_1(x) = (x - x_0)(x - x_1)$ é

$$g'(x) = 2x - (x_0 + x_1),$$

de modo que a parábola $y = \omega_1(x)$ atinge seu mínimo em $\bar{x} = (x_0 + x_1)/2$. Como

$$g(\bar{x}) = -\frac{(x_1 - x_0)^2}{4} < 0,$$

e $g(x_0) = g(x_1) = 0$, podemos concluir que

$$|\omega_1(x)| \leq |g(\bar{x})| = \frac{(x_1 - x_0)^2}{4},$$

para $x \in [x_0, x_1]$. Portanto,

$$|E_1(x)| \leq \frac{M|g(\bar{x})|}{2!} \leq \frac{M(x_1 - x_0)^2}{8},$$

qualquer que seja $x \in [x_0, x_1]$. Por esta fórmula, se usarmos um segmento de reta para aproximar $\sin(x)$ em um intervalo de comprimento 0.1, cometeremos um erro inferior a

$$\frac{(0.1)^2}{8} = 0.00125,$$

pois a $|\sin''(x)| = |\sin(x)| \leq 1$. Por exemplo, se $x_1 - x_0$ mede 1 mm, o erro será inferior a 0.0125 mm. Como nosso olho não consegue distinguir um erro tão pequeno, a curva polinomial obtida ligando pontos da forma $(x_i, \sin(x_i))$ cujas abscissas distam de 1 mm será indistinguível do gráfico de $y = \sin(x)$ no mesmo intervalo.

3. Integração: regra do trapézio

Ao contrário das ideias-chaves do cálculo diferencial, só descobertas no século XVI, o cálculo de áreas por aproximações poligonais, que está na raiz do cálculo integral, já era conhecido dos gregos antigos. Na proposição 2 do livro XII dos *Elementos* de Euclides, a área do círculo é calculada usando polígonos inscritos e circunscritos, ao passo que Arquimedes dedicou o tratado *Quadratura da parábola* ao cálculo da área de um setor de parábola. Na seção 5 do capítulo 6, estudamos a regra do retângulo, que é uma adaptação ao cálculo de áreas sob curvas do método utilizado por Euclides e Arquimedes. Nesta seção veremos que, ao utilizar trapézios em vez de retângulos, podemos controlar o erro cometido na aproximação a partir da fórmula do erro de interpolação.

Seja $f : [a, b] \rightarrow \mathbb{R}$ uma função contínua e seja n um inteiro positivo. Como já se tornou usual, denotaremos por x_i o ponto $a + ih$, em que $h = (b - a)/n$. Para calcular a integral de $f(x)$ entre a e b aproximaremos a curva $y = f(x)$ pelo segmento

de reta que liga $(x_i, f(x_i))$ a $(x_{i+1}, f(x_{i+1}))$. A área sob a curva $y = f(x)$ no intervalo $[x_i, x_{i+1}]$, será, então, aproximada pelo trapézio com vértices em

$$(x_i, 0), (x_{i+1}, 0), (x_i, f(x_i)) \text{ e } (x_{i+1}, f(x_{i+1})).$$

Para implementar esta estratégia, começamos calculando o polinômio interpolador entre $(x_i, f(x_i))$ e $(x_{i+1}, f(x_{i+1}))$ e o erro incorrido em utilizá-lo como aproximação de $y = f(x)$ entre x_i e x_{i+1} . Usando a fórmula de Lagrange, o polinômio interpolador é

$$P_i(x) = f(x_i) \frac{x - x_{i+1}}{x_i - x_{i+1}} + f(x_{i+1}) \frac{x - x_i}{x_{i+1} - x_i}$$

e o erro correspondente é dado por

$$(133) \quad f(x) - P_i(x) = -\frac{f''(\xi_i(x))}{2} \omega_i(x),$$

em que

$$x, \xi_i(x) \in [x_i, x_{i+1}] \quad \text{e} \quad \omega_i(x) = (x - x_i)(x - x_{i+1})$$

Note que estamos utilizando o índice i , subscrito aos polinômios $P_i(x)$ e $\omega_i(x)$, para indicar que correspondem à interpolação no intervalo $[x_i, x_{i+1}]$ e não para indicar o grau do polinômio, como fizemos, algumas vezes, em seções anteriores. Integrando os dois lados de (133) entre x_i e x_{i+1} , obtemos

$$\int_{x_i}^{x_{i+1}} f(t) dt - \int_a^b P_i(t) dt = -\frac{1}{2} \int_a^b f''(\xi_i(t)) \omega_i(x) dt.$$

Levando em conta que $x_{i+1} - x_i = h$, obtemos de

$$\int_{x_i}^{x_{i+1}} P_i(t) dt = \frac{f(x_i)}{h} \int_{x_i}^{x_{i+1}} (t - x_{i+1}) dt - \frac{f(x_{i+1})}{h} \int_{x_i}^{x_{i+1}} (t - x_i) dt,$$

que

$$\int_{x_i}^{x_{i+1}} P_i(t) dt = \frac{h}{2} (f(x_i) + f(x_{i+1}));$$

donde

$$(134) \quad \int_{x_i}^{x_{i+1}} f(t) dt = \frac{h}{2} (f(x_i) + f(x_{i+1})) - \frac{1}{2} \int_{x_i}^{x_{i+1}} f''(\xi_i(t)) \omega_i(x) dt.$$

Este ainda não é o final da história, porque só calculamos o valor da área em um dos segmentos em que dividimos o intervalo de integração; mas, antes de prosseguir, vamos fazer um exemplo bem simples. Digamos que queremos integrar $\sin(x)$ no intervalo $[0, 2]$ e que, para isso, dividimos este intervalo em 20 partes iguais, de modo que

$$h = \frac{2}{20} = 0.1.$$

Usando a fórmula (134), obtemos a seguinte aproximação da integral do seno no subintervalo $[0, 0.1]$,

$$\int_0^{0.1} \sin(t) dt \approx \frac{0.1}{2} (\sin(0) + \sin(0.1)) = 0.0049916708.$$

A mesma fórmula também nos permite estimar o erro cometido nesta aproximação como

$$\left| \frac{1}{2} \int_0^{0.1} \sin''(\xi_i(t)) t(t-0.1) dt \right| \leq \frac{1}{2} \int_0^{0.1} |\sin''(\xi_i(t))| |t(t-0.1)| dt.$$

Levando em conta que

$$|\sin''(x)| = |\sin(x)| \leq 1$$

e que $|t(t-0.1)| = t(0.1-t)$, quando $t \in [0, 0.1]$, concluímos que o erro, neste exemplo, é inferior a

$$\frac{1}{2} \int_0^{0.1} |\sin''(\xi_i(t))| |t(t-0.1)| dt \leq \frac{1}{2} \int_0^{0.1} t(0.1-t) dt = 0.8 \cdot 10^{-4}.$$

Entretanto, sabemos que

$$\int_0^{0.1} \sin(t) dt = -\cos(t) \Big|_{t=0}^{0.1} = 0.0049958347,$$

de modo que o erro realmente incorrido neste exemplo é

$$|0.0049958347 - 0.0049916708| = 0.4 \cdot 10^{-5},$$

que é um pouco menos de um décimo do error estimado.

Voltando ao caso geral, lembre-se que o que a fórmula (134) nos dá é uma aproximação para a integral de $f(x)$ em um dos pequenos intervalo $[x_i, x_{i+1}]$ em que subdividimos $[a, b]$. Porém, como

$$\int_a^b f(t) dt = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(t) dt,$$

a fórmula da integral desejada, e seu erro, podem ser obtidos somando as fórmulas para cada um dos subintervalos; isto é,

$$(135) \quad \int_a^b f(t) dt = \frac{h}{2} \sum_{i=0}^{n-1} (f(x_i) + f(x_{i+1})) - \frac{1}{2} \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f''(\xi_i(t)) dt.$$

Porém, o primeiro somatório na fórmula acima é igual a

$$f(x_0) + \underbrace{f(x_1) + f(x_1)} + \underbrace{f(x_2) + f(x_2)} + \underbrace{f(x_3) + \cdots + f(x_{n-1})} + f(x_n),$$

no qual cada parcela diferente de $f(x_0)$ e $f(x_n)$ aparece duas vezes. Isto nos permite reescrever (135) como

$$(136) \quad \int_a^b f(t)dt = \frac{h}{2} \left(f(x_0) + 2 \sum_{i=1}^{n-1} f(x_i) + f(x_n) \right) - \frac{1}{2} \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f''(\xi_i(t))dt,$$

que é mais fácil de lembrar, exceto pelo o termo do erro. Para podermos simplificar este último, note que uma das poucas coisas que sabemos sobre a função ξ_i é que $\xi_i(x) \in [x_i, x_{i+1}]$, quando $x \in (x_i, x_{i+1})$. Portanto, realisticamente falando, o melhor a fazer é substituir $f''(\xi_i(t))$ pelo máximo M de $|f''(x)|$ no intervalo $[a, b]$. Fazendo isto, obtemos

$$(137) \quad \left| \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f''(\xi_i(t))(t - x_i)(t - x_{i+1})dt \right| \leq M \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} |(t - x_i)(t - x_{i+1})|dt.$$

Contudo, como $|(t - x_i)(t - x_{i+1})| = (t - x_i)(x_{i+1} - t)$,

$$\int_{x_i}^{x_{i+1}} |(t - x_i)(t - x_{i+1})|dt = \int_{x_i}^{x_{i+1}} (t - x_i)(x_{i+1} - t)dt = \frac{h^3}{6},$$

para $i = 0, \dots, n-1$. Substituindo isto em (137), descobrimos que o erro no cálculo da integral de $f(x)$ em $[a, b]$ é menor ou igual que

$$M \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} |(t - x_i)(t - x_{i+1})|dt = \frac{nh^3M}{6} = \frac{h^2(b-a)M}{6},$$

pois $nh = (b-a)$. Obtemos, com isto, a seguinte desigualdade para o valor absoluto do erro na fórmula (136)

$$\left| \frac{1}{2} \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f''(\xi(t))dt \right| \leq \frac{h^2(b-a)M}{12}$$

Resumindo, temos a seguinte regra de integração.

REGRA DO TRAPÉZIO. *Sejam $n > 0$ um número inteiro e $f : [a, b] \rightarrow \mathbb{R}$ uma função cujas primeiras duas derivadas existem e são contínuas. Se $h = (b-a)/n$, então*

$$(138) \quad \frac{h}{2} \left(f(x_0) + 2 \sum_{i=1}^{n-1} f(x_i) + f(x_n) \right)$$

é uma aproximação da integral de $f(x)$ no intervalo $[a, b]$, com erro inferior a

$$(139) \quad \frac{h^2(b-a)M}{12},$$

em que M é o valor máximo de $f''(x)$ no intervalo $[a, b]$.

Aplicando a regra do trapézio ao exemplo que consideramos anteriormente, temos que

$$\int_0^2 \sin(t) dt \approx \frac{0.1}{2} (\sin(0) + 2 \sum_{i=1}^{n-1} \sin(i/10) + \sin(2)) \approx 1.4149665174,$$

com erro inferior a

$$\frac{(0.1)^2 \cdot 2 \cdot 1}{12} = 0.0016666666$$

pois

$$M = \max\{|\sin(x)| \mid x \in [0, 2]\} = 1,$$

já que $\pi/2 \in [0, 2]$ e $\sin(\pi/2) = 1$. Contudo,

$$\int_0^2 \sin(t) dt = -\cos(t) \Big|_{t=0}^{t=2} = 1.4161468365,$$

de modo que o erro realmente incorrido nesta aproximação é

$$|1.4149665174 - 1.4161468365| = 0.0011803191.$$

Para nosso próximo exemplo, continuamos considerando a integral de $\sin(x)$ em $[0, 2]$, mas desta vez nosso objetivo é calculá-la com erro inferior a 10^{-6} . Para isto, partimos da estimativa para o erro na regra do trapézio, quando $[0, 2]$ é dividido em n partes iguais. Como sabemos que o máximo da segunda derivada do seno em $[0, 2]$ é igual a um, a fórmula (139) nos diz que o erro é inferior a

$$\frac{h^2(2-0)M}{12} \leq \frac{h^2}{6} \leq \frac{2}{3n^2},$$

pois $h = 2/n$. Portanto, para que o erro seja menor que 10^{-6} , basta que

$$\frac{2}{3n^2} < 10^{-6};$$

isto é, que $n > 1154.7$. Como n tem que ser inteiro, precisamos tomar $n = 1155$ para garantir que o valor desta integral, calculado pelo método do trapézio, seja inferior a 10^{-6} . Executando os cálculos para este valor de n , obtemos que

$$\int_0^2 \sin(t) dt \approx 1.4161461293,$$

que corresponde a um erro real de

$$|1.4161461293 - 1.4161468365| = 0.7072 \cdot 10^{-6} < 10^{-6}$$

como desejado.

Encerraremos a seção usando a regra do trapézio para calcular a integral

$$\int_0^1 \exp(-t^2) dt.$$

Na verdade esta é a única saída para calcular esta integral, porque $f(x) = \exp(-x^2)$ não admite primitiva, veja [11, p. 971] para mais detalhes. Digamos que queremos calcular o valor desta integral com erro inferior a 10^{-6} . De acordo com (139), o erro cometido no cálculo desta integral é inferior a

$$\frac{Mh^2}{12},$$

em que M é o máximo de

$$f''(x) = \frac{d^2 \exp(-x^2)}{dx^2} = (4x^2 - 2) \exp(-x^2)$$

no intervalo $[0, 1]$. Mas,

$$f'''(x) = \frac{d^3 \exp(-x^2)}{dx^3} = (-8x^2 + 12)x \exp(-x^2)$$

é positiva entre as raízes

$$\pm \sqrt{\frac{3}{2}} \approx \pm 1.2247$$

de $-8x^2 + 12 = 0$. Logo, a segunda derivada de $\exp(-x^2)$ é crescente no intervalo $[0, 1]$. Portanto,

$$f''(x) \leq f''(1) = \frac{2}{e} \leq 0.73576,$$

de modo que o erro no cálculo da integral é inferior a

$$\frac{Mh^2}{12} \leq \frac{0.73576 \cdot h^2}{12} \leq 0.06132 \frac{1}{n^2}.$$

Fazendo

$$0.06132 \frac{1}{n^2} \leq 10^{-6}$$

obtemos $n \geq 247.6288$. Logo, como n tem que ser inteiro, precisamos que seja maior ou igual que 248. Aplicando o método do trapézio com $n = 248$ e $h = 1/248$, obtemos

$$\int_0^1 \exp(-t^2) dt \approx 0.74682313591347$$

como a aproximação desejada.

4. Regra de Simpson

Como vimos na seção anterior, na regra do trapézio calculamos a aproximação de uma integral substituindo o integrando pelo polinômio interpolador de grau um. Nesta seção veremos que, usando um polinômio interpolador de grau dois, é possível obter um algoritmo numérico de integração mais eficiente do que a regra do trapézio.

Como na seção anterior, suporemos que $f : [a, b] \rightarrow \mathbb{R}$ é uma função que, desta vez, precisa ter suas três primeiras derivadas contínuas. Como a interpolação por um polinômio de grau dois requer três pontos a partir dos quais definir o polinômio, teremos que subdividir o intervalo $[a, b]$ em uma quantidade par de partes iguais. Se tomarmos $n = 2m$, teremos que

$$h = \frac{b-a}{2m} \quad \text{e} \quad x_i = a + ih, \quad \text{para} \quad i = 0, \dots, 2m-1.$$

Se fôssemos repetir os passos que seguimos para obter a regra do trapézio, a próxima coisa a fazer seria aplicar a fórmula (132) da página 167 para escrever $f(x)$ em termos do polinômio interpolador de *grau dois* e de seu erro. Integrando os dois lados obteríamos a fórmula para integração numérica que buscamos. Embora sejam estes os passos que vamos seguir, nosso ponto de partida será a fórmula correspondente ao polinômio interpolador de *grau três*. Como o que está escrito acima pode parecer meio paradoxal, vou repetir para não haver sombra de dúvidas:

embora a fórmula que vamos obter seja derivada a partir do polinômio interpolador de grau dois, que usa três nós consecutivos, a dedução parte de uma aplicação da fórmula de interpolação supondo que o polinômio interpolador tenha grau três.

Como se esta afirmação já não fosse confusa o bastante, nos deparamos imediatamente com outro problema: para obter um polinômio interpolador de grau três precisamos de quatro pontos, mas na afirmação em destaque acima nos referimos apenas a “três nós consecutivos”; *três* e não quatro. Resolvemos este problema acrescentando um quarto ponto $\eta \in [x_{2j}, x_{2(j+1)}]$, *que precisa apenas ser diferente dos outros três nós*. Sejam $Q_j(x)$ o polinômio que interpola os pontos

$$(x_{2j}, f(x_{2j})), (x_{2j+1}, f(x_{2j+1})) \text{ e } (\eta, f(\eta)),$$

e $\omega_j(x, \eta)$ o polinômio

$$(x - x_{2j})(x - x_{2j+1})(x - x_{2(j+1)})(x - \eta).$$

Note que, como na seção anterior, os índices em $Q_j(x)$ e $\omega_j(x, \eta)$ indicam qual das triplas de nós consecutivos estão sendo usadas para construir este polinômios, e não seu grau. Talvez você tenha estranhado que η tenha aparecido como um dos argumentos de ω_j , junto com x . Afinal, η é só um nó adicional que usamos para construir os polinômios; se não usamos x_{2j} , x_{2j+1} e $x_{2(j+1)}$ como argumentos de ω_j , por que usar η ? A verdade é que, como veremos adiante, η não é um nó em pé de igualdade com os outros três, porque x_{2j} , x_{2j+1} e $x_{2(j+1)}$ não podem ser alterados depois que a subdivisão do intervalo foi construída, mas η está livre e pode (e vai!) variar livremente em $[x_{2j}, x_{2(j+1)}]$, *desde que não coincida com nenhum dos outros três nós deste intervalo*.

Aplicando a fórmula (132) a x_{2j} , x_{2j+1} , $x_{2(j+1)}$ e η , obtemos

$$(140) \quad f(x) = Q_j(x) - \frac{f^{(iv)}(\xi(x))}{(n+1)!} \omega_j(x), \quad \text{para } j = 0, \dots, k,$$

em que x e $\xi_j(x)$ pertencem a $[x_j, x_{2j}]$. Integrando os dois lados de (140), obtemos

$$(141) \quad \int_{x_{2j}}^{x_{2(j+1)}} f(t) dt = \int_{x_{2j}}^{x_{2(j+1)}} Q_j(t) dt - \int_{x_{2j}}^{x_{2(j+1)}} \frac{f^{(iv)}(\xi_j(t))}{4!} \omega_j(t, \eta) dt.$$

Nosso próximo passo consistirá em calcular as integrais do lado direito. Porém, em vez de calcular $Q_j(x)$ usando o método de Lagrange, vamos usar a equação (126) para escrevê-lo na forma

$$Q_j(x) = P_j(x) + c_3(x - x_{2j})(x - x_{2j} - h)(x - x_{2j} - 2h),$$

em que c_3 é um número real que, como você logo verá, não precisamos calcular. Integrando os dois lados desta última equação, obtemos

$$\int_{x_{2j}}^{x_{2(j+1)}} Q_j(t) dt = \int_{x_{2j}}^{x_{2(j+1)}} P_j(t) dt + c_3 \int_{x_{2j}}^{x_{2(j+1)}} (t - x_{2j})(t - x_{2j} - h)(t - x_{2j} - 2h) dt.$$

Com isto estamos prontos para descobrir porque escolhemos partir da fórmula para o polinômio interpolador de grau três. Como você pode constatar da figura 1, o gráfico de $y = (x - x_{2j})(x - x_{2j} - h)(x - x_{2j} - 2h)$ é formado de duas partes idênticas, uma negativa, a outra positiva.

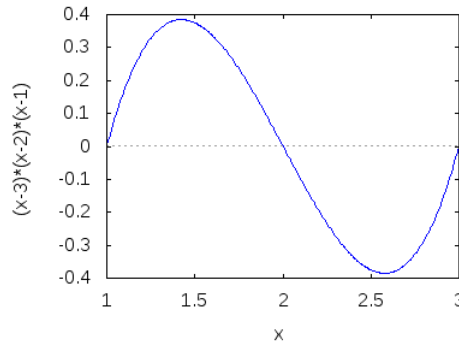


FIGURA 1. Gráfico de $y = (x - x_{2j})(x - x_{2j} - h)(x - x_{2j} - 2h)$.

Resulta disto que

$$\int_{x_{2j}}^{x_{2(j+1)}} (t - x_{2j})(t - x_{2j} - h)(t - x_{2j} - 2h) dt = 0,$$

como é fácil de constatar, expandindo o integrando e efetuando a integração. Como consequência disto,

$$\int_{x_{2j}}^{x_{2(j+1)}} Q_j(t) dt = \int_{x_{2j}}^{x_{2(j+1)}} P_j(t) dt.$$

A grande vantagem de proceder desta maneira é que, como o erro em (140) depende da quarta derivada de $f(x)$, podemos obter uma estimativa melhor do que a que teríamos conseguido usando a fórmula da interpolação para $P_j(x)$; veja exercício ???. Integrando $P_j(x)$ no intervalo $[x_j, x_{2j}]$, obtemos

$$\int_{x_{2j}}^{x_{2(j+1)}} P_j(t) dt = \frac{h}{3} (f(x_{2j+2}) + 4f(x_{2j+1}) + f(x_{2j})).$$

Em seguida precisamos analisar o comportamento do erro, que é igual ao valor absoluto de

$$E_j = \int_{x_{2j}}^{x_{2(j+1)}} \frac{f^{(iv)}(\xi_j(t))}{4!} \omega_j(t, \eta) dt.$$

Um detalhe sutil, mais importante, é que embora $\omega_j(t, \eta)$ dependa de η , o erro não depende. Para constatar que isto é verdade, basta notar que (141) também nos permite escrever

$$E_j = \int_{x_{2j}}^{x_{2(j+1)}} Q_j(t) dt - \int_{x_{2j}}^{x_{2(j+1)}} f(t) dt = \int_{x_{2j}}^{x_{2(j+1)}} P_j(t) dt - \int_{x_{2j}}^{x_{2(j+1)}} f(t) dt,$$

que é uma expressão na qual η não aparece em lugar algum.

Como no caso da regra do trapézio, nosso ponto de partida para chegar à estimativa do erro é a desigualdade

$$|E_j| = \left| \int_{x_{2j}}^{x_{2(j+1)}} \frac{f^{(iv)}(\xi_j(t))}{4!} \omega_j(t, \eta) dt \right| \leq \int_{x_{2j}}^{x_{2(j+1)}} \frac{|f^{(iv)}(\xi_j(t))|}{4!} |\omega_j(t, \eta)| dt,$$

Supondo que conhecemos

$$M = \max\{|f^{(iv)}(x)| \mid x \in [a, b]\},$$

temos que

$$\int_{x_{2j}}^{x_{2(j+1)}} \frac{|f^{(iv)}(\xi_j(t))|}{4!} |\omega_j(t, \eta)| dt \leq \frac{M}{4!} \int_{x_{2j}}^{x_{2(j+1)}} |\omega_j(t, \eta)| dt,$$

donde

$$(142) \quad |E_j| \leq \frac{M}{4!} \int_{x_{2j}}^{x_{2(j+1)}} |\omega_j(t, \eta)| dt,$$

para qualquer escolha legítima de η que desejemos fazer. O que complica a análise do erro é que $|\omega_j(t, \eta)|$ será igual a $\omega_j(t, \eta)$ ou a $-\omega_j(t, \eta)$, dependendo de onde $x \in [x_{2j}, x_{2(j+1)}]$ está, relativamente a x_{2j+1} e η . Porém, como o integrando de

$$\int_{x_{2j}}^{x_{2(j+1)}} |\omega_j(t, \eta)| dt$$

é contínuo, segue-se que

$$\lim_{\eta \rightarrow x_{2j+1}} \int_{x_{2j}}^{x_{2(j+1)}} |\omega_j(t, \eta)| dt = \int_{x_{2j}}^{x_{2(j+1)}} \lim_{\eta \rightarrow x_{2j+1}} |\omega_j(t, \eta)| dt = \int_{x_{2j}}^{x_{2(j+1)}} |\omega_j(t, x_{2j+1})| dt.$$

Contudo,

$$\omega_j(t, x_{2j+1}) = (t - x_{2j})(t - x_{2j+1})^2(t - x_{2(j+1)})$$

é negativo *qualquer que seja* $t \in [x_{2j}, x_{2(j+1)}]$, de modo que

$$\int_{x_{2j}}^{x_{2(j+1)}} |\omega_j(t, x_{2j+1})| dt = - \int_{x_{2j}}^{x_{2(j+1)}} \omega_j(t, x_{2j+1}) dt = \frac{4h^5}{15}.$$

Substituindo isto em (142), obtemos

$$|E_j| \leq \frac{M}{4!} \cdot \frac{4h^5}{15} = \frac{Mh^5}{90}$$

que é a estimativa que estávamos buscando.

Para obter a regra de Simpson basta somarmos

$$\int_{x_{2j}}^{x_{2(j+1)}} f(t) dt = \frac{h}{3} (f(x_{2j+2}) + 4f(x_{2j+1}) + f(x_{2j})) - E_j$$

para $j = 0, \dots, m-1$, que nos dá

$$\int_a^b f(t) dt = \frac{h}{3} \sum_{j=0}^{m-1} (f(x_{2(j+1)}) + 4f(x_{2j+1}) + f(x_{2j})) - \sum_{j=0}^{m-1} E_j.$$

O primeiro somatório pode ser reescrito, de maneira mais compacta, na forma

$$f(a) + f(b) + 2 \sum_{i \text{ par}} f(x_i) + 4 \sum_{i \text{ ímpar}} f(x_i),$$

ao passo que, para o erro, obtemos

$$\left| \sum_{j=0}^{m-1} E_j \right| \leq \sum_{j=0}^{m-1} |E_j| \leq m \frac{Mh^5}{90} = \frac{M(b-a)h^4}{180},$$

pois

$$mh = \frac{nh}{2} = \frac{b-a}{2}.$$

Resumindo, o argumento acima nos dá a seguinte regra de integração numérica.

REGRA DE SIMPSON. *Sejam $n > 0$ um número inteiro par e $f : [a, b] \rightarrow \mathbb{R}$ uma função cujas primeiras quatro derivadas existem e são contínuas. Se $h = (b - a)/n$, então*

$$(143) \quad \frac{h}{3} \left(f(a) + f(b) + 2 \sum_{i \text{ par}} f(x_i) + 4 \sum_{i \text{ impar}} f(x_i) \right),$$

é uma aproximação da integral de $f(x)$ no intervalo $[a, b]$, com erro inferior a

$$(144) \quad \frac{M(b-a)h^4}{180},$$

em que M é o valor máximo de $f^{(iv)}(x)$ no intervalo $[a, b]$.

Para encerrar, recalcularemos os dois exemplos do final da seção 3 usando a regra de Simpson. Como no caso da regra do trapézio, começamos nosso cálculo da integral de $\sin(x)$ determinando em quantas partes o intervalo $[0, 2]$ tem que ser dividido para que o erro seja menor ou igual que 10^{-6} . Como a quarta derivada de $\sin(x)$ é $-\cos(x)$, cujo módulo é sempre menor ou igual a 1, a fórmula (144) requer que tomemos

$$\frac{2h^4}{180} = \frac{2^4}{90n^4} < 10^{-6},$$

donde $n > 20.5338$. Como n tem que ser um inteiro par, o menor valor que podemos escolher é $n = 22$. Lembre-se que, no cálculo desta integral com a regra do trapézio, precisamos tomar $n = 1155$ para garantir que teríamos a mesma precisão. Aplicando a fórmula (143),

$$\int_0^2 \sin(t) dt \approx \frac{h}{3} \left(\sin(2) + 2 \sum_{i=1}^{10} \sin(x_{2i}) + 4 \sum_{i=0}^{10} \sin(x_{2i+1}) \right) \approx 1.416147374435926.$$

Comparando este valor com o valor desta integral calculado a partir de sua primitiva, verificamos que o erro cometido foi, de fato

$$|1.416147374435926 - 1.416146836547142| = 5.37888783957996 \cdot 10^{-7}$$

Passando, agora, ao cálculo de

$$\int_0^1 \exp(-t^2) dt$$

com erro menor que 10^{-6} , começamos calculando a quarta derivada de $f(x) = \exp(-x^2)$, que é igual a

$$f^{(iv)}(x) = (16x^4 - 48x^2 + 12) \exp(-x^2).$$

Para achar o máximo desta função, precisamos encontrar seus pontos críticos, que são os zeros de

$$f^{(v)}(x) = (-32x^5 + 160x^3 - 120x) e^{-x^2}.$$

Usando o Maxima, verificamos que as raízes reais de $-32x^5 + 160x^3 - 120x$ são

$$-2.0202, -0.9586, 0.009586, \text{ e } 2.0202;$$

das quais 0 e 0.9586 estão no intervalo de integração. Como

$$f^{(iv)}(0) = 0, \quad f^{(iv)}(0.9586) = 0.0011 \text{ e } f^{(iv)}(2) = 0.2931$$

podemos concluir que

$$M = \max\{f^{(v)}(x) \mid x \in [0, 2]\} = 0.2931.$$

Logo, pela fórmula (144), o erro no cálculo desta integral pela regra de Simpson será menor que 10^{-6} se

$$\frac{0.2931 \cdot 1 \cdot h^4}{180} = \frac{0.2931}{180n^4} \leq 10^{-6};$$

isto é, quando $n > 6.3524$. Como n tem que ser um inteiro par, o menor valor que podemos escolher é $n = 8$. Finalmente, aplicando (143) com $n = 8$ e $h = 1/8 = 0.125$, obtemos

$$\int_0^1 \exp(-t^2) dt \approx 0.74682425743573.$$

CAPÍTULO 9

O problema de valor inicial

Neste capítulo voltamos a tratar dos problemas de valores iniciais, introduzidos no capítulo 7, onde vimos como resolvê-los pelo método de Euler. Nossa meta é estudar três algoritmos que pertencem a uma família de algoritmos para resolver estes problemas, conhecidos conjuntamente como *métodos de Runge-Kutta*. Como veremos neste capítulo, estes métodos estão relacionados, de perto, às regras de integração apresentadas no capítulo 8.

1. O método de Euler Modificado

Começaremos reinterpretando o método de Euler de uma maneira adequada a facilitar sua generalização. Digamos que queremos resolver o problema de valor inicial definido pela equação diferencial

$$\dot{y} = f(t, y) \quad \text{e} \quad y(a) = \alpha,$$

em que $f : [a, b] \rightarrow \mathbb{R}$ é uma função contínua. Começamos dividindo o intervalo $[a, b]$ em n partes iguais e tomando $h = (b - a)/n$. Como usual, denotaremos por t_i o ponto $a + ih$. De acordo com o *teorema fundamental do cálculo*,

$$y(t_{i+1}) - y(t_i) = \int_{t_i}^{t_{i+1}} \frac{dy(s)}{ds} ds.$$

A equação diferencial $\dot{y} = f(t, y)$ nos permite reescrever isto na forma

$$(145) \quad y(t_{i+1}) - y(t_i) = \int_{t_i}^{t_{i+1}} f(y(s), s) ds.$$

Aproximando esta integral pelo retângulo de vértices $(t_i, 0)$, $(t_{i+1}, 0)$, (t_i, y_i) e (t_{i+1}, y_{i+1}) , obtemos

$$y(t_{i+1}) - y(t_i) \approx hf(y(t_i), t_i).$$

Substituindo $y(t_i)$ por sua aproximação y_i e \approx pela igualdade,

$$y_{i+1} - y_i = hf(y_i, t_i),$$

que é a iteração do método de Euler. Em outras palavras, o método de Euler consiste apenas da aplicação da regra do retângulo para aproximar a integral (145), obtida

integrando a derivada da solução da equação $\dot{y} = f(t, y)$. Isto sugere a pergunta: o que acontece se aplicarmos outra regra de integração à mesma integral?

Começaremos nossa investigação pela regra do trapézio. Aplicando-a à equação (145), obtemos

$$y(t_{i+1}) - y(t_i) \approx \frac{h}{2}(f(y(t_i), t_i) + f(y(t_{i+1}), t_{i+1})).$$

Procedendo como no caso da regra do retângulo, substituiremos $y(t_i)$ por sua aproximação y_i e transformemos \approx em uma igualdade, obtendo, assim, a iteração

$$(146) \quad y_{i+1} - y_i = \frac{h}{2}(f(y_i, t_i) + f(y_{i+1}, t_{i+1})),$$

que é apenas um pouco mais complicada do que a correspondente ao método de Euler. Vejamos como aplicá-la ao problema de valor inicial

$$\dot{y} = y \quad \text{e} \quad y(0) = 1.$$

Como, $f(t, y) = y$, a equação (146) para este problema específico será

$$y_{i+1} - y_i = \frac{h}{2}(y_i + y_{i+1}).$$

Note que, ao contrário do que ocorre no método de Euler, y_{i+1} aparece dos dois lados da equação, o que nos obriga a rearrumar a equação para podermos escrevê-lo em função de y_i . Fazendo isto,

$$y_{i+1} = y_i \cdot \left(\frac{2+h}{2-h} \right).$$

Para obter $y(1)$, tomamos $h = 1/n$, de modo que a iteração se torna

$$y_{i+1} = y_i \cdot \left(\frac{2n+1}{2n-1} \right).$$

Portanto,

$$y_n = y_{n-1} \cdot \left(\frac{2n+1}{2n-1} \right) = y_{n-2} \cdot \left(\frac{2n+1}{2n-1} \right)^2 = y_{n-3} \cdot \left(\frac{2n+1}{2n-1} \right)^3.$$

continuando desta maneira, obtemos

$$y_n = y_0 \cdot \left(\frac{2n+1}{2n-1} \right)^n = \left(\frac{2n+1}{2n-1} \right)^n,$$

pois estamos supondo que $y_0 = 1$. Contudo,

$$(147) \quad \lim_{n \rightarrow \infty} y_n = \lim_{n \rightarrow \infty} \left(\frac{2n+1}{2n-1} \right)^n = e,$$

como esperávamos. Detalhes de como calcular este limite podem ser encontrados no exercício 1.

Infelizmente não é tão fácil aplicar este método quanto este primeiro exemplo faz crer. Para entender porque, considere o problema de valor inicial

$$\dot{y} = \sin(y) \quad \text{e} \quad y(0) = 1.57.$$

Desta vez a equação (146) nos dá

$$y_{i+1} - y_i = \frac{h}{2}(\sin(y_i) + \sin(y_{i+1})).$$

Passando as expressões em y_{i+1} para um lado e as expressões em y_i para o outro,

$$(148) \quad y_{i+1} - \frac{h}{2} \sin(y_{i+1}) = y_i + \frac{h}{2} \sin(y_i).$$

Mas como escrever y_{i+1} em função de y_i a partir desta expressão? A resposta é que, infelizmente, isto não é possível, o que nos deixa uma única saída: achar os zeros de (148), como uma função de y_{i+1} . Com um pouco de sorte, o zero é único no intervalo em que estamos resolvendo a equação diferencial, o que nos permite achar a solução sem ambiguidade. Por exemplo, podemos usar o método de Newton-Raphson para resolver (148) relativamente à variável y_{i+1} , uma vez que y_i tenha sido calculado. Fazendo isto a partir de $y_0 = 1.57$, por 100 etapas consecutivas com $h = 0.1$, obtemos o gráfico da figura 1.

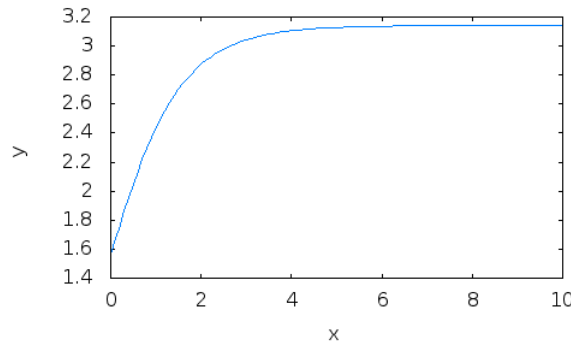


FIGURA 1. Curva solução de $\dot{y} = \sin(y)$ com $y(0) = 1.57$.

Este é um exemplo de um método *implícito*, assim chamado porque é necessário resolver uma equação para achar o valor de y_{i+1} ; ao contrário do que acontece nos métodos *explícitos*, em que y_{i+1} é diretamente escrita como função de y_i , t_i e t_{i+1} . O método de Euler é um exemplo de método explícito. Os métodos implícitos têm o inconveniente de que o custo em aplicá-los depende muito da equação que y_{i+1} satisfaz. Se o custo de achar o zero da equação for alto, o método implícito pode se tornar muito lento; a saída é transformá-lo em explícito. Para isso, procederemos em duas etapas: na primeira, calcularemos uma aproximação para y_{i+1} usando o

método de Euler; na segunda, melhoraremos esta aproximação usando a equação (146). Aplicando isto ao problema de valor inicial

$$\dot{y} = f(t, y) \quad \text{e} \quad y(a) = \alpha,$$

com que começamos a seção, obtemos na primeira etapa

$$y_{i+1} = y_i + hf(t_i, y_i)$$

e na segunda

$$(149) \quad y_{i+1} = y_i + \frac{h}{2}(f(y_i, t_i) + f(y_i + hf(t_i, y_i), t_{i+1})),$$

em que apenas substituímos a equação acima no lado direito de (146). Esta fórmula define o *método de Euler melhorado*, que pertence à família dos métodos de Runge-Kutta.

Vejamos o que acontece quando aplicamos o método de Euler melhorado ao problema de valor inicial

$$\dot{y} = \sin(y) \quad \text{e} \quad y(0) = 1.57.$$

A iteração resultante da aplicação do método de Euler melhorado a este problema é

$$y_{i+1} = y_i + \frac{h}{2}(\sin(y_i) + \sin(y_i + h \sin(y_i))) \quad \text{e} \quad y_0 = 1.57.$$

O resultado das oito primeiras iterações com $h = 0.1$ está resumido na tabela 1.

i	1	2	3	4	5	6	7	8
t_i	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
y_i	1.6697	1.7685	1.8653	1.9594	2.05	2.1365	2.2185	2.2957

TABELA 1. Método de Euler melhorado para $\dot{y} = \sin(y)$ com $y(0) = 1.57$.

Encerraremos a seção com uma análise rápida do método de Runge-Kutta de quarta ordem. A iteração deste método, quando aplicada ao problema de valor inicial

$$\dot{y} = f(t, y) \quad \text{e} \quad y(a) = \alpha,$$

é dada pela fórmula

$$(150) \quad y_{i+1} - y_i = \frac{h}{6}(K_1 + 2K_2 + 2K_3 + K_4),$$

em que

$$\begin{aligned} K_1 &= f(x_k, y_k); & K_2 &= f(x_k + h/2, y_k + hK_1/2); \\ K_3 &= f(x_k + h/2, y_k + hK_2/2); & K_4 &= f(x_k + h, y_k + hK_3). \end{aligned}$$

Da mesma forma que o método de Euler melhorado é uma generalização da regra do trapézio, o método de Runge-Kutta de quarta ordem é uma generalização da regra de Simpson. Provaremos isto, aplicando este método para calcular a integral

$$\int_a^b \phi(s) ds.$$

Para isto subdividimos $[a, b]$ em n intervalos iguais de comprimento $h = (b - a)/n$, que escreveremos na forma $[t_i, t_{i+1}]$, em que $t_i = a + ni$. Com isto, basta aplicar o método de Runge-Kutta de quarta ordem às integrais,

$$\int_{t_i}^{t_{i+1}} \phi(s) ds, \quad \text{com} \quad i = 0, \dots, n-1.$$

Para converter o cálculo desta última integral em um problema de valor inicial, definimos

$$y(t) = \int_{t_i}^t \phi(s) ds,$$

de modo que, pelo teorema fundamental do cálculo,

$$(151) \quad y(t_i) - y(t_{i+1}) = \int_{t_i}^{t_{i+1}} \phi(s) ds.$$

Mas, isto equivale a dizer que

$$\dot{y} = \phi(t) \quad \text{com} \quad y(t_i) = y_i,$$

que é o problema de valor inicial que vamos resolver. Aplicando o método de Runge-Kutta de quarta ordem a este problema, verificamos que

$$\begin{aligned} K_1 &= \phi(t_k); & K_2 &= \phi(t_k + h/2); \\ K_3 &= \phi(t_k + h/2); & K_4 &= \phi(t_k + h). \end{aligned}$$

Substituindo isto em (150) e observando que

$$K_2 = K_3 = \phi(t_k + h/2),$$

obtemos

$$y_{i+1} - y_i = \frac{h}{6}(\phi(t_k) + 4\phi(t_k + h/2) + \phi(t_k + h)).$$

Contudo, por (151), esta última equação equivale à fórmula (143) (p. 178) aplicada a subintervalos de largura $h/2$, confirmando nossa afirmação original.

Como fizemos para o método de Euler melhorado, aplicaremos o método de Runge-Kutta de quarta ordem ao problema de valor inicial

$$\dot{y} = \sin(y) \quad \text{e} \quad y(0) = 1.57,$$

que nos dá a iteração (150), com

$$\begin{aligned} K_1 &= \sin(y_k); & K_2 &= \sin(y_k + hK_1/2); \\ K_3 &= \sin(y_k + hK_2/2); & K_4 &= \sin(y_k + hK_3). \end{aligned}$$

A tabela 2 lista os valores de y_i e dos quatro K s para as oito primeiras iterações.

i	t_i	K_1	K_2	K_3	K_4	y_i
1	0.1	0.999	0.998	0.998	0.995	1.6365
2	0.2	0.997	0.993	0.993	0.986	1.7027
3	0.3	0.991	0.983	0.983	0.973	1.7682
4	0.4	0.98	0.969	0.969	0.956	1.8328
5	0.5	0.965	0.952	0.952	0.936	1.8963
6	0.6	0.947	0.931	0.931	0.913	1.9584
7	0.7	0.925	0.907	0.907	0.887	2.0188
8	0.8	0.901	0.88	0.881	0.859	2.0775

TABELA 2. Runge-Kutta de quarta ordem para $\dot{y} = \sin(y)$ com $y(0) = 1.57$.

Para tornar mais fácil comparar a rapidez com que os três métodos explícitos que estudamos convergem, desenhamos na figura 2 as aproximações da curva solução deste problema de valor inicial calculadas usando o método de Euler (vermelho), o método de Euler melhorado (verde) e o método de Runge-Kutta de quarta (azul), juntamente com a solução analítica (preto). A figura 3 mostra o zoom do mesmo gráfico no intervalo $[1, 1.5]$.

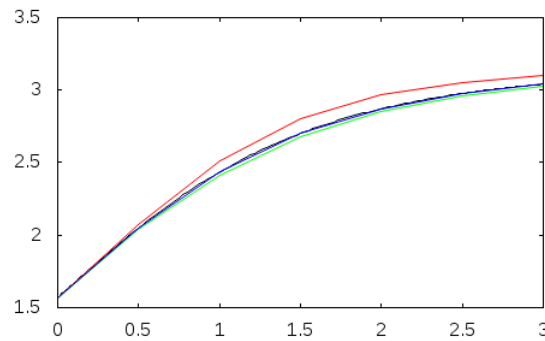


FIGURA 2. Os métodos explícitos comparados.

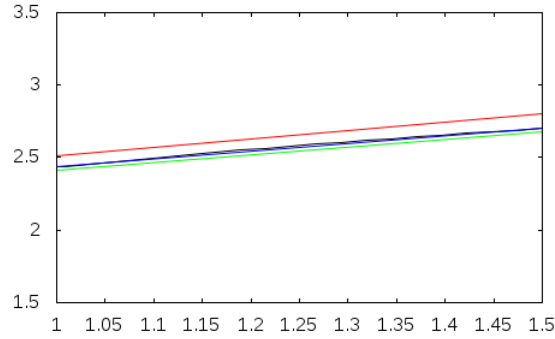


FIGURA 3. Zoom do trecho entre 1 e 1.5.

Analisando criticamente o que fizemos nesta seção, verificamos que a maneira como tratamos os métodos numéricos para solução do problema de valor inicial é bastante imprevidente. Afinal, nada nos impede de inventar expressões iterativas mirabolantes usando y_i , t_i , t_{i+1} e $f(t, y)$ como ingredientes. Porém uma iteração, seja lá como foi inventada, precisa convergir para a solução do problema de valor inicial, de preferência rapidamente, caso contrário não terá utilidade. Portanto, antes de usar qualquer um destes métodos numéricos, deveríamos nos certificar de que sempre converge para a solução exata. Esta deficiência será suprida na seção 2, onde provaremos a convergência e veremos como analisar a eficiência dos vários métodos que estamos estudando, explicando, pelo caminho, o sentido em que a palavra *ordem* é usada para qualificar o método de Runge-Kutta.

2. Convergência e ordem

Considere o *problema de valor inicial*

$$(152) \quad \dot{y} = f(t, y) \quad \text{e} \quad y(t_0) = y_0,$$

em que $f(t, y)$ é uma função analítica. Em outras palavras, f admite uma expansão em série de potências. A vantagem desta hipótese é que ela implica que a solução $y(t)$, do problema de valor inicial, também é analítica. Em particular, *tanto f , quanto $y(t)$ têm derivadas contínuas de todas as ordens*. A restrição a funções analíticas não impõe uma restrição muito forte; por exemplo, quase todas as funções estudadas nos cursos de cálculo são analíticas, como já tivemos ocasião de observar na página 50.

Nesta seção analisaremos os três métodos explícitos para solução do problema de valor inicial que estudamos na seção 1: o método de Euler, o método de Euler melhorado e o método de Runge-Kutta de quarta ordem (RK-4). A iteração para todos estes três métodos pode ser escrita na forma

$$(153) \quad y_{k+1} = y_k + h\Phi(t_k, y_k, h),$$

com $\Phi(t_k, y_k, h)$ tomando as seguintes formas:

método de Euler: $\Phi(t_k, y_k, h) = f(t_k, y_k)$;

método de Euler melhorado: $\Phi(t_k, y_k, h) = (f(t_k, y_k) + f(t_k + h, y_k + hf(t_k, y_k)))/2$;

método RK4: $\Phi(t_k, y_k, h) = (K_1 + 2K_2 + 2K_3 + K_4)/6$,

em que

$$\begin{aligned} K_1 &= f(t_k, y_k); & K_2 &= f(t_k + h/2, y_k + hK_1/2); \\ K_3 &= f(t_k + h/2, y_k + hK_2/2); & K_4 &= f(t_k + h, y_k + hK_3). \end{aligned}$$

A análise que faremos destes métodos requer a introdução de dois tipos de erros. Digamos que $y(t)$ seja uma solução exata do problema (152). O *erro global* cometido quando tentamos resolver este problema usando a iteração (153) é

$$(154) \quad e_k = y(t_k) - y_k;$$

já o *erro de truncamento* correspondente à mesma iteração é

$$(155) \quad T_k = \frac{y(t_{k+1}) - y(t_k)}{h} - \Phi(t_k, y(t_k), h).$$

Observe que o erro global diz respeito à qualidade da aproximação da solução correta relativamente à aproximação numérica, enquanto o erro de truncamento está relacionado à qualidade da aproximação do quociente de Newton

$$\frac{y(t_{k+1}) - y(t_k)}{h} = \frac{y(x_n + h) - y(t_k)}{h}$$

pela função Φ . Para entender porque é natural comparar estes dois últimos valores, basta reescrever (153) na forma

$$\Phi(t_k, y_k, h) = \frac{y_{k+1} - y_k}{h}.$$

Para simplificar a exposição, suporemos, de agora em diante, que a equação diferencial que estamos considerando é *autônoma*; isto é, que f é independente de t . Sob esta hipótese, as funções Φ dos três métodos que estamos analisando têm a forma

método de Euler: $\Phi(t_k, y_k, h) = f(y_k)$;

método RK-2: $\Phi(t_k, y_k, h) = (f(y_k) + f(y_k + hf(y_k)))/2$;

método RK4: $\Phi(t_k, y_k, h) = (K_1 + 2K_2 + 2K_3 + K_4)/6$,

em que

$$\begin{aligned} K_1 &= f(y_k); & K_2 &= f(y_k + hK_1/2); \\ K_3 &= f(y_k + hK_2/2); & K_4 &= f(y_k + hK_3). \end{aligned}$$

Em particular, $\Phi(t_k, y_k, h)$ independe da variável t_k em todos os três casos. Portanto, podemos supor, de agora em diante, que Φ independe de t_k e escreveremos $\Phi = \Phi(y_k, h)$.

Como seria de esperar, o erro global e o erro de truncamento estão relacionados entre si. Reescrevendo (155) na forma

$$y(t_{k+1}) - y(t_k) = h\Phi(y(t_k), h) + T_k h$$

e subtraindo

$$y_{k+1} - y_k = h\Phi(t_k, y_k, h),$$

disto, obtemos,

$$(156) \quad e_{k+1} - e_k = h(\Phi(y(t_k), h) - \Phi(y_k, h)) + T_k h.$$

Seja I um intervalo fechado que contém os pontos $y(t_0), \dots, y(t_n)$, assim como, y_0, \dots, y_n . Supondo que $\Phi(y, h)$ tem primeira derivada contínua como função de y , e levando em conta que I é fechado e limitado, podemos concluir que existe uma constante positiva C tal que

$$\left| \frac{\partial \Phi}{\partial y}(y, h) \right| < C$$

para todo $y \in I$. Contudo, a fórmula de Taylor nos permite escrever

$$(157) \quad |\Phi(y(t_k), h) - \Phi(y_k, h)| \leq C|y(t_k) - y_k| = C e_k,$$

pois

$$\max \left\{ \left| \frac{\partial \Phi}{\partial y}(y, h) \right| \mid (y, h) \in I \right\} \leq C.$$

Mas, segundo (156),

$$e_{k+1} = e_k + h(\Phi(y(t_k), h) - \Phi(y_k, h)) + T_k h.$$

Usando a desigualdade triangular e (157), obtemos

$$(158) \quad |e_{k+1}| < |e_k|(1 + hC) + |T_k|h \leq |e_k|(1 + hC) + Th,$$

em que

$$T = \max\{|T_k| \mid 0 \leq k \leq n\},$$

Definindo

$$(159) \quad \theta_{k+1} = (1 + Ch)\theta_k + Th \quad \text{e} \quad \theta_0 = e_0 = y(x_0) - y_0 = 0,$$

temos de (158) que

$$|e_{k+1}| \leq \theta_k$$

para todo $k \geq 0$. Mas a recorrência (159) tem como solução

$$\theta_n = \frac{(Ch + 1)^n T}{Ch} - \frac{T}{Ch};$$

para mais detalhes veja o exercício 2. Como $T \geq 0$ e $C \geq 0$, concluímos que

$$|e_{k+1}| \leq \theta_k < \frac{(Ch + 1)^k T}{C} \leq \frac{(Ch + 1)^n T}{C}.$$

Finalmente, levando em conta que

$$1 + Ch \leq 1 + Ch + \frac{(Ch)^2}{2!} + \frac{(Ch)^3}{3!} + \cdots = \exp(Ch),$$

e que

$$nh = t_n - t_0$$

obtemos a desigualdade

$$(160) \quad |e_n| < \exp(C(t_n - t_0)) \frac{T}{C},$$

que relaciona o erro global e_n ao máximo T dos erros de truncamento.

⚠ Como cota superior para o erro, (160) é excepcionalmente ruim, porque a exponencial faz com que o lado direito seja muito grande. Em particular, *não é adequado usá-la para estimar em quantas partes o intervalo de integração deve ser dividido para que o erro fique abaixo de uma dada tolerância*. Apesar disto esta desigualdade nos ajuda a comparar a eficiência dos métodos que estamos estudando, além de nos permitir mostrar que convergem.

A velocidade com que um método converge depende de sua *ordem*, que é definida como o maior inteiro positivo p para o qual o módulo do erro de truncamento é menor ou igual a λh^p , para alguma constante positiva λ . Para entender de que forma a eficiência de um método está relacionada à sua ordem, note que se o método tem ordem p , então

$$T \leq \lambda h^p.$$

Substituindo isto em (160),

$$|e_n| < \exp(C(t_n - t_0)) \frac{\lambda h^p}{C},$$

donde

$$\lim_{n \rightarrow \infty} |e_n| < \lim_{h \rightarrow 0} \left(\exp(C(x_n - x_0)) \frac{\lambda h^p}{C} \right) = \exp(C(x_n - x_0)) \lim_{h \rightarrow 0} \frac{\lambda h^p}{C} = 0,$$

sempre que $|h| < 1$. Naturalmente este limite tenderá a zero tão mais rápido, quanto maior for p .

Resta-nos determinar a ordem de cada um dos três métodos que estudamos. Para ser honesto esta é, do ponto de vista dos cálculos que precisam ser feitos, a parte mais difícil do estudo do erro. Não que os cálculos sejam conceitualmente difíceis; o problema é que são longos e complicados. Vejamos o que acontece quando analisamos

o método de Euler. Como $\Phi(y, h) = f(y)$ para este método, teremos que o erro de truncamento, dado pela fórmula (155), será

$$(161) \quad T_k = \frac{y(t_{k+1}) - y(t_k)}{h} - f(y(t_k)).$$

Como $y(t)$ é analítica, a fórmula de Taylor nos dá,

$$y(t_{k+1}) = y(t_k) + \dot{y}(t_k)(t_{k+1} - t_k) + \frac{M_k}{2}(t_{k+1} - t_k)^2,$$

em que M_k é uma constante que satisfaz

$$|M_k| \leq \max\{\ddot{y}(t) \mid t \in [a, b]\}.$$

Substituindo $t_{k+1} - t_k$ por h ,

$$(162) \quad y(t_{k+1}) - y(t_k) = \dot{y}(t_k)h + \frac{M_2}{2}h^2 = h \left(\dot{y}(t_k) + \frac{M_2}{2}h \right),$$

donde

$$T_k = \dot{y}(t_k) + \frac{M_2}{2}h - f(y(t_k)).$$

Mas, a equação diferencial $\dot{y} = f(y)$ nos dá $\dot{y}(t_k) = f(y(t_k))$, donde

$$T_k = \frac{M_2}{2}h,$$

o que mostra que o método de Euler é de primeira ordem.

Vejamos o que acontece quando o mesmo argumento é aplicado ao método de Runge-Kutta de segunda ordem, para o qual

$$\Phi(y, h) = \frac{1}{2}(f(y) + f(y + hf(y))).$$

Desta vez, para podermos explicitar $\Phi(y(t_k), h)$ como função de h , precisamos usar a fórmula de Taylor de f , segundo a qual

$$f(y(t_k) + \dot{y}(t_k)h) = f(y(t_k)) + f'(y(t_k))\dot{y}(t_k)h + \frac{E_1}{2}(\dot{y}(t_k)h)^2,$$

em que E_k é uma constante que satisfaz

$$|E_k| \leq \max\{f''(t) \mid t \in [y(t_k), y(t_k) + \dot{y}(t_k)h]\}.$$

Note que substituímos $f(y(t_k))$ por $\dot{y}(t)$, na fórmula acima. Segue disto que

$$\Phi(y_k, h) = f(y(t_k)) + \frac{f'(y(t_k))\dot{y}(t_k)}{2}h + \frac{E_k}{4}(\dot{y}(t_k)h)^2,$$

Por outro lado, pela fórmula de Taylor de ordem dois para y ,

$$y(t_{k+1}) - y(t_k) = \dot{y}(t_k)h + \frac{\ddot{y}(y(t_k))}{2}h^2 + \frac{C_k}{6}h^3,$$

para uma constante C_k que satisfaz

$$|C_k| \leq \max\{\ddot{y}(t) \mid t \in [t_k, t_{k+1}]\}.$$

Substituindo estas duas últimas expressões em (161), agrupando as potências comuns de h e levando em conta que $\dot{y}(t_k) = f(y(t_k))$, obtemos

$$T_k = \frac{h}{2} (\ddot{y}(t_k) - f'(y(t_k))\dot{y}(t_k)) + \left(\frac{C_k}{6} - \frac{E_k}{2!} (\dot{y}(t_k))^2 \right) h^2$$

Contudo, da equação diferencial (152) e da regra da cadeia, temos também que

$$(163) \quad \ddot{y}(t) = \frac{df(y(t))}{dt} = f'(y(t))\dot{y}(t) = f'(y(t))f(y(t)),$$

em que f' denota a derivada de $f(y)$ relativamente a y . Levando isto em conta, a expressão para T_k acima torna-se

$$T_k = \left(\frac{C_k}{6} - \frac{E_k}{2!} (\dot{y}(t_k))^2 \right) h^2,$$

o que mostra que o método de Runge-Kutta de segunda ordem tem mesmo a ordem que lhe é atribuída.

Para encerrar, trataremos do método de Runge-Kutta de quarta ordem. Desta vez precisamos aplicar a fórmula de Taylor a $y(t)$ até quarta ordem e a $f(y)$ até terceira ordem. Como as expressões são longas e complicadas, vamos deixar os cálculos por conta do sistema de computação algébrica AXIOM. Este sistema é de domínio público e você pode baixá-lo a partir da página

<http://axiom-developer.org/axiom-website/download.html>

e instalá-lo em seu computador. Para ajudá-lo a repetir os cálculos, se desejar, apresentamos em cada etapa os comandos do AXIOM necessários para executá-los.

Começamos informando ao AXIOM que f e y devem ser considerados como funções

```
f:= operator 'f
y:= operator 'y
```

Em seguida preciso ensinar algumas o AXIOM a relacionar a primeira, segunda e terceira derivadas de y à função f e suas derivadas. A equação diferencial nos dá $\dot{y}(t) = f(y(t))$, e já vimos como relacionar \ddot{y} à primeira derivada de f em (163). De maneira semelhante,

$$(164) \quad \ddot{y}(t) = f''(y(t))f(y(t))^2 + f'(y(t))^2 f(y(t)).$$

Para ensinar o sistema a reconhecer estas igualdades, criaremos quatro regras

```
ypara f1:= rule
  y(tk)==yk
```

```

yparaf2:= rule
  D(y(tk),tk)==f(yk)

yparaf3:= rule
  D(y(tk),tk,2)==D(f(yk),yk)*D(y(tk),tk)

yparaf4:= rule
  D(y(tk),tk,3)==D(f(yk),yk,2)*D(y(tk),tk)^2+D(f(yk),yk)*D(y(tk),tk,2)

```

a primeira das quais é usada para substituir a função $y(t_k)$ pela variável y_k . Em seguida, reunimos todas estas regras sob uma única função, tomando cuidado para que sejam aplicadas da última derivada para a primeira, para que $\ddot{y}(t)$ seja substituída em $\ddot{y}(t)$ pela expressão correta em termos das derivadas de f , e assim por diante.

```

simplifica(exp) ==
  exp:= yparaf4(exp)
  exp:= yparaf3(exp)
  exp:= yparaf2(exp)
  exp:= yparaf1(exp)
  return(exp)

```

As simplificações correspondentes a estas regras serão efetuadas apenas ao final dos cálculos.

Com isso estamos prontos para calcular T_k . Para isto precisaremos das fórmulas de Taylor de $y(t_{k+1}) = y(t_k + h)$ e de $\Phi(y(t_k), h)$. Para mostrar que o método de Runge-Kutta de quarta ordem tem mesmo esta ordem basta mostrar que os coeficientes de h^i na expressão de T_k são nulos para $i = 0, \dots, 3$. Isto requer que calculemos as fórmulas de Taylor de $y(t_k + h)$ e $\Phi(y(t_k), h)$ até ordem 4 e 3, respectivamente. A razão pela qual precisamos de $y(t_k + h)$ até ordem 4 é que $y(t_k + h) - y(t_k)$ aparece dividido por h na definição de T_k . Calcularemos cada parcela da fórmula de Taylor de $y(t_k + h)$ separadamente, reunindo-as todas ao final:

```

B0:= y(tk)
B1:= f(y(tk))
B2:= D(B1,tk)
B3:= D(B2,tk)
B4:= D(B3,tk)

STy:= y(tk)+ h*D(y(tk),tk) + ((h^2)/2)*B2+ (h^3/6)*B3+ (h^4/factorial(4))*B4

```

Para achar uma expressão para $\Phi(y(t_k), h)$ em função de h , calculamos, separadamente, as fórmulas de Taylor de K_1 , K_2 , K_3 e K_4 até ordem três. Por exemplo, para calcular a fórmula de Taylor de $K_2 = f(t_k + (h/2)K_1)$, expandimos $f(t_k + (h/2)K_1 \cdot s)$ como função de s e tomamos $s = 0$ ao final. Como há três K 's diferentes cujas

fórmulas precisamos encontrar, criaremos uma função para executar os cálculos necessários.

```
taylorNextK(j,K) ==
  if member?(j,[2,3]) then
    A:= f(yk+(K*(h/2))*t)
    lK:= [D(A,t,i)/factorial(i) for i in 0..3]
    lK:=eval(lK,t=0)
    FTK:= reduce(+,lK)
  if j=4 then
    A:= f(yk+(K*h)*t)
    lK:= [D(A,t,i)/factorial(i) for i in 0..3]
    lK:=eval(lK,t=0)
    FTK:= reduce(+,lK)
  return(FTK)
```

Note que as fórmulas foram calculadas em duas etapas: primeiro geramos uma lista lK cujas entradas são as parcelas da fórmula de Taylor; em seguida as parcelas foram somadas usando $\text{FTK} := \text{reduce}(+, \text{lK})$. Observe, também, que precisamos tratar separadamente os cálculos de K_4 e de K_2 e K_3 , porque nos dois últimos casos o K anterior aparece multiplicado por $h/2$, e não por h .

Com isto estamos prontos para calcular as fórmulas de Taylor para os diversos K s e somá-las para obter $\Phi(y_k, h)$:

```
K1:= f(yk)
K2:= taylorNextK(2,K1)
K3:= taylorNextK(3,K2)
K4:= taylorNextK(4,K3)
```

```
Phi:= (1/6)*(K1+2*K2+2*K3+K4);
```

Como a expressão para Φ é bastante grande, adicionamos um ponto-e-vírgula ao final da linha em que é definida, para que o sistema não escreva a equação no terminal quando acabar de calculá-la. Em seguida, calculamos T_k e determinamos os coeficientes de h^0 , h^1 , h^2 e h^3 em sua fórmula de Taylor, que denotaremos por **par0**, **par1**, **par2** e **par3**, respectivamente.

```
Tk:= ((STy-yk)/h)-Phi; Tk0:= eval(Tk,yk=y(tk))
par0:=eval(Tk1,h=0)
Tk1:= (Tk1-par0)/h
par1:= eval(Tk2,h=0)
Tk2:= (Tk2-par1)/h
par2:= eval(Tk3,h=0)
```

```
Tk3:= (Tk3-par2)/h
par3:= eval(Tk4,h=0)
```

Estes coeficientes *não* são nulos, porque ainda precisamos fazer as simplificações resultantes de $\dot{y}(t) = f(y(t))$ e das equações (163) e (164). Para isso basta usar a função `simplifica` definida acima. Como

```
simplifica(par0)
simplifica(par1)
simplifica(par2)
simplifica(par3)
```

retorna zero para todos os coeficientes de h^i , com $i = 0, \dots, 3$, podemos concluir que o coeficiente não nulo da menor potência de h na expressão de T_k é h^4 , o que mostra que o método de Runge-Kutta de quarta ordem tem mesmo a ordem que lhe é atribuída, ao menos no caso em que a equação diferencial é autônoma.

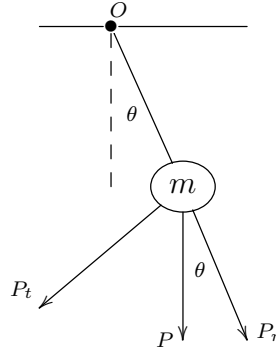


Note que nossa análise do erro para todos os três métodos pressupõe que o problema de valor inicial seja referente a uma equação diferencial *autônoma* e de *dimensão um*. Para tratar equações que não sejam autônomas ou que tenham dimensão maior que um, precisaríamos utilizar a fórmula de Taylor em mais de uma variável. Embora isto possa ser feito de maneira razoavelmente simples, não traria nenhuma contribuição expressiva para seu entendimento de porquê estes métodos têm as ordens que lhe são atribuídas; por isso optamos por uma enfoque mais simples. Um tratamento teórico e geral da determinação da ordem dos métodos que estudamos pode ser encontrado no livro [1] relacionado na bibliografia abaixo. Por sua vez, nossa apresentação é baseada no capítulo 12 do livro [12].

3. O pêndulo revisitado

Encerramos revisitando o pêndulo. Começaremos com o caso em que não há, nem atrito, nem resistência do ar, já analisado na seção 1 do capítulo 6. Mas, ao invés do princípio de conservação da energia, usaremos o diagrama usual de equilíbrio de forças para obter a equação diferencial. Considere a figura abaixo, na qual um pêndulo de massa m oscila em torno de um ponto fixo O . O vetor P representa o peso da bola, ao passo que P_r e P_t correspondem aos componentes de P nas direções

radial e tangente à trajetória da bola.



Admitindo que o sistema de eixos foi posicionado da maneira usual, temos que $P = -mg$ e que

$$(165) \quad P_t = P \sin(\theta) = -mg \sin(\theta).$$

Por outro lado, se s é o comprimento de arco, medido a partir da vertical, e ℓ é o comprimento da haste, temos que $s = \ell\theta$. Como a haste tem comprimento fixo, a aceleração ao longo da tangente à trajetória será igual a

$$\ddot{s} = \ell\ddot{\theta},$$

em que os pontos são usados para denotar a derivada relativamente ao tempo. Segue disto que $P_t = -m\ell\ddot{\theta}$. Igualando isto à expressão de P_t em (165) e dividindo tudo por $m\ell$, obtemos

$$(166) \quad \ddot{\theta} = -\frac{g}{\ell} \sin(\theta),$$

que é a equação do pêndulo simples. Como esta é uma equação de segunda ordem, precisamos de duas condições inicial: uma para a posição e outra para a velocidade do pêndulo no instante inicial. Supondo, para simplificar, que o pêndulo foi solto, a partir do repouso, quando formava um ângulo θ_0 com a vertical e que começamos a contar o tempo a partir deste momento, obtemos

$$(167) \quad \ddot{\theta} = -\frac{g}{\ell} \sin(\theta), \quad \theta(0) = \theta_0 \quad \text{e} \quad \dot{\theta}(0) = 0.$$

Como já observado acima, (166) é uma equação diferencial de segunda ordem, ao passo que a que havíamos obtido no capítulo 6 era de primeira ordem. Isto explica porque escolhemos o enfoque via conservação da energia, em vez do que adotamos nesta seção, que provavelmente lhe parece mais natural. Mas isto também põe dois problemas: de que forma as duas equações estão relacionadas? como resolver equações de segunda ordem usando os métodos estudados neste capítulo?

Começaremos respondendo a primeira pergunta; a segunda será o tema da próxima seção. Vejamos o que acontece se multiplicarmos os dois lados da equação (166)

por $\dot{\theta}$:

$$(168) \quad \dot{\theta}\ddot{\theta} = -\frac{g}{\ell} \sin(\theta)\dot{\theta}.$$

Como $\ddot{\theta}$ é a derivada de $\dot{\theta}$, o lado direito de (168) é igual à derivada de $\dot{\theta}^2/2$. Mas, pela regra da cadeia, $\sin(\theta)\dot{\theta}$ é igual à derivada de $\cos(\theta)$. Portanto, integrando os dois lados de (168) em relação a t obtemos

$$\frac{\dot{\theta}^2}{2} = \frac{g}{\ell} \cos(\theta) + c,$$

em que c é a constante de integração. Escolhendo as unidades como na página 114 e extraindo a raiz quadrada a equação toma a forma

$$\dot{\theta} = \sqrt{c + 2 \cos(\theta)},$$

que coincide com (80), a menos da escolha de c . Isto ajuda a explicar um fenômeno com que nos deparamos em nossa análise do pêndulo no capítulo 6, onde verificamos que (80) admite duas soluções quando o pêndulo atinge sua altura máxima. Acontece que, nestes pontos, a velocidade do pêndulo se anula. Mas isto significa que $\dot{\theta} = 0$ nestes pontos. Contudo, para passar de uma equação à outra, multiplicamos os dois lados da equação de segunda ordem $\ddot{\theta}$, com a consequência de que a equação de primeira ordem terá que se anular em todos os pontos em que a derivada de θ se anula.

4. Aplicando o método de Euler ao pêndulo

Vejamos como aplicar o método de Euler ao problema de valor inicial

$$\ddot{\theta} = -\sin(\theta), \quad \theta(0) = \theta_0 \quad \text{e} \quad \dot{\theta}(0) = 0.$$

Tudo o que fizermos nesta seção aplica-se igualmente aos métodos de Runge-Kutta de ordens dois e quatro; vamos nos concentrar no método de Euler apenas porque produz uma recorrência mais simples. Criando uma nova variável v para denotar a velocidade do pêndulo, temos que

$$v = \dot{\theta}, \quad \text{mas também que} \quad \dot{v} = \ddot{\theta}.$$

Denotando por X o vetor coluna

$$\begin{bmatrix} \theta \\ v \end{bmatrix}$$

e por $F(t, X)$ a função vetorial

$$\begin{bmatrix} v \\ \sin(\theta) \end{bmatrix},$$

o problema de valor inicial com o qual começamos pode ser reescrito na forma

$$\dot{X} = F(t, X) \quad \text{e} \quad X(0) = X_0,$$

em que

$$\dot{X} = \begin{bmatrix} \dot{\theta} \\ \dot{v} \end{bmatrix} \quad \text{e} \quad X_0 = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}.$$

Neste caso a recorrência (84) toma a forma

$$X_{k+1} = X_k + hF(X_k),$$

para o valor de X_0 definido acima. Escrevendo a mesma recorrência em termos das coordenadas dos vetores obtemos,

$$\begin{bmatrix} \theta_{k+1} \\ v_{k+1} \end{bmatrix} = \begin{bmatrix} \theta_k \\ v_k \end{bmatrix} + h \begin{bmatrix} v_k \\ -\sin(\theta_k) \end{bmatrix}$$

Supondo que $\theta_0 = \pi/4$ e que $h = 0.1$, teremos, por exemplo, que

$$X_1 = \begin{bmatrix} 0.785 \\ 0 \end{bmatrix} + 0.1 \begin{bmatrix} 0 \\ -0.706 \end{bmatrix} = \begin{bmatrix} 0.785 \\ -0.071 \end{bmatrix}$$

e que

$$X_2 = \begin{bmatrix} 0.785 \\ -0.071 \end{bmatrix} + 0.1 \begin{bmatrix} -0.071 \\ -0.706 \end{bmatrix} = \begin{bmatrix} 0.778 \\ -0.140 \end{bmatrix}$$

Plotando as 100 primeiras iterações, com os valores de t nas abscissas e os valores de θ nas ordenadas, teremos o gráfico da figura 4.

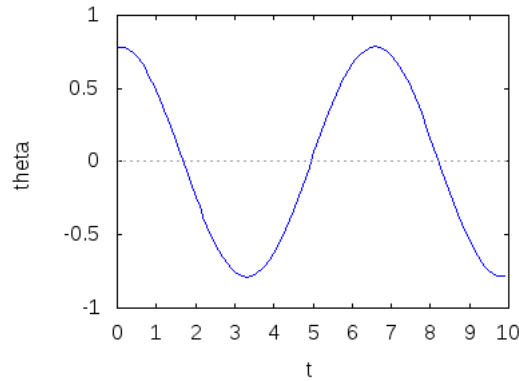


FIGURA 4. Solução numérica do pêndulo.

Podemos usar os dados calculados ao longo da iteração do método de Euler para desenhar um outro gráfico que nos diz muito sobre o comportamento do pêndulo. Desta vez plotaremos os valores da posição do pêndulo nas abscissas e as velocidades nas ordenadas. O plano resultante desta escolha das coordenadas é o *espaço de fase* da equação do pêndulo. No caso da iteração acima, a curva obtida no espaço de fase é ilustrada na figura 5.

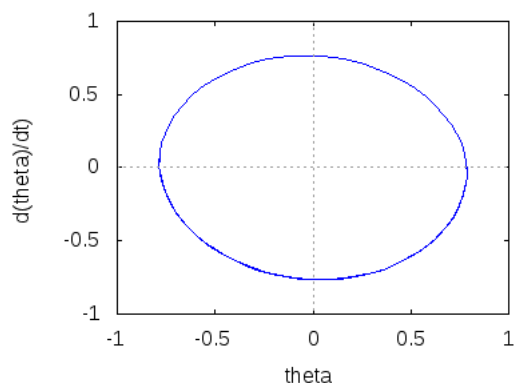


FIGURA 5. Solução numérica do pêndulo no espaço de fase.

A curva descrita pelo pêndulo no espaço de fase é fechada, porque estamos supondo que o pêndulo não está sujeito a nenhuma forma de amortecimento; isto é, não há nenhuma perda de energia. Com isso a posição e a velocidade do pêndulo oscilam entre valores máximos e mínimos que se repetem ao longo das oscilações. Além disso, o valor de θ é zero justamente quando o pêndulo atinge seu ponto mais baixo, que é quando a velocidade é máxima. Por outro lado, a haste forma um ângulo de $\pi/4$ radianos exatamente quando o pêndulo chega a seu ponto mais alto, caso em que a velocidade é nula. Por causa disto, a curva no espaço de fase é uma oval.

O método de Euler nos permite explorar versões mais complicadas do pêndulo, como aquela em que há amortecimento causado pela resistência do ar. Um modelo bastante simples consiste em supor que a resistência do ar é proporcional à velocidade. Neste caso, a equação do pêndulo terá a forma

$$(169) \quad \ddot{\theta} = -\sin(\theta) - c\dot{\theta},$$

em que c é uma constante. Note que, como há amortecimento, o pêndulo acabará parando de oscilar. Tomando $c = 0.1$, o gráfico da variação do ângulo ao longo do tempo é ilustrado na figura 6.

Como o ângulo e a velocidade do pêndulo decrescem à medida que o tempo passa, a trajetória descrita pelo pêndulo amortecido no espaço de fase não é mais uma oval, mas sim uma espiral, como mostra a figura 7 da página 200.

Exercícios

1. Nesta questão veremos como calcular o limite utilizado na equação (147) da página 182.

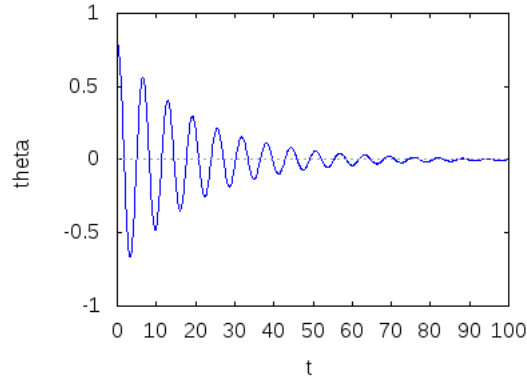


FIGURA 6. Solução numérica do pêndulo amortecido.

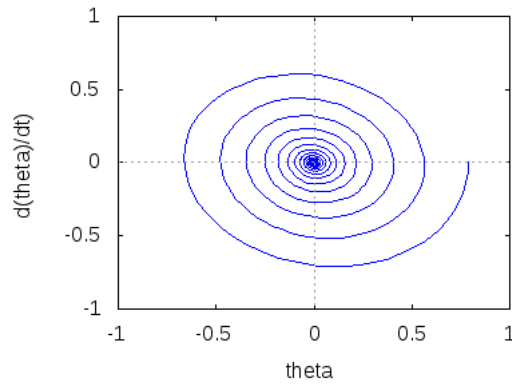


FIGURA 7. Espaço de fase do pêndulo amortecido.

- (a) Mostre que a mudança de variáveis $m = (2n - 1)/2$ nos permite escrever

$$\left(\frac{2n+1}{2n-1}\right)^n = \left(1 + \frac{1}{m}\right)^m \cdot \left(1 + \frac{1}{m}\right)^{1/2}.$$

- (b) Mostre que o limite do lado *esquerdo* da fórmula obtida em (a), quando n tende a infinito, é igual ao limite do seu lado *direito* quando m tende a infinito.
 (c) Calcule os limites dos dois fatores do lado direito fórmula obtida em (a) e deduza disto que

$$\lim_{n \rightarrow \infty} \left(\frac{2n+1}{2n-1}\right)^n = e.$$

2. Nesta questão determinamos a solução da recorrência (159) definida por

$$\theta_{k+1} = (1 + Ch)\theta_k + Th \quad \text{e} \quad \theta_0 = 0.$$

- (a) Denotando $1 + Ch$ por α e usando o método de substituições sucessivas, mostre que

$$\theta_{k+1} = (1 + Ch)^k \theta_0 + (1 + \alpha + \cdots + \alpha^{k-1})Th.$$

- (b) Deduza da fórmula da soma de uma progressão geométrica que

$$\theta_k = (1 + Ch)^k \theta_0 + \left(\frac{(1 + Ch)^k - 1}{(1 + Ch) - 1} \right) Th.$$

- (c) Levando em conta que $\theta_0 = 0$, mostre que

$$\theta_k = \frac{T}{C}(1 + Ch)^k - \frac{T}{C}.$$

Referências Bibliográficas

1. J. C. Butcher, *Numerical methods for ordinary differential equations*, second ed., John Wiley & Sons, Ltd., Chichester, 2008, With a foreword by J. M. Sanz-Serna.
2. H. C. Corben and Philip Stehle, *Classical mechanics*, Dover Publications, Inc., New York, 1994, Revised reprint of the second 1960 edition.
3. James F. Epperson, *On the Runge example*, Amer. Math. Monthly **94** (1987), no. 4, 329–341.
4. Gene H. Golub and Charles F. Van Loan, *Matrix computations*, fourth ed., Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, 2013.
5. Nicholas J. Higham, *Accuracy and stability of numerical algorithms*, second ed., Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002.
6. L. V. King, *On the numerical calculation of elliptic functions and integrals*, Cambridge University Press, 1924, Revised reprint of the second 1960 edition.
7. A. J. Lotka, *The frequency distribution of scientific productivity*, Journal of the Washington Academy of Sciences **16** (1926), 317–324.
8. Kim Plofker, *Mathematics in India*, Princeton University Press, Princeton, NJ, 2009.
9. Gay Robins and Charles Shute, *The Rhind mathematical papyrus*, British Museum Publications, Ltd., London, 1987, An ancient Egyptian text, With a note by T. G. H. James.
10. N. Rösch, *The derivation of algorithms to compute elliptic integrals of the first and second kind*, Bol. Ciênc. Geod. **17** (2011), 03–22.
11. Maxwell Rosenlicht, *Integration in finite terms*, Amer. Math. Monthly **79** (1972), 963–972.
12. Endre Süli and David F. Mayers, *An introduction to numerical analysis*, Cambridge University Press, Cambridge, 2003.
13. Richard S. Westfall, *Never at rest*, Cambridge University Press, Cambridge, 1980, A biography of Isaac Newton. MR 741027
14. E. T. Whittaker, *A treatise on the analytical dynamics of particles and rigid bodies*, Cambridge Mathematical Library, Cambridge University Press, Cambridge, 1988, With an introduction to the problem of three bodies, Reprint of the 1937 edition, With a foreword by William McCrea.