Departamento de Produção e Sistemas

Escola de Engenharia

Texto de Apoio

Métodos Numéricos e Otimização Não Linear

Isabel Espírito Santo

Conteúdo

Lista de Algoritmos

Capítulo 1

Erros e números

A base da computação numérica são os números. É inevitável que se cometam erros nos cálculos que se vão efetuando ao longo de qualquer processo numérico. Basta, para isso, o facto de serem usados números, isto para além dos erros que são inerentes aos dados e ao próprio método numérico usado para resolver o problema (Figura1.1).

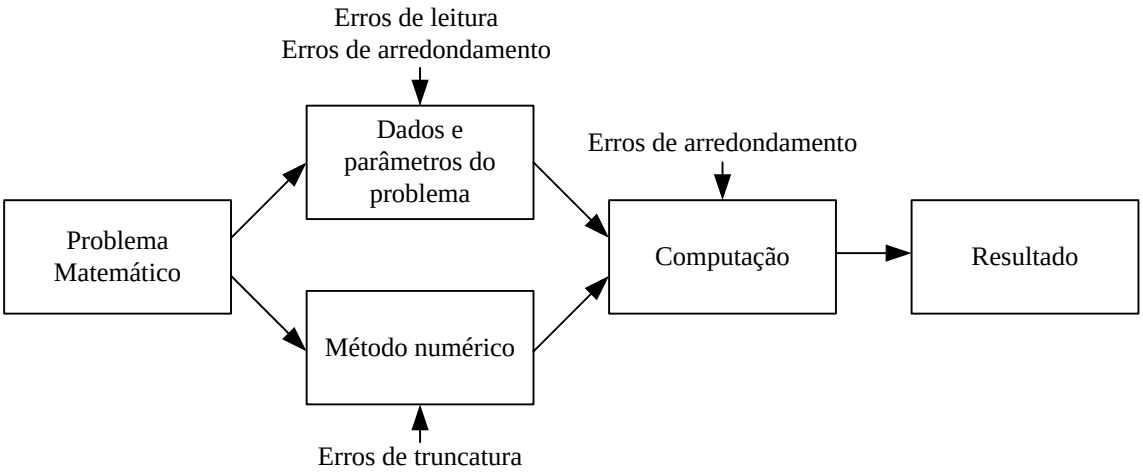


Figura 1.1: Representação dos erros cometidos ao longo da resolução de um problema com um método numérico

Assim, o primeiro tipo de erro que se comete, surge da representação do próprio número.

Quando se usam números que podem ser representados por uma sequência finita de dígitos, estes podem ser usados exatamente nos cálculos e, por isso, não se comete qualquer erro na sua representação (por exemplo, $\frac{1}{8} = 0.125$). No entanto, outros há que originam uma sequência

infinita de dígitos, pelo que a sua representação não é exata, sendo apenas possível utilizar-se um número limitado de dígitos (por exemplo $\sqrt{2} = 1.414213562...$). Assim sendo, comete-se um erro na sua representação.

1.1 Formato de vírgula flutuante

Uma das formas mais usuais de se representar um número é o formato de vírgula flutuante,

$$fl(x) = \pm 0.d_1 d_2 \dots d_k \times b^e,$$

em que d_k, $k = 1, \dots, k$ representam os k dígitos da mantissa, b é a base de representação (o mais usual é considerar-se $b = 10$) e e é o expoente.

A mantissa contém um número limitado de dígitos e esse número k define o comprimento da palavra no computador. Os cálculos podem ser feitos com precisão simples, quando se usam os k dígitos da mantissa ou com precisão dupla, caso se usem o dobro de dígitos da mantissa. No caso da precisão dupla, garante-se uma maior precisão nos cálculos.

1.2 Erro de arredondamento

Para se obter um número no formato de vírgula flutuante com uma mantissa de k dígitos pode recorrer-se a duas técnicas.

1. A **truncatura**, em que o formato da mantissa é o que está mais próximo de x e que se encontra entre x e 0.

   Exemplo 1.1

   $$x = \sqrt{2} = 1.414213562... = 0.1414213562... \times 10^1$$

   $$se\ k = 8,\ fl(x) = 0.14142135 \times 10^1$$

   $$se\ k = 6,\ fl(x) = 0.141421 \times 10^1$$

   O erro de aproximação será, assim, a diferença $x - fl(x)$.

2. O **arredondamento**, em que o formato da mantissa é o que está mais próximo de x.

Exemplo 1.2

$$x = \sqrt{2} = 1.414213562... = 0.1414213562... \times 10^1$$

$$se\ k = 8,\ fl(x) = 0.14142136 \times 10^1$$

$$se\ k = 6,\ fl(x) = 0.141421 \times 10^1$$

O erro de aproximação será, assim, a diferença $x - fl(x)$.

1.3 Erros inerentes aos dados

Por vezes, surgem erros por não ser possível atribuir, com exactidão, valores numéricos corretos aos dados obtidos por leitura experimental, proveniente de um equipamento. Por exemplo, medir um segmento de reta com uma régua, efetuar uma pesagem com uma balança... estes são os erros que são inerentes aos próprios dados.

1.4 Erro absoluto e erro relativo

A primeira questão que surge quando se fala de erros, quer de arredondamento, quer inerentes aos dados, é: como se avalia a proximidade do valor aproximado em relação ao valor exato?

Se x for o valor aproximado usado nos cálculos e x^* o valor exato, então

$$\text{erro absoluto} = x^* - x.$$

Esta diferença pode ser positiva, negativa ou nula.

O que acontece, em geral, é que o valor exato x^* é desconhecido e por isso também o erro. No entanto, é possível estabelecer um limite máximo para o erro que se comete, o que permite ter uma ideia da sua grandeza. Este conhecimento é, na maior parte das vezes, suficiente para se conhecer a exactidão do valor em causa.

Define-se, assim, como limite superior do erro absoluto a quantidade não negativa, δ_x, tal que

$$|\, x - x^*| \leq \delta_x$$

ou

$$x - \delta_x \leq x^* \leq x + \delta_x.$$

Apesar de útil, o erro absoluto, por si só, não é suficiente para que se perceba a exatidão do valor aproximado. Um certo erro absoluto pode ser considerado pequeno numas situações e grande noutras. Tudo depende do valor exato, x^*, que se está a medir. Para ultrapassar este problema, define-se o erro relativo, r_x, que depende, não apenas do erro absoluto, mas também do valor de x^*:

$$r_x = \frac{|x^* - x|}{|x^*|}.$$

Se δ_x for pequeno quando comparado com x^*, então define-se como limite superior do erro relativo a quantidade

$$r_x \simeq \frac{|x^* - x|}{|x|} \leq \frac{\delta_x}{|x|}.$$

Esta aproximação é muito importante e é a mais usada na prática porque, mais uma vez, o valor de x^* é quase sempre desconhecido.

1.5 Algarismos significativos

Quando se usa um valor aproximado x nem todos os algarismos são de confiança. Apenas o são os denominados algarismos significativos.

Os algarismos significativos são os que se encontram na parte de confiança do número e são

- diferentes de zero;
- iguais a zero mas não estão no número a indicar a posição do ponto decimal.

Para se determinar a parte de confiança de x, usa-se o limite superior do erro absoluto δ_x. Considera-se o majorante do erro imediatamente superior ao calculado ou conhecido e que comece pelo algarismo 5, isto é, que esteja na forma 0.5×10^n. Coloca-se x a concordar, isto é, na forma $\ldots \times 10^n$. A parte de confiança de x inclui os algarismos que se encontram à esquerda da posição que é equivalente à posição onde se encontra o algarismo 5 em δ_x.

Exemplo 1.3 *Seja* $x = 50.1234 \times 10^0$ *e* $\delta_x = 0.5 \times 10^{-3}$. *Então*

$$\begin{aligned} x &= 50123.4 \quad \times 10^{-3} \\ \delta_x &= 0.5 \quad \times 10^{-3} \end{aligned}$$

ou

$$x = 50.1234 \quad \times 10^0$$
$$\delta_x = 0.0005 \quad \times 10^0.$$

Logo, x tem 5 algarismos significativos.

Exemplo 1.4 *Seja $x = 13.5550 \times 10^0$. Então*

$$x = 13.5550 \quad \times 10^0$$
$$\delta_x = 0.0005 \quad \times 10^0.$$

Logo, x tem 5 algarismos significativos se $\delta_x = 0.5 \times 10^{-3}$, ou

$$x = 13. \quad 5550 \quad \times 10^0$$
$$\delta_x = 0. \quad 00005 \quad \times 10^0.$$

Logo, x tem 6 algarismos significativos se $\delta_x = 0.5 \times 10^{-4}$.

1.6 Fórmula fundamental dos erros

Quando se efetua um certo número de operações, há necessidade também de se avaliar o erro cometido. Para isso usa-se a fórmula fundamental dos erros, que permite calcular o limite superior do erro absoluto de uma expressão que envolve várias operações e várias variáveis, $f(x, y, z, \ldots)$. Conhecidos os valores aproximados das variáveis envolvidas na operação $x, y, z, \ldots$ e os correspondentes limites superiores dos erros absolutos $\delta_x, \delta_y, \delta_z, \ldots$, calculam-se os intervalos de valores possíveis para as variáveis,

$$I = \begin{cases} x - \delta_x \leq x^* \leq x + \delta_x \\ y - \delta_y \leq y^* \leq y + \delta_y \\ z - \delta_z \leq z^* \leq z + \delta_z \\ \ldots \end{cases}$$

De seguida, calculam-se os majorantes das derivadas parciais de f em ordem a cada uma das variáveis $x, y, z, \ldots$:

$$\left|[\frac{\partial f}{\partial x}]_I\right| \leq M_x\,, \left|[\frac{\partial f}{\partial y}]_I\right| \leq M_y\,, \left|[\frac{\partial f}{\partial z}]_I\right| \leq M_z\,, \ldots.$$

O limite superior do erro absoluto em f é dado por

$$\delta_f \leq M_x\delta_x + M_y\delta_y + M_z\delta_z + \cdots \tag{1.1}$$

Exemplo 1.5 *Seja* $x = 1.1$ *com 2 algarismos significativos,* $y = 2.04$ *com 3 algarismos significativos e* $z = 0.5$ *com 1 algarismo significativo. Quantos algarismos significativos tem o resultado de* $f(x, y, z) = -x + y + \sin z$*?*

$$\begin{cases} x = 1.1 & \Rightarrow \delta_x = 0.05 \\ y = 2.04 & \Rightarrow \delta_y = 0.005 \\ z = 0.5 & \Rightarrow \delta_z = 0.05 \end{cases} \quad e \quad I = \begin{cases} 1.05 & \leq x^* < 1.15 \\ 2.035 & \leq y^* < 2.045 \\ 0.45 & \leq z^* < 0.55. \end{cases}$$

Sendo que

$$f(x, y, z) = -x + y^2 + \sin z,$$

vem

$$\frac{\partial f}{\partial x} = -1\,, \quad \frac{\partial f}{\partial y} = 2y\,, \quad \frac{\partial f}{\partial z} = \cos(z),$$

ou seja,

$$M_x = 1,\ M_y = 2 \times 2.045 = 4.09,\ M_z = \cos(0.45) = 0.9004471024.$$

Assim, pela fórmula fundamental dos erros (1.1) *obtém-se*

$$\begin{aligned} \delta_f &\leq 1 \times 0.05 + 4.09 \times 0.005 + 0.9004471024 \times 0.05 \\ &= 0.115423551 \\ &\leq 0.5 \times 10^0. \end{aligned}$$

Como

$$f(1.1, 2.04, 0.5) = -1.1 + 2.04^2 + \sin(0.5) = 3.541025539,$$

$$\begin{aligned} f &= 3.\mathbf{5}41025539 \times 10^0 \\ \delta_f &= 0.\mathbf{5} \times 10^0 \\ &\Rightarrow 1 \textit{ algarismo significativo.} \end{aligned}$$

1.7 Erros de truncatura

Quando se resolvem problemas matemáticos recorrendo a um certo tipo de Métodos Numéricos, podem cometer-se erros de truncatura. Estes acontecem quando se utilizam métodos iterativos ou métodos de discretização.

1.7.1 Métodos iterativos

Um método iterativo é definido por uma equação iterativa, ou seja,

$$x_{k+1} = F(x_k),$$

a partir da qual é gerada uma sucessão de aproximações à solução exata do problema, x^*. k indica o índice da iteração e x_k representa a aproximação à solução numa determinada iteração k.

Um método iterativo gera uma sucessão de aproximações, que é iniciada numa aproximação inicial x_1:

$$x_1 \curvearrowright x_2 \curvearrowright x_3 \curvearrowright \ldots \curvearrowright x_n \curvearrowright x_{n+1} \ldots.$$

Se o processo iterativo estiver a convergir, esta sucessão converge para x^*.

Ao usar-se um método iterativo, a solução x^* é atingida ao fim de um número infinito de operações. Por isso, face aos recursos limitados em termos de tempo e de memória, o processo iterativo tem de ser terminado ao fim de um número finito de iterações. Para isso usa-se um critério de paragem que garante que a aproximação calculada, x_{n+1}, está suficientemente próxima de x^*.

O erro de truncatura é dado pela diferença

$$x^* - x_{n+1}.$$

1.7.2 Métodos de discretização

Este tipo de métodos transforma um problema matemático que envolve conceitos de natureza contínua, tais como a diferenciação e a integração, num problema discreto, que envolve apenas operações algébricas. A este processo chama-se discretização.

Exemplo 1.6 *Dado o integral*

$$I = \int_1^{1.2} \sqrt{x}\, dx,$$

este pode ser transformado em

$$I \approx S(0.05) = \frac{0.05}{3}[f(1) + 4f(1.05) + 2f(1.1) + 4f(1.15) + f(1.2)]$$

através da fórmula composta de Simpson.

O erro de truncatura, neste caso, é dado pela diferença $I - S(0.05)$.

Capítulo 2

Sistemas de equações lineares

2.1 Forma geral do problema

O problema que se aborda neste capítulo é um sistema de n equações e n variáveis que tem a forma geral

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ a_{31}x_1 + a_{32}x_2 + \ldots + a_{3n}x_n &= b_3 \\ \ldots & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n \end{cases}$$

ou

$$Ax = b,$$

em que $A_{n\times n}$ é a matriz dos coeficientes do sistema com n linhas e n colunas, $x = (x_1\, x_2\, x_3\, \ldots, x_n)^T$ é o vetor solução, $b = (b_1\, b_2\, b_3\, \ldots, b_n)^T$ é o vetor dos termos independentes e $(A|b)$ é a matriz ampliada do sistema.

2.2 Existência e unicidade da solução

Um sistema de equações lineares, em termos do número de soluções, pode ser classificado como

- **possível e determinado**, quando tem uma solução única,
- **possível e indeterminado** quando tem um número infinito de soluções e

- **impossível** quando não tem solução.

A existência e unicidade da solução do sistema linear depende da característica das matrizes A e $(A|b)$.

A característica de uma matriz A, $c(A)$, é dada pelo número de linhas ou colunas linearmente independentes.

Existe uma relação direta entre $c(A)$, o determinante de A, $det(A)$, e a existência da inversa de A, A^{-1}.

- Se $c(A) = n$, então $det(A) \neq 0$, A^{-1} existe e o sistema é **possível e determinado**, ou seja, tem uma solução única.
- Se $c(A) < n$, então $det(A) = 0$, A^{-1} não existe e
  - se $c(A) = c(A|b)$ o sistema é **possível e indeterminado**, ou seja, tem uma infinidade de soluções,
  - se $c(A) < c(A|b)$ o sistema é **impossível**, ou seja, não tem solução.

2.3 Métodos para a resolução de $Ax = b$

Há vários métodos para a resolução de sistemas de equações lineares, no entanto, de um modo geral, podem agrupar-se em dois tipos:

- **Métodos diretos**

  São aplicados a sistemas de pequena ou média dimensão e a solução exata é obtida num número finito de operações.

- **Métodos iterativos**

  São aplicados a sistemas de grande dimensão e a solução exata só se obtém ao fim de um número infinito de operações.

2.4 Método directo de eliminação de Gauss com pivotagem parcial para resolução de sistemas lineares

Um dos métodos mais usados para resolver sistemas de equações lineares é o método de eliminação de Gauss com pivotagem parcial (EGPP). Este distingue-se do método de eliminação de

Gauss (EG) por permitir trocas de linhas, que garantem que o pivot tenha um valor numérico não superior a um. O sucesso deste método deve-se ao facto de ser direto, simples e numericamente estável, ao contrário, por exemplo, do EG. Ao resolver-se um sistema por EGPP, uma vez que se garante que o multiplicador nunca é superior a um, os erros de arredondamento nunca serão propagados, podendo ser, pelo contrário, minimizados. Atente-se ao exemplo 2.1

Exemplo 2.1 *Considere o sistema linear, cuja solução é* $(1,1)^T$

$$\begin{cases} 10^{-20}x_1 + x_2 &= 1 \\ x_1 + x_2 &= 2 \end{cases}$$

- *Se for resolvido por EG,*

$$\left(\begin{array}{cc|c} 10^{-20} & 1 & 1 \\ 1 & 1 & 2 \end{array}\right) \underset{m_{21}=-10^{20}}{\longrightarrow} \left(\begin{array}{cc|c} 10^{-20} & 1 & 1 \\ 0 & -10^{20} & -10^{20} \end{array}\right)$$

$$\rightarrow \begin{cases} 10^{-20}x_1 + x_2 &= 1 \\ -10^{20}x_2 &= -10^{20} \end{cases} \rightarrow \begin{cases} x_1 &= 0 \\ x_2 &= 1 \end{cases}$$

- *Se for resolvido por EGPP,*

$$\left(\begin{array}{cc|c} 10^{-20} & 1 & 1 \\ 1 & 1 & 2 \end{array}\right) \rightarrow \left(\begin{array}{cc|c} 1 & 1 & 2 \\ 10^{-20} & 1 & 1 \end{array}\right) \underset{m_{21}=-10^{-20}}{\longrightarrow} \left(\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 1 & 1 \end{array}\right)$$

$$\rightarrow \begin{cases} x_1 + x_2 &= 2 \\ x_2 &= 1 \end{cases} \rightarrow \begin{cases} x_1 &= 1 \\ x_2 &= 1 \end{cases}$$

Com base neste exemplo pode verificar-se que o método EGPP é numericamente estável, ao contrário do EG, que pode tornar-se numericamente instável.

O método de EGPP para resolver um sistema de equações lineares, pode dividir-se em 2 passos:

1. **transformar** $Ax = b$ **em** $Ux = c$ **usando EGPP**

   em que U é uma matriz triangular superior. Os sistemas $Ax = b$ e $Ux = c$ são equivalentes, isto é, têm a a mesma solução.

   Há operações elementares de matrizes que se podem efetuar de forma a transformar A em U:

   (a) trocar duas linhas paralelas,

(b) multiplicar uma linha por um escalar diferente de zero,

(c) substituir uma linha pela que dela é obtida adicionando o produto de outra linha paralela por um escalar.

2. resolver $Ux = c$ por substituição inversa

Exemplo 2.2 *Considere a matriz ampliada do sistema*

$$\left(\begin{array}{ccc|c} 0.3 & -0.2 & 10 & 71.4 \\ 0.1 & 7 & -0.3 & -19.3 \\ 3 & -0.1 & -0.2 & 7.85 \end{array}\right)$$

Este sistema tem dimensão $n = 3$, o que significa que o primeiro passo de resolução, isto é, a transformação de A em U, tem $n - 1 = 2$ etapas.

Etapa 1

- *Colocar 'pivot' na posição $(1, 1)$. O 'pivot' é o valor de maior módulo de entre todos os valores da primeira coluna.*
- *Trocar a linhas 1 com a linha 3.*

$$\left(\begin{array}{ccc|c} \mathbf{3} & -0.1 & -0.2 & 7.85 \\ 0.1 & 7 & -0.3 & -19.3 \\ 0.3 & -0.2 & 10 & 71.4 \end{array}\right) \underset{I_{1,3}(A|b)}{\Leftrightarrow} \quad com \quad I_{1,3} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Para reduzir a zero os elementos 0.1 e 0.3, calculam-se os escalares - denominados multiplicadores:

- $m_{21} = -\dfrac{0.1}{\mathbf{3}} = -0.033333$
- $m_{31} = -\dfrac{0.3}{\mathbf{3}} = -0.1$

Nota: *não esquecer que $|$ multiplicador $| \leq 1$ conserva a estabilidade numérica.*

$m_{21}\times$ (linha 1) + linha 2 e $m_{31}\times$ (linha 1) + linha 3 $\Rightarrow$

$$\left(\begin{array}{ccc|c} \mathbf{3} & -0.1 & -0.2 & 7.85 \\ 0 & 7.003333 & -0.293333 & -19.561664 \\ 0 & -0.19 & 10.02 & 70.615 \end{array}\right) \underset{J_1\, I_{1,3}(A|b)}{\Leftrightarrow} \quad com \quad J_1 = \begin{pmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & 0 & 1 \end{pmatrix}.$$

Etapa 2 colocar o 'pivot' na posição $(2,2)$*. O pivot é o elemento de maior módulo na coluna dois, abaixo da linha 1.*

- *não é necessário trocarem-se linhas,*
- *para reduzir a zero o elemento* -0.19*, calcula-se o multiplicador* $m_{32} = \dfrac{-0.19}{\mathbf{7.003333}} = 0.027130$
- $m_{32} \times$ *(linha 2)* $+$ *(linha 3)* $\Rightarrow$

$$\begin{pmatrix} 3 & -0.1 & -0.2 & | & 7.85 \\ 0 & \mathbf{7.003333} & -0.293333 & | & -19.561664 \\ 0 & 0 & 10.012042 & | & 70.084292 \end{pmatrix} \underset{J_2\, J_1\, I_{1,3}(A|b)}{\Leftrightarrow} \quad com \quad J_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & m_{32} & 1 \end{pmatrix}.$$

- *A matriz ampliada está, nesta fase, na forma* $(U|c)$ *que corresponde ao sistema* $Ux = c$*.*
- *Os cálculos efetuados sobre a matriz ampliada correspondem ao procedimento:*

$$J_2\, J_1\, I_{1,3}(A|b) = (U|c).$$

Para resolver o sistema $Ux = c$ *por substituição inversa:*

$$\begin{cases} 3x_1 & -0.1x_2 & -0.2x_3 & = & 7.85 \\ & 7.003333x_2 & -0.293333x_3 & = & -19.561664 \\ & & 10.012042x_3 & = & 70.084292 \end{cases}$$

$$x_3 = \frac{70.084292}{10.012042} = 7,$$

$$x_2 = \frac{-19.561664 + 0.293333(7)}{7.003333} = -2.5,$$

$$x_1 = \frac{7.85 + 0.2(7) + 0.1(-2.5)}{3} = 3.$$

A solução do sistema é

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3 \\ -2.5 \\ 7 \end{pmatrix}.$$

2.5 EGPP para o cálculo do determinante de uma matriz

Para calcular o determinante de uma matriz quadrada pode usar-se o método de eliminação de Gauss com pivotagem parcial. Para isso usa-se EGPP para transformar A na matriz triangular superior U. De seguida, usam-se as propriedades do determinante. Atente-se no Exemplo 2.3.

Exemplo 2.3 *Calcular o determinante de*

$$A = \begin{pmatrix} 0.3 & -0.2 & 10 \\ 0.1 & 7 & -0.3 \\ 3 & -0.1 & -0.2 \end{pmatrix}$$

O procedimento para transformar A em U foi (Exemplo 2.2)

$$J_2 J_1 I_{1,3} A = U.$$

Então,

$$\begin{aligned} & \det(J_2 J_1 I_{1,3} A) = \det(U) \Leftrightarrow \\ \Leftrightarrow \quad & \det(J_1) \times \det(J_1) \times \det(I_{1,3}) \times \det(A) = \det(U) \Leftrightarrow \\ \Leftrightarrow \quad & 1 \times 1 \times (-1)\det(A) = 3 \times 7.003333 \times 10.012042 \Leftrightarrow \\ \Leftrightarrow \quad & \det(A) = -210.352994. \end{aligned}$$

Como regra, estabelece-se que

$$det(A) = (-1)^r \times U_{ii},\ i = 1 \ldots, n,$$

sendo r o número de trocas de linhas efetuadas durante o processo de EGPP.

2.6 EGPP para o cálculo da inversa de uma matriz

Dada uma matriz quadrada $A_{n\times n}$, se $\det(A) \neq 0$, então existe uma matriz inversa de A, A^{-1}, tal que $A \times A^{-1} = A^{-1} \times A = I$, sendo I a matriz identidade. Quando existe a inversa de uma matriz A, esta matriz A diz-se não singular.

Quando $\det(A) = 0$, a matriz A não tem inversa e diz-se singular.

Para calcular a inversa de uma matriz A pode usar-se o método de eliminação de Gauss com pivotagem parcial. As n colunas de A^{-1} são as soluções dos n sistemas

$$AX = I \quad \Rightarrow \quad X = A^{-1}.$$

Usa-se EGPP para transformar A na matriz triangular superior U, fazendo as mesmas operações em simultâneo sobre a matriz I, ou seja, transforma-se a matriz ampliada $(A|I)$ em $(U|J)$.

De seguida, resolvem-se os n sistemas, em que os termos independentes são as colunas de J, por substituição inversa.

Atente-se no Exemplo 2.4.

Exemplo 2.4 *Calcular* A^{-1} *sendo*

$$A = \begin{pmatrix} 2.71 & 1.63 & 0.32 \\ 4.11 & 2.44 & 0.19 \\ 2.69 & 1.64 & 0.36 \end{pmatrix}$$

A matriz ampliada dos $n = 3$ *sistemas é*

$$(A|I) = \left(\begin{array}{ccc|ccc} 2.71 & 1.63 & 0.32 & 1 & 0 & 0 \\ 4.11 & 2.44 & 0.19 & 0 & 1 & 0 \\ 2.69 & 1.64 & 0.36 & 0 & 0 & 1 \end{array}\right)$$

Etapa 1:

$$I_{1,2}(A|I) \Leftrightarrow \left(\begin{array}{ccc|ccc} 4.11 & 2.44 & 0.19 & 0 & 1 & 0 \\ 2.71 & 1.63 & 0.32 & 1 & 0 & 0 \\ 2.69 & 1.64 & 0.36 & 0 & 0 & 1 \end{array}\right) \begin{array}{l} \\ m_{21} = -0.65937 \\ m_{31} = -0.65450 \end{array}$$

$$J_1 I_{1,2}(A|I) \Leftrightarrow \left(\begin{array}{ccc|ccc} 4.11 & 2.44 & 0.19 & 0 & 1 & 0 \\ 0 & 0.02114 & 0.19472 & 1 & -0.65937 & 0 \\ 0 & 0.04302 & 0.23565 & 0 & -0.65450 & 1 \end{array}\right)$$

Etapa 2:

$$I_{2,3}J_1I_{1,2}(A|I) \Leftrightarrow \left(\begin{array}{ccc|ccc} 4.11 & 2.44 & 0.19 & 0 & 1 & 0 \\ 0 & 0.04302 & 0.23565 & 0 & -0.65450 & 1 \\ 0 & 0.02114 & 0.19472 & 1 & -0.65937 & 0 \end{array}\right) m_{32} = -0.49140$$

$$J_2I_{2,3}J_1I_{1,2}(A|I) \Leftrightarrow \left(\begin{array}{ccc|ccc} 4.11 & 2.44 & 0.19 & 0 & 1 & 0 \\ 0 & 0.04302 & 0.23565 & 0 & -0.65450 & 1 \\ 0 & 0 & 0.07892 & 1 & -0.33775 & -0.49140 \end{array}\right)$$

A inversa calcula-se coluna a coluna resolvendo os $n = 3$ sistemas por substituição inversa:

- ***1ª coluna***

$$\left(\begin{array}{ccc|c} 4.11 & 2.44 & 0.19 & 0 \\ 0 & 0.04302 & 0.23565 & 0 \\ 0 & 0 & 0.07892 & 1 \end{array}\right) \longrightarrow \begin{pmatrix} 40.62 \\ -69.40807 \\ 12.67106 \end{pmatrix}$$

- ***2ª coluna***

$$\left(\begin{array}{ccc|c} 4.11 & 2.44 & 0.19 & 1 \\ 0 & 0.04302 & 0.23565 & -0.65450 \\ 0 & 0 & 0.07892 & -0.33775 \end{array}\right) \longrightarrow \begin{pmatrix} 4.44403 \\ 8.22872 \\ -4.27965 \end{pmatrix}$$

- ***3ª coluna***

$$\left(\begin{array}{ccc|c} 4.11 & 2.44 & 0.19 & 0 \\ 0 & 0.04302 & 0.23565 & 1 \\ 0 & 0 & 0.07892 & -0.49140 \end{array}\right) \longrightarrow \begin{pmatrix} -33.76063 \\ 57.35214 \\ -6.22656 \end{pmatrix}$$

A inversa da matriz A é, assim,

$$A^{-1} = \begin{pmatrix} 40.62 & 4.44403 & -33.76063 \\ -69.40807 & 8.22872 & 57.35214 \\ 12.67106 & -4.27965 & -6.22656 \end{pmatrix}$$

Capítulo 3

Equações não lineares

3.1 Forma geral do problema

O problema que se pretende tratar neste capítulo tem a forma geral

$$f(x) = 0, \;\; f : \mathbb{R}^n \to \mathbb{R}^n.$$

No caso de uma equação não linear com uma variável, $n = 1$, o problema reduz-se a

$$f(x) = 0, \;\; \text{com} \;\; f : \mathbb{R} \to \mathbb{R},$$

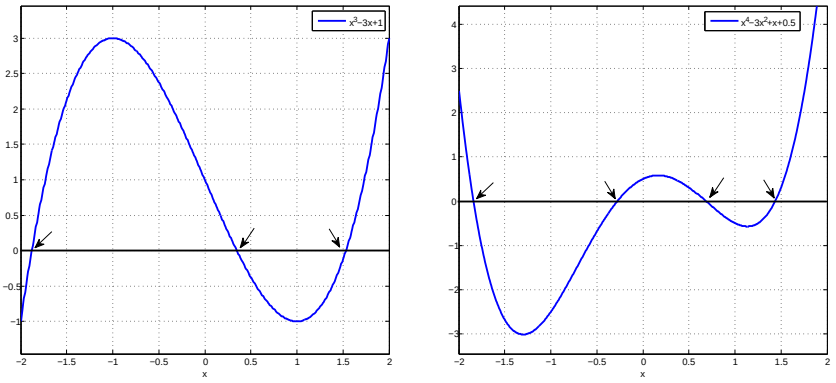
em que x é a variável independente do problema e $y = f(x)$ é a variável dependente do problema.

Quando f é não linear em x, esta equação pode não ter soluções reais, ter uma só solução real, ter várias soluções reais, ter soluções complexas e ter soluções reais e complexas. Uma solução real da equação $f(x) = 0$ é equivalente a uma raiz real da equação $f(x) = 0$, a um zero real da função f ou ainda à interseção de f com o eixo do X no plano real XOY.

Podem dividir-se as equações não lineares em dois tipos genéricos.

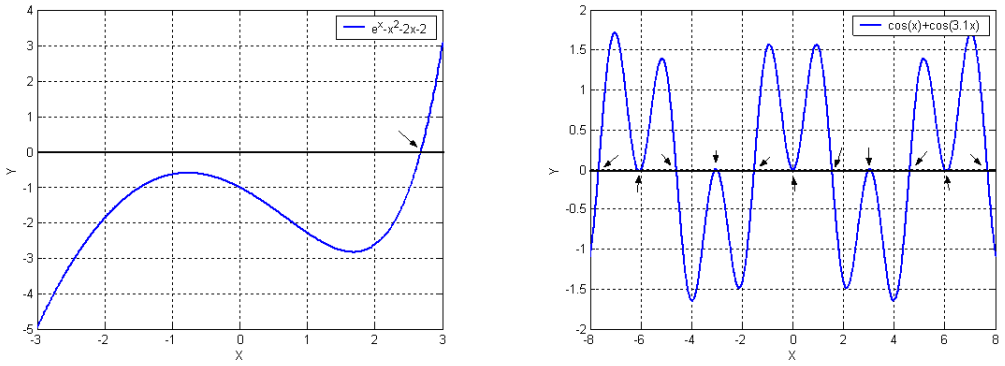
1. **as algébricas**, que envolvem apenas operações aritméticas básicas, de que são exemplo os polinómios (Figura 3.1),

2. **as transcendentes**, que envolvem funções trigonométricas, exponenciais, logarítmicas, entre outras, como são exemplos as equações $e^x - x^2 - 2x - 2 = 0$ e $\cos(x) + \cos(3.1x) = 0$ (Figura 3.2).



(a) $x^3 - 3x + 1 = 0$ (b) $x^4 - 3 * x^2 + x + 0.5 = 0$

Figura 3.1: Soluções de duas equações algébricas.



(a) $e^x - x^2 - 2x - 2 = 0$ (b) $\cos(x) + \cos(3.1x) = 0$

Figura 3.2: Soluções de duas equações transcendentes.

3.2 Solução de uma equação não linear

Nem sempre é fácil, ou mesmo possível, resolver uma equação não linear analiticamente. Os métodos mais usados na resolução deste tipo de problemas são os **métodos iterativos**. No entanto, estes métodos exigem que seja fornecida uma ou várias aproximações iniciais. Estas podem ser identificadas através da localização de raízes, por exemplo, recorrendo a **métodos gráficos**.

3.2.1 Métodos gráficos

Os métodos gráficos não permitem conhecer em rigor a solução de uma equação não linear, mas permitem localizá-la. Tornam-se, assim, de extrema importância quando se pretende

saber onde se encontra a solução ou soluções de uma equação não linear, como é o caso de se determinar uma aproximação ou aproximações iniciais que servirão de ponto de partida para a implementação de um método iterativo. Este processo pode fazer-se através da representação gráfica de:

- $f(x)$ no plano XOY se f for fácil de representar,
- $g(x)$ e $h(x)$ no plano XOY, em que

$$f(x) = 0 \Leftrightarrow g(x) = h(x)$$

se f for difícil de representar e g e h forem fáceis de representar.

Exemplo 3.1 *Localização gráfica de raízes*

Considere a equação $x^3 - 3x + 1 = 0$. *Para as suas raízes pode fazer-se a representação gráfica de* $f(x) = x^3 - 3x + 1$.

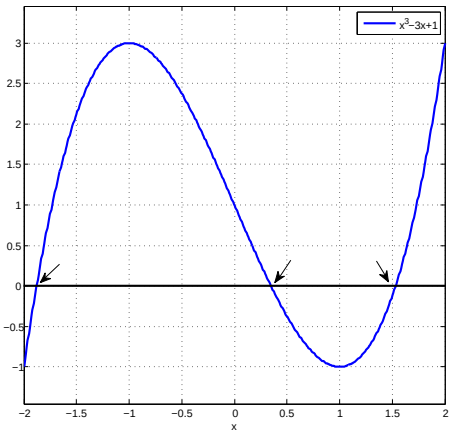


Figura 3.3: Representação gráfica de $f(x)$.

Pode verificar-se na Figura 3.3 a localização das três soluções desta equação. Uma próxima de -2, outra próxima de 0.5 e outra ainda próxima de 1.5.

No entanto, esta equação pode ser reformulada:

$$x^3 - 3x + 1 = 0 \Leftrightarrow x^3 - (3x - 1) = 0 \Leftrightarrow \begin{cases} x^3 & = & 3x - 1 \\ g(x) & = & h(x) \end{cases}$$

Desta forma, os zeros de $f(x) = x^3 - 3x + 1$ *são os pontos de interseção de* $g(x) = x^3$ *com* $h(x) = 3x - 1$. *Como pode verificar-se na Figura 3.4, as soluções encontradas são exatamente as mesmas que pelo processo anterior.*

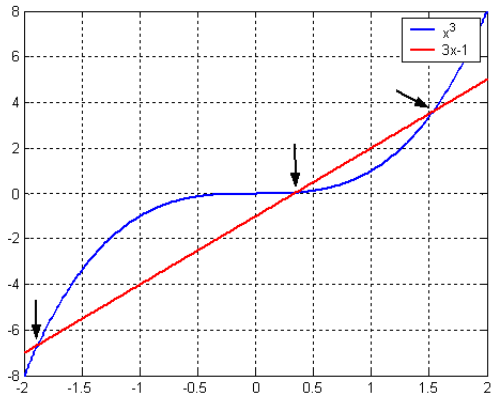


Figura 3.4: Representação gráfica de $g(x)$ e $h(x)$.

Exemplo 3.2 *Localização gráfica de raízes*

Considere a equação $\sin x + x - 2 = 0$. *Esta equação pode resolver-se por métodos gráficos representando no plano XOY* $f(x) = \sin x + x - 2$. *Pode ver-se na Figura 3.5 que esta equação*

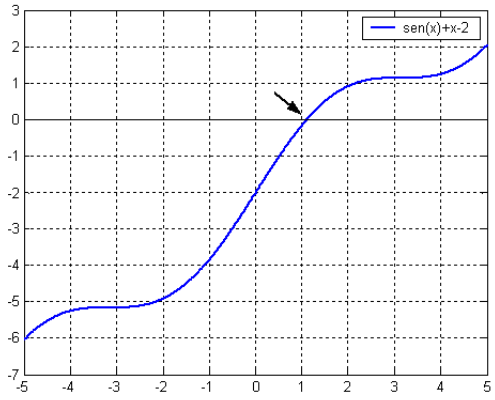


Figura 3.5: Representação gráfica de $f(x)$.

tem uma única solução, próxima de 1.

Da mesma forma que no Exemplo 3.1, pode desdobrar-se a função $f(x) = \sin x + x - 2$ *em duas mais simples.*

$$\sin x + x - 2 = 0 \Leftrightarrow \sin x - (-x + 2) = 0 \Leftrightarrow \begin{cases} \sin x &= -x + 2 \\ g(x) &= h(x) \end{cases}$$

Assim, os zeros de $f(x) = \sin x + x - 2$ *são os pontos onde se verifica a interseção de* $g(x) =$

$\sin x$ *com* $h(x) = -x + 2$. *Mais uma vez, a solução obtida é exatamente a mesma que na representação gráfica anterior, como se pode verificar na Figura 3.6* ($x \approx 1$).

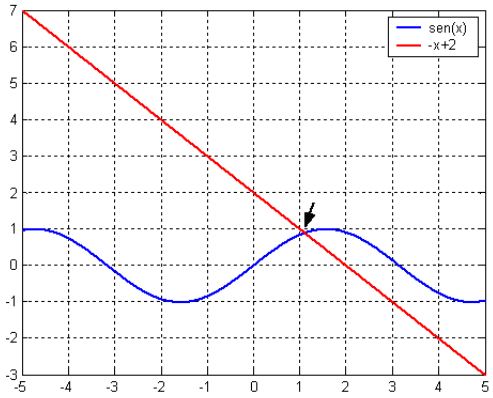


Figura 3.6: Representação gráfica de $g(x)$ e $h(x)$.

3.2.2 Métodos iterativos

Da representação gráfica, como pode verificar-se na Subsecção 3.2.1, é possível retirar um intervalo que contenha a solução pretendida ou apenas um valor que esteja próximo da solução pretendida. A aproximação obtida por métodos gráficos pode ser melhorada através de métodos iterativos. Entre estes métodos encontram-se

- o método da secante;
- o método de Newton.

Qualquer um destes métodos pode ser usado para calcular raízes reais de $f(x) = 0$. No entanto, podem também ser usados para calcular raízes complexas, desde que se introduza aritmética complexa nos cálculos e as aproximações iniciais sejam números complexos.

Como qualquer processo iterativo, o método da secante e o método de Newton exigem um critério de paragem, já que a solução exata só é obtida para um número infinito de iterações, o que em termos computacionais não é exequível. Os critérios de paragem usados nestes métodos são os que se apresentam em (3.1) e (3.2). Os valores ε_1 e ε_2 são quantidades positivas e próximas de zero. Quanto menores forem estas quantidades, mais próxima será a aproximação x_{k+1} da solução x^*.

Critérios de paragem

São usados como critérios de paragem dos métodos iterativos para a resolução de equações não lineares

- a estimativa do erro relativo da aproximação próxima de zero

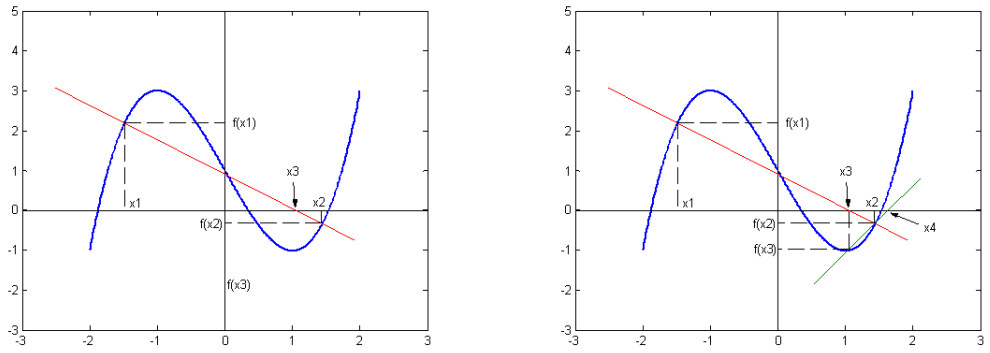
$$\frac{|x_{k+1} - x_k|}{|x_{k+1}|} \leq \varepsilon_1, \tag{3.1}$$

- o valor absoluto da função na última aproximação próximo de zero

$$|f(x_{k+1})| \leq \varepsilon_2. \tag{3.2}$$

3.2.3 Método da secante

O método da secante precisa de duas aproximações iniciais para iniciar o processo iterativo. Em cada iteração, e com base em dois pontos (aproximações), o método aproxima $f(x)$ por uma reta definida por esses dois pontos (secante). O ponto de interseção desta reta com o eixo do X fornece a aproximação seguinte. Em termos gráficos, podem ver-se na Figura 3.7 as duas primeiras iterações do método da secante. Na primeira iteração (Figura 3.7(a)), parte-se



(a) primeira iteração (b) segunda iteração

Figura 3.7: Duas iterações do método da secante

de duas aproximações iniciais $(x_1, f(x_1))$ e $(x_2, f(x_2))$. Traça-se a reta que passa nestes dois pontos e obtém-se a aproximação seguinte $(x_3, f(x_3))$. Na segunda iteração (Figura 3.7(b)), a reta é traçada a partir dos pontos $(x_2, f(x_2))$ e $(x_3, f(x_3))$. A nova aproximação é $(x_4, f(x_4))$. O processo iterativo é repetido até se obter uma solução que verifique os critérios de paragem.

Equação iterativa do método da secante

Sejam dois pontos $(x_{k-1}, f(x_{k-1}))$ e $(x_k, f(x_k))$. O ponto de interseção da reta secante que passa por estes dois pontos e o eixo do X é x_{k+1} dado por

$$x_{k+1} = x_k - \frac{(x_k - x_{k-1})f(x_k)}{f(x_k) - f(x_{k-1})}, \quad k = 2, 3, \ldots$$

Descreve-se, de forma detalhada, o método da secante no Algoritmo 3.1.

Algoritmo 3.1 Método da secante

ler: x_1 e x_2 (aproximações iniciais)

$k \leftarrow 1$

repetir

$k \leftarrow k + 1$

$x_{k+1} \leftarrow x_k - \dfrac{(x_k - x_{k-1})f(x_k)}{f(x_k) - f(x_{k-1})}$

até $\dfrac{|x_{k+1} - x_k|}{|x_{k+1}|} \leq \varepsilon_1 \wedge |f(x_{k+1})| \leq \varepsilon_2$

$x^* \leftarrow x_{k+1}$

$f(x^*) \leftarrow f(x_{k+1})$

Exemplo 3.3 *Um certo equipamento de* 20000 *euros vai ser pago durante* 6 *anos. O pagamento anual é de* 4000 *euros. A relação entre o custo do equipamento P, o pagamento anual A, o número de anos n e a taxa de juro i é a seguinte:*

$$A = P\frac{i(1+i)^n}{(1+i)^n - 1}.$$

Utilize o método que não recorre à derivada para determinar a taxa de juro utilizada nos cálculos. O valor da taxa de juro pertence ao intervalo $[0.05, 0.15]$. *Use* $\varepsilon_1 = \varepsilon_2 = 0.005$. *Use seis casas decimais nos cálculos.*

Resolução:

Substituindo os valores de P, A e n, vem

$$4000 = 20000\frac{i(1+i)^6}{(1+i)^6 - 1} \quad \Leftrightarrow \quad 20000\frac{i(1+i)^6}{(1+i)^6 - 1} - 4000 = 0 \quad \Leftrightarrow \quad \frac{5i(1+i)^6}{(1+i)^6 - 1} - 1 = 0$$

Logo,

$$f(i) = \frac{5i(1+i)^6}{(1+i)^6 - 1} - 1.$$

- **1ª iteração** *($k = 2$)*

$i_1 = 0.05, \quad f(i_1) = -0.014913$

$i_2 = 0.15, \quad f(i_2) = 0.321185$

$$i_3 = i_2 - \frac{(i_2 - i_1)f(i_2)}{f(i_2) - f(i_1)} = 0.054437$$

Critério de Paragem

$f(i_3) = -0.000891$

$|f(i_3)| \leq \varepsilon_2 \Leftrightarrow 0.000891 \leq 0.005 \quad (verdadeiro)$

`Nota:` *Uma vez que se verifica esta condição, para o processo iterativo poder terminar tem que se verificar também a outra.*

$$\frac{|i_3 - i_2|}{|i_3|} \leq \varepsilon_1 \Leftrightarrow 1.755479 \leq 0.005 \quad (falso)$$

- **2ª iteração** *($k = 3$)*

$i_2 = 0.15, \quad f(i_2) = 0.321185$

$i_3 = 0.054437, \quad f(i_3) = -0.000891$

$$i_4 = i_3 - \frac{(i_3 - i_2)f(i_3)}{f(i_3) - f(i_2)} = 0.054701$$

Critério de Paragem

$f(i_4) = -0.000054$

$|f(i_4)| \leq \varepsilon_2 \Leftrightarrow 0.000054 \leq 0.005 \quad (verdadeiro)$

`Nota:` *Uma vez que se verifica esta condição, para o processo iterativo poder terminar tem que se verificar também a outra.*

$$\frac{|i_4 - i_3|}{|i_4|} \leq \varepsilon_1 \Leftrightarrow 0.004826 \leq 0.005 \quad (verdadeiro)$$

Uma vez que se verificam ambas as condições do critério de paragem, o processo iterativo termina com $i \approx 0.054701$ *e* $f(i) \approx 0.000054$.

Condições de convergência do método da secante

Nem sempre se pode garantir que o método da secante convirja. Para garantir a convergência para a solução é necessário que

- x^* é tal que $f(x^*) = 0$,
- $f(x)$ é continuamente diferenciável,
- $f'(x^*) \neq 0$,
- as aproximações iniciais x_1 e x_2 têm de se encontrar na vizinhança de x^* (convergência local).

Se se verificarem as condições anteriores, o método iterativo da secante converge e

$$\lim_{k\to\infty} \frac{|x^* - x_{k+1}|}{|x^* - x_k|^p} = L, \quad L > 0, \ p = 1.618.$$

Por essa razão diz-se que o método da secante exibe convergência superlinear.

Situação de divergência

Ao longo do processo iterativo, pode acontecer que duas aproximações tenham valores de f muito próximos. Esta situação leva a uma divergência do método (Figura 3.8).

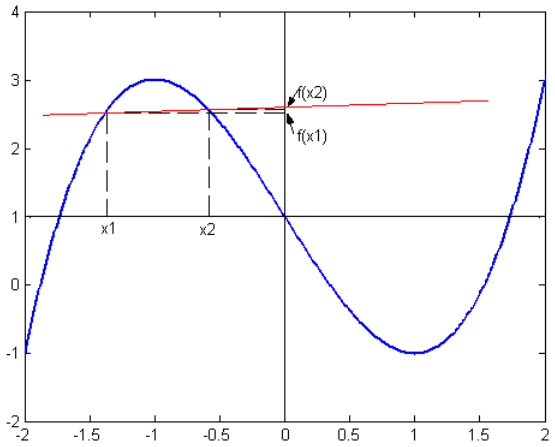
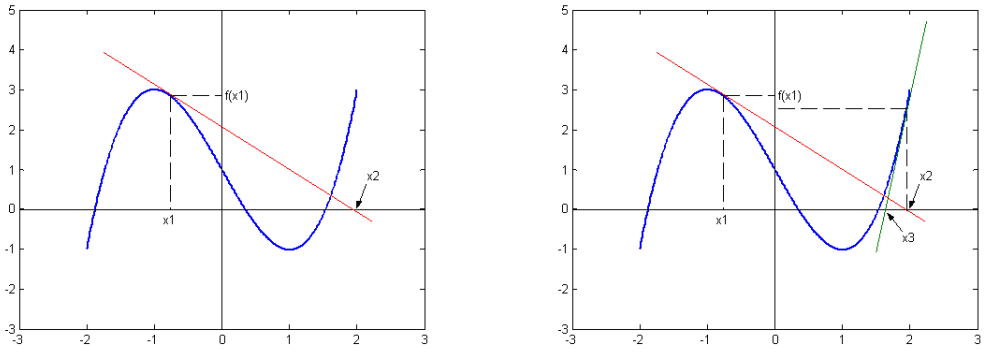


Figura 3.8: Situação de divergência no método da secante.

3.2.4 Método de Newton

O método de Newton precisa apenas de uma aproximação inicial. Em cada iteração usa informação de f e de f' relativa a um ponto, aproximando a função $f(x)$ por uma reta que é tangente a f nesse ponto. O ponto de interseção dessa recta com o eixo do X fornece a aproximação seguinte.

Na primeira iteração (Figura 3.9(a)), parte-se de uma aproximação inicial $(x_1, f(x_1))$. Traça-se a reta tangente a $f(x)$ nesse ponto e obtém-se a aproximação seguinte $(x_2, f(x_2))$. Na segunda iteração (Figura 3.9(b)), a reta é traçada a partir do ponto $(x_2, f(x_2))$. A nova aproximação é $(x_3, f(x_3))$. O processo iterativo é repetido até se obter uma solução com a precisão desejada.



(a) primeira iteração (b) segunda iteração

Figura 3.9: Duas iterações do método de Newton

Equação iterativa do método de Newton

Considere-se o ponto $(x_k, f(x_k))$ e $f'(x_k)$. Então, o ponto de interseção da reta, que passa por este ponto, com declive definido por $f'(x_k)$ com o eixo do X é x_{k+1}, sendo

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 1, 2, \ldots$$

Descreve-se, de forma detalhada, o método de Newton no Algoritmo 3.2.

Algoritmo 3.2 Método de Newton

ler: x_1 (aproximação inicial)

$k \leftarrow 0$

repetir

$k \leftarrow k + 1$

$x_{k+1} \leftarrow x_k - \dfrac{f(x_k)}{f'(x_k)}$

até $\dfrac{|x_{k+1} - x_k|}{|x_{k+1}|} \leq \varepsilon_1 \wedge |f(x_{k+1})| \leq \varepsilon_2$

$x^* \leftarrow x_{k+1}$

$f(x^*) \leftarrow f(x_{k+1})$

Exemplo 3.4 *Considere a seguinte função* $f : \mathbb{R} \to \mathbb{R}$ *definida por*

$$f(x) = x^2 - e^x.$$

Calcule a raiz negativa utilizando o método iterativo de Newton. Considere $x_1 = 0.25$. *Faça duas iterações e apresente o erro relativo.*

Resolução:

- **1ª iteração** *(*$k = 1$*)*

  $x_1 = 0.25, \quad f(x_1) = -1.2215, \quad f'(x_1) = -0.7840$

  $$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = -1.3080$$

- **2ª iteração** *(*$k = 2$*)*

  $x_2 = -1.3080, \quad f(x_2) = 1.4405, \quad f'(x_2) = -2.8864$

  $$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)} = -0.8089$$

Ao fim da segunda iteração, $x \approx -0.8089$ *e o erro relativo é*

$$\frac{|x_3 - x_2|}{|x_3|} = 0.6170.$$

Condições de convergência do método de Newton

O método de Newton nem sempre converge. Para se garantir uma convergência para a solução de uma equação não linear através deste método é necessário que

- x^* é tal que $f(x^*) = 0$,
- $f(x)$ é continuamente diferenciável,
- $f'(x^*) \neq 0$ e
- a aproximação inicial x_1 tem de se encontrar na vizinhança de x^* (convergência local).

Se se verificarem todas as condições anteriores, o método de Newton converge e

$$\lim_{k\to\infty} \frac{|x^* - x_{k+1}|}{|x^* - x_k|^p} = L, \quad L > 0, \ p = 2.$$

Por esta razão diz-se que o método de Newton exibe convergência quadrática.

Situação de divergência

Ao longo do processo iterativo pode acontecer que o declive da reta que é tangente a f na aproximação, que vai ser usada para gerar a nova aproximação, seja um valor próximo de zero. Esta situação leva a uma divergência do método (Figura 3.10).

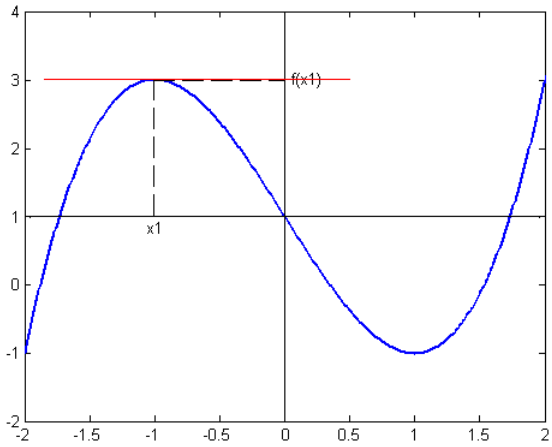


Figura 3.10: Situação de divergência no método de Newton.

3.2.5 Método da secante *versus* método de Newton

Os métodos iterativos da secante e de Newton têm vantagens e desvantagens. Como tal, devem ser escolhidos de forma a adequarem-se o melhor possível ao problema que se está a resolver. Apresentam-se de seguida algumas linhas gerais que podem ser seguidas de forma a ajudar na escolha de um ou outro em determinadas situações.

- Quando ambos os métodos convergem, o método de Newton é, em geral, mais rápido, já que exibe convergência quadrática, ao passo que o método da secante tem convergência superlinear.

- O método da secante necessita apenas de informação sobre $f(x)$ ao passo que o método de Newton exige também informação sobre $f'(x)$. Se $f(x)$ não for diferenciável ou se a sua expressão analítica for muito complexa, deve optar-se pelo método da secante, já que, ou não é possível o cálculo da derivada ou este é demasiado dispendioso, em termos de esforço computacional.

3.3 Solução de um sistema de equações não lineares

3.3.1 Forma geral do problema

O problema que se pretende resolver tem a forma geral

$$f(x) = 0,\ f : \mathbb{R}^n \to \mathbb{R}^n \Leftrightarrow \begin{cases} f_1(x_1, x_2, \ldots, x_n) &= 0 \\ f_2(x_1, x_2, \ldots, x_n) &= 0 \\ \cdots & \\ f_n(x_1, x_2, \ldots, x_n) &= 0, \end{cases}$$

em que f é o vetor $(f_1\ f_2\ \cdots\ f_n)^T$, x é o vetor $(x_1\ x_2\ \cdots\ x_n)^T$ e pelo menos uma das funções de f é não linear (Exemplo 3.5).

Exemplo 3.5 *Considerando o sistema de equações não lineares com três equações e três variáveis* (x_1,x_2 *e* x_3)

$$\begin{cases} 3x_1 - \cos(x_2 x_3) - 0.5 &= 0 \\ x_1^2 - 625x_2^2 &= 0 \\ e^{-x_1 x_2} + 20x_3 + 9 &= 0, \end{cases}$$

este pode ser reescrito na forma genérica

$$\begin{cases} f_1(x_1, x_2, \ldots, x_n) &= 0 \\ f_2(x_1, x_2, \ldots, x_n) &= 0 \\ f_3(x_1, x_2, \ldots, x_n) &= 0 \end{cases} \Leftrightarrow f(x) = 0.$$

Assim, f é o vetor que contém as três funções f_1, f_2 e f_3, ou seja, $f = (f_1\ f_2\ f_3)^T$ e x é o vetor que contém as três variáveis x_1, x_2 e x_3, ou seja, $x = (x_1\ x_2\ x_3)^T$. Trata-se de um problema de dimensão $n = 3$.

3.3.2 Método de Newton

Para resolver um sistema de equações não lineares ($n > 1$) pode usar-se o método de Newton, que não é mais que uma generalização do caso em que este método se aplica à resolução de uma equação não linear ($n = 1$).

A equação iterativa do método de Newton para resolver uma equação não linear, $f(x) = 0$, $f : \mathbb{R} \to \mathbb{R}$, pode ser reformulada e escrita na forma

$$x_{k+1} = x_k + \Delta x_k, \ \ k = 1, 2, \ldots$$

com

$$\begin{aligned} \Delta x_k &= -\frac{1}{f'(x_k)} f(x_k) \\ &= -\left(f'(x_k)\right)^{-1} f(x_k). \end{aligned}$$

Assim, para um sistema de n equações em n variáveis, a primeira derivada da função, que é um escalar, é substituída pela matriz do Jacobiano $n \times n$ do vetor de funções, e a equação iterativa fica

$$x_{k+1} = x_k + \Delta x_k, \ \ k = 1, 2, \ldots$$

com

$$\Delta x_k = -\left(J(x_k)\right)^{-1} f(x_k), \tag{3.3}$$

em que $x \in \mathbb{R}^n$, $\Delta x \in \mathbb{R}^n$, $f \in \mathbb{R}^n$ e $J(x) \in \mathbb{R}^{n \times n}$.

A matriz do Jacobiano contém as primeiras derivadas parciais das funções f_1, f_2,...,f_n

em ordem às variáveis x_1, x_2,...,x_n.

$$J(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \dots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \dots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}.$$

Equação iterativa do método de Newton

A equação (3.3) pode ser reescrita da seguinte forma:

$$J(x_k)\Delta x_k = -f(x_k). \tag{3.4}$$

Significa que o vetor Δx_k em cada iteração se obtém através da resolução do sistema linear (3.4) pelo método de eliminação de Gauss com pivotagem parcial, por exemplo.

Critério de paragem

Tal como para as equações não lineares, podem usar-se como critério de paragem as medidas

- estimativa do erro relativo da aproximação

$$\frac{\|\Delta x_k\|_2}{\|x_{k+1}\|_2} \leq \varepsilon_1 \tag{3.5}$$

- aproximação ao zero de f

$$\|f(x_{k+1})\|_2 \leq \varepsilon_2 \tag{3.6}$$

sendo ε_1 e ε_2 quantidades positivas e próximas de zero. $\|.\|_2$ representa a norma 2.
A implementação do método de Newton encontra-se descrita no Algoritmo 3.3.

Algoritmo 3.3 Método de Newton para sistemas de equações não lineares

ler: x_1 (aproximação inicial)

$k \leftarrow 0$

calcular $J(x)$

repetir

$k \leftarrow k + 1$

calcular $J(x_k)$

calcular $f(x_k)$

resolver o sistema linear $J(x_k)\Delta x_k = -f(x_k)$ por EGPP para calcular o vetor Δx_k

$x_{k+1} \leftarrow x_k + \Delta x_k$

até $\dfrac{\|\Delta x_k\|_2}{\|x_{k+1}\|_2} \leq \varepsilon_1 \wedge \|f(x_{k+1})\|_2 \leq \varepsilon_2$

$x^* \leftarrow x_{k+1}$

$f(x^*) \leftarrow f(x_{k+1})$

Exemplo 3.6 *Considere o sistema do Exemplo 3.5. Tome* $x^1 = (0, 1, 0)^T$ *para valor inicial e faça apenas uma iteração. Apresente uma medida da proximidade ao zero das funções.*

Resolução:

1ª iteração *(*$k = 1$*)*

$$x_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad f(x_1) = \begin{pmatrix} -1.5 \\ -625 \\ 10 \end{pmatrix}, \quad J(x_1) = \begin{pmatrix} 3 & 0 & 0 \\ 0 & -1250 & 0 \\ -1 & 0 & 20 \end{pmatrix}$$

Resolver o sistema linear $J(x_1)\Delta_1 = -f(x_1)$ *por EGPP para calcular* Δ_1*:*

$$\left(\begin{array}{ccc|c} 3 & 0 & 0 & 1.5 \\ 0 & -1250 & 0 & 625 \\ -1 & 0 & 20 & -10 \end{array}\right)$$

$$\Delta_1 = \begin{pmatrix} -0.5 \\ -0.5 \\ -0.475 \end{pmatrix}$$

$$x_2 = x_1 + \Delta_1 = \begin{pmatrix} 0.5 \\ 0.5 \\ -0.475 \end{pmatrix}$$

Aproximação ao zero absoluto das funções

$$f(x_2) = \begin{pmatrix} 0.0281 \\ -156 \\ 0.7840 \end{pmatrix}$$

$\|f(x_2)\|_2 = 156.0020$

Condições de convergência do método de Newton

Para que o método de Newton convirja quando aplicado à resolução do sistema

$$\begin{cases} f_1(x_1, x_2, \ldots, x_n) &= 0 \\ \ldots \\ f_n(x_1, x_2, \ldots, x_n) &= 0 \end{cases}$$

em que $f = (f_1 \ \ldots \ f_n)^T$ e $x = (x_1 \ \ldots \ x_n)^T$, é necessário que

- x^* é tal que $f(x^*) = 0$,
- f é um vetor de funções continuamente diferenciáveis,
- $J(x^*)$ é uma matriz não singular $(\exists (J(x^*))^{-1})$,
- $(J(x^*))^{-1}$ é limitada $(\|J(x_k) - J(x^*)\| \leq \beta,\ \beta > 0)$
- $J(x)$ é matriz Lipschitz contínua na vizinhança de x^*, ou seja, $\exists \gamma > 0 : \|J(x_k) - J(x^*)\| \leq \gamma \|x_k - x^*\|$,
- A aproximação inicial, x_1, tem de estar na vizinhança de x^* - convergência local.

Verificando-se todas estas condições,

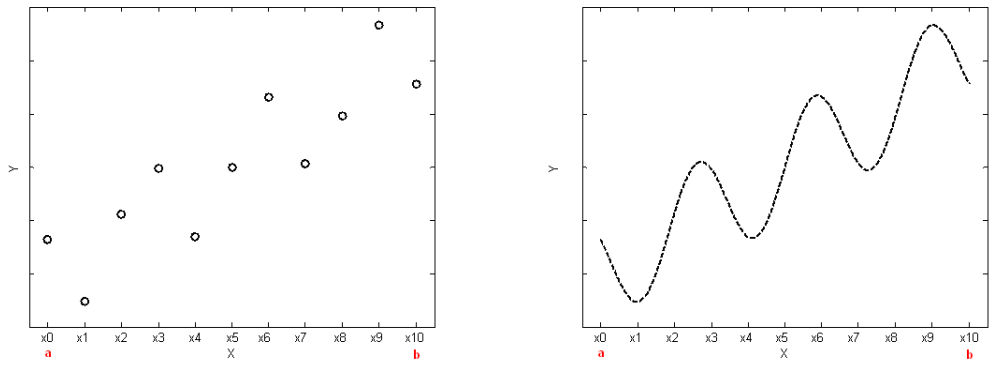
$$\lim_{k \to \infty} \frac{\|x^* - x_{k+1}\|}{\|x^* - x_k\|^p} = L, \ \ L > 0, \ p = 2,$$

o que significa que o método de Newton para a resolução de sistemas não lineares exibe convergência quadrática.

Capítulo 4

Polinómio interpolador de Newton

O polinómio interpolador é uma das técnicas disponíveis para a aproximação de funções. O objetivo é encontrar uma aproximação, neste caso concreto através de um polinómio, $p_n(x)$, à função dada, $f(x)$, com o menor erro possível. Faz-se este tipo de aproximação em duas situações:



(a) Função dada por um conjunto de 11 pontos (b) Função dada por uma expressão matemática

Figura 4.1: Aproximação de funções.

1. Dado um conjunto discreto de valores (Figura 4.1(a))

$$(x_i, f_i), \quad i = 0, 1, \ldots, n \quad (n+1 \quad \text{pontos})$$

pretende-se encontrar uma relação funcional (expressão matemática) entre as variáveis x e f para prever o comportamento entre as variáveis e estimar valores, em que x é a variável independente e f é a variável dependente.

2. Dada uma função $f(x)$ por uma expressão matemática complicada (Figura 4.1(b)), pretende-se conhecer uma expressão mais simples que descreva o melhor possível o comportamento de f como função de x.

4.1 Erro da aproximação

Teorema 4.1.1 *Teorema de Weirstrass*

Dadas a função $f(x)$, contínua num intervalo $[a, b]$, e uma quantidade $\varepsilon > 0$, existe sempre um polinómio $p_n(x)$, de grau menor ou igual a n, tal que o erro da aproximação $\|e_n(x)\| = \|f(x) - p_n(x)\| \leq \varepsilon$.

Pelo Teorema 4.1.1 pode assegurar-se que o erro seja igual a zero para um conjunto de $n + 1$ pontos selecionados do intervalo $[a, b]$, isto é, o polinómio passa por esses $n + 1$ pontos da função,

$$f_i \equiv f(x_i) = p_n(x_i), \quad \text{para } i = 0, 1, \ldots, n$$

e que este polinómio é único e de grau menor ou igual a n.

Os polinómios interpoladores mais conhecidos são

- o polinómio interpolador de Newton baseado em diferenças divididas,
- o polinómio interpolador de Lagrange,

que são ambos polinómios de colocação.

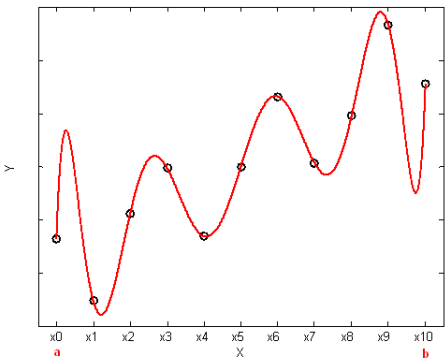
Pode ainda definir-se o polinómio dos mínimos quadrados, em que se assegura que a soma dos quadrados dos erros é mínima no intervalo $[a, b]$, isto é,

$$\min \sum_{i=0}^{n} (e_n(x))^2 \equiv \sum_{i=0}^{n} (f_i - p_n(x_i))^2,$$

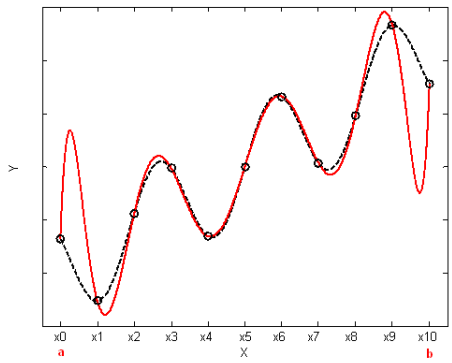
mas este assunto será discutido em detalhe noutro capítulo. Pode ver-se na Figura 4.2 a diferença entre usar-se um polinómio interpolador (Figuras 4.2(a) e 4.2(b)) ou um polinómio dos mínimos quadrados (Figuras 4.2(c) e 4.2(d)).

Deve escolher-se um ou outro método de acordo com a situação que se tem em mãos.

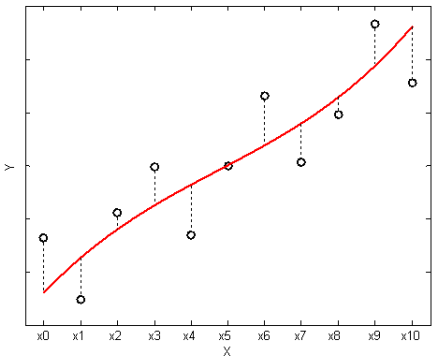
Se os dados forem precisos, isto é, não contêm erros de observação, é mais vantajoso usar-se uma função que passe pelos pontos dados, e por isso o método mais adequado, neste caso, é o polinómio interpolador.



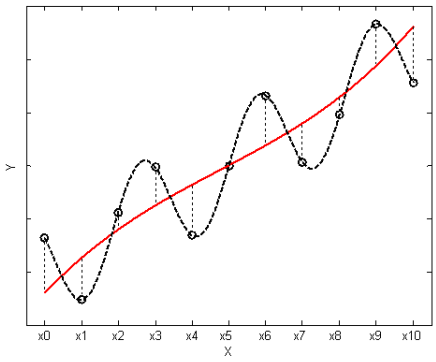
(a) Polinómio interpolador de grau 10 (conjunto de 11 pontos)



(b) Polinómio interpolador de grau 10 (função dada por uma expressão matemática)



(c) Polinómio de grau 3 dos mínimos quadrados (conjunto de 11 pontos)



(d) Polinómio de grau 3 dos mínimos quadrados (função dada por uma expressão matemática)

Figura 4.2: Aproximação polinomial.

Se, pelo contrário, os dados possuem erros de observação, torna-se mais vantajoso encontrar uma função que descreva o comportamento dos dados, sem a preocupação da curva passar pelos pontos, e por isso deve usar-se o polinómio dos mínimos quadrados. Esta situação ocorre, por exemplo, quando se estão a recolher dados experimentais associados a um aparelho com um erro associado.

Deve ainda ter-se em conta que quando o número de pontos disponível é muito grande, o polinómio interpolador será de um grau muito elevado, tornando-se, por isso, muito ruidoso e irregular, e o seu uso é também, neste caso, desaconselhável. O que se faz muitas vezes é selecionar um número limitado de pontos na região que interessa interpolar. Podendo assim usar-se um polinómio interpolador de grau mais baixo.

4.2 Diferenças divididas

4.2.1 Definição

Considere-se um conjunto de $n+1$ pontos diferentes entre si (x_0, f_0), $(x_1, f_1), \ldots, (x_n, f_n)$.

Podem definir-se as diferenças divididas como

- Diferenças divididas de 1ª ordem

$$[x_0, x_1] = \frac{f_0 - f_1}{x_0 - x_1} = \frac{f_1 - f_0}{x_1 - x_0}$$

$\ldots$

$$[x_{n-1}, x_n] = \frac{f_{n-1} - f_n}{x_{n-1} - x_n} = \frac{f_n - f_{n-1}}{x_n - x_{n-1}}$$

- Diferenças divididas de 2ª ordem

$$[x_0, x_1, x_2] = \frac{[x_0, x_1] - [x_1, x_2]}{x_0 - x_2} = \frac{[x_1, x_2] - [x_0, x_1]}{x_2 - x_0}$$

$\ldots$

$$[x_{n-2}, x_{n-1}, x_n] = \frac{[x_{n-2}, x_{n-1}] - [x_{n-1}, x_n]}{x_{n-2} - x_n} = \frac{[x_{n-1}, x_n] - [x_{n-2}, x_{n-1}]}{x_n - x_{n-2}}$$

- $\ldots$

- Diferença dividida de nª ordem

$$[x_0, x_1, \ldots, x_{n-1}, x_n] = \frac{[x_0, x_1, \ldots, x_{n-1}] - [x_1, x_2, \ldots, x_n]}{x_0 - x_n} = \frac{[x_1, x_2, \ldots, x_n] - [x_0, x_1, \ldots, x_{n-1}]}{x_n - x_0}$$

As diferenças divididas podem ser colocadas numa tabela com a forma

$$
\begin{array}{llllllll}
x_0 & f_0 & & & & & & \\
 & & [x_0, x_1] & & & & & \\
x_1 & f_1 & & [x_0, x_1, x_2] & & & & \\
 & & [x_1, x_2] & & [x_0, x_1, x_2, x_3] & & & \\
x_2 & f_2 & & [x_1, x_2, x_3] & & \dots & & \\
 & & [x_2, x_3] & & [x_1, x_2, x_3, x_4] & & [x_0, x_1, \dots, x_n] & \\
x_3 & f_3 & & \dots & & \dots & & \\
\vdots & \vdots & \dots & & \dots & & & \\
x_{n-1} & f_{n-1} & & [x_{n-2}, x_{n-1}, x_n] & & & & \\
 & & [x_{n-1}, x_n] & & & & & \\
x_n & f_n & & & & & &
\end{array}
$$

4.2.2 Propriedades das diferenças divididas

As diferenças divididas gozam de algumas propriedades importantes que a seguir se listam.

1. Podem ser calculadas para qualquer espaçamento, mesmo sendo não constante, entre os pontos disponíveis $x_0, x_1, \dots, x_{n-1}, x_n$.

2. As diferenças divididas são funções simétricas dos seus argumentos, isto é

   $[x_0, x_1] = [x_1, x_0]$

   $[x_0, x_1, x_2] = [x_2, x_1, x_0]$

   $\dots$

3. As diferenças divididas de n^{a} ordem de um polinómio de grau n são iguais entre si e diferentes de zero. Por consequência, as diferenças divididas de $(n+1)^{\text{a}}$ ordem são iguais a zero neste caso.

4.3 Polinómio interpolador de Newton baseado em diferenças divididas

O polinómio de grau menor ou igual a n, $p_n(x)$, é construído com base em $n+1$ pontos e é único.

Para simplificar a notação, considere-se $f_i \equiv f(x_i)$. Sejam os $n+1$ pontos:

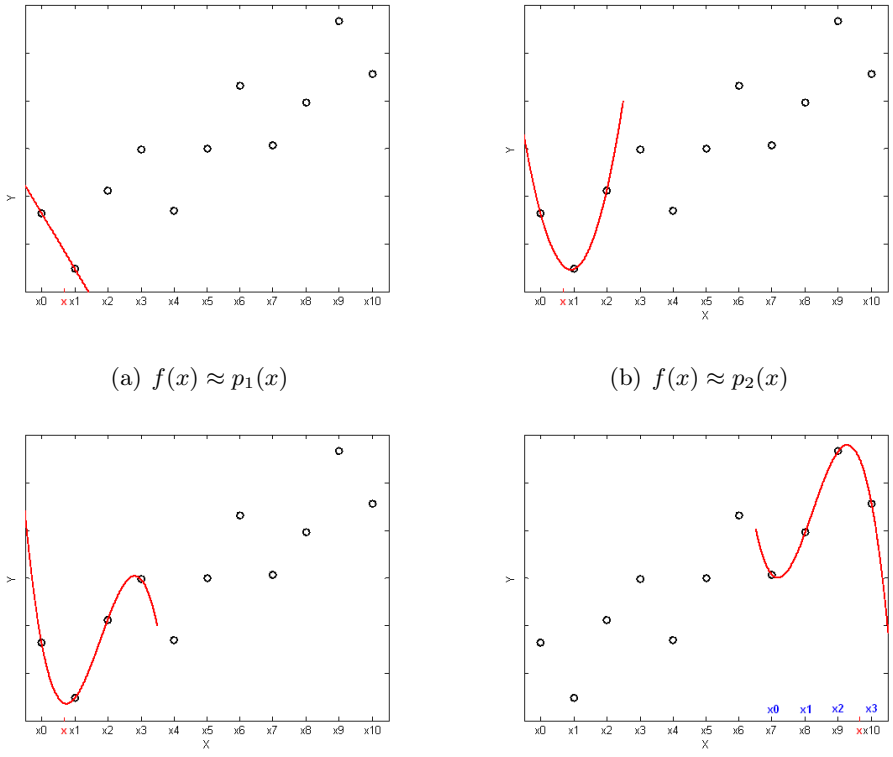
x_0	x_1	x_2	x_3	x_4	$\cdots$	x_{n-2}	x_{n-1}	x_n
f_0	f_1	f_2	f_3	f_4	$\cdots$	f_{n-2}	f_{n-1}	f_n

O polinómio interpolador de Newton de grau $\leq n$ é

$$\begin{aligned} p_n(x) &= f_0 + (x-x_0)[x_0,x_1] + (x-x_0)(x-x_1)[x_0,x_1,x_2] + (x-x_0)(x-x_1)(x-x_2)[x_0,x_1,x_2,x_3] \\ &+ \cdots + (x-x_0)(x-x_1)(x-x_2)\cdots(x-x_{n-1})[x_0,x_1,x_2,\ldots,x_{n-1},x_n] \end{aligned}$$

4.3.1 Interpolação direta

O objetivo da interpolação direta é estimar o valor de $f(\overline{x})$, sendo $\overline{x}$ um ponto que não está na tabela.



(a) $f(x) \approx p_1(x)$ (b) $f(x) \approx p_2(x)$

(c) $f(x) \approx p_3(x)$ (d) $f(x) \approx p_3(x)$

Figura 4.3: Alguns exemplos de polinómios interpoladores

Para construir um polinómio de grau n, devem escolher-se $n+1$ pontos da tabela de pontos disponíveis, garantindo que pelo menos um dos pontos está à direita de $\overline{x}$ e outro à

sua esquerda, se isso for possível. Os restantes pontos da tabela escolhem-se de forma a que estejam o mais próximo possível de $\overline{x}$.

Na Figura 4.3(a) apresentam-se os pontos selecionados (dois) e o polinómio de grau um resultante para estimar o valor da função no ponto representado a vermelho. Na Figura 4.3(b) apresenta-se a mesma situação mas para um polinómio de grau dois (três pontos). Nas Figuras 4.3(c) e 4.3(d) apresentam-se dois polinómios de grau três diferentes (quatro pontos) para dois pontos interpoladores diferentes, representados a vermelho.

No Exemplo 4.1 é apresentada a forma como os pontos que entram na construção de um polinómio interpolador devem ser escolhidos.

Exemplo 4.1 *Escolha de quatro pontos para construir um polinómio de grau três.*

Dada a tabela

x_i	-1	0	1	2	8	10	12	15	20
f_i	-5	-2	-1	3	0	-2	-1	4	6

que pontos devem ser escolhidos para construir um polinómio de grau três se o ponto interpolador for

a) $\overline{x} = 3$*?*

$$\begin{array}{cccc} x_0 & x_1 & x_2 & x_3 \\ \hline f_0 & f_1 & f_2 & f_3 \end{array} \quad \Leftrightarrow \quad \begin{array}{cccc} 0 & 1 & 2 & 8 \\ \hline -2 & -1 & 3 & 0 \end{array}$$

b) $\overline{x} = 13$*?*

$$\begin{array}{cccc} x_0 & x_1 & x_2 & x_3 \\ \hline f_0 & f_1 & f_2 & f_3 \end{array} \quad \Leftrightarrow \quad \begin{array}{cccc} 8 & 10 & 12 & 15 \\ \hline 0 & -2 & -1 & 4 \end{array}$$

4.4 Erro de truncatura

O erro de truncatura do polinómio interpolador é dado por

$$e_n(x) = f(x) - p_n(x) = (x - x_0)(x - x_1) \cdots (x - x_{n-1})(x - x_n) \frac{f^{(n+1)}(\xi)}{(n+1)!},$$

com $\xi \in [a, b]$.

O majorante do erro de truncatura cometido com a aproximação, para um certo x do intervalo $[a, b]$, que contém os pontos usados para construir o polinómio de grau n, x_0, x_1, x_2, ..., x_{n-1}, x_n, pode ser estimado de duas formas:

1. Se $f(x)$ for dada por uma expressão, então

$$|e_n(x)| \leq |(x - x_0)(x - x_1) \cdots (x - x_{n-1})(x - x_n)| \, \frac{1}{(n+1)!} M_{n+1}$$

em que

$$\left| \left[f^{(n+1)}(x) \right]_{[a,b]} \right| \leq M_{n+1}.$$

2. Se $f(x)$ for dada por um conjunto discreto de pontos,

$$|e_n(x)| \leq |(x - x_0) \cdots (x - x_{n-1})(x - x_n)| \, |(\text{dd de ordem } n+1)|$$

em que

$$(\text{dd de ordem } n+1) = [x_0, x_1, \ldots, x_{n-1}, x_n, x_z]$$

se só existir uma, ou a maior delas em valor absoluto se existirem mais que uma.

Exemplo 4.2 *Dada a tabela de valores de uma função* $f(x)$

x_i	0.0	0.1	0.2	0.3	0.4	0.5	0.8	1.0
$f(x_i)$	0	1	1	2	2	3	3	4

a) Pretende-se aproximar $f(0.6)$ *usando um polinómio de grau 3. Use a fórmula interpoladora de Newton baseada em diferenças divididas.*

b) Estime o erro de truncatura cometido na alínea anterior.

Resolução:

a) Para estimar $f(0.6)$ *usando um polinómio de grau 3 são necessários 4 pontos. A tabela das diferenças divididas para os 4 pontos mais próximos de 0.6 é*

x_i	f_i	$dd1$	$dd2$	$dd3$
0.3	2			
		0		
0.4	2		50	
		10		-150
0.5	3		-25	
		0		
0.8	3			

$$p_3(x) = 2 + 50(x - 0.3)(x - 0.4) - 150(x - 0.3)(x - 0.4)(x - 0.5).$$

A aproximação para $f(0.6)$ *é*

$$f(0.6) \approx p_3(0.6) = 4.1.$$

b) Tem de se acrescentar à tabela anterior o ponto mais próximo do ponto interpolador que ainda não tenha sido usado no cálculo do polinómio, por exemplo, $x = 1$*. Poder-se-ia usar também o ponto* $x = 0.2$*, já que está à mesma distância do ponto interpolador.*

x_i	f_i	$dd1$	$dd2$	$dd3$	$dd4$
0.3	2				
		0			
0.4	2		50		
		10		-150	
0.5	3		-25		**297.619047**
		0		58.333333	
0.8	3		10		
		5			
1.0	4				

O erro de truncatura é dado por

$$\begin{aligned}|e_3| &\leq |(x-x_0)(x-x_1)(x-x_2)(x-x_3))| \times |dd4| \\ &= |(0.6-0.3)(0.6-0.4)(0.6-0.5)(0.6-0.8))| \times 297.619047 = 0.357143\end{aligned}$$

Capítulo 5

Interpolação segmentada - 'spline'

Uma 'spline' é uma régua de madeira utilizada para traçar curvas suaves entre dois pontos dados.

Esta técnica surgiu nos anos 40 do século XX na engenharia náutica, para a elaboração de trajetórias de grandes navios. Estas devem ser curvas suaves que passam pelos vários pontos de paragem. É muito usada também na indústria naval para apurar a forma dos cascos a partir de esboços grosseiros e na área da robótica para a definição de trajetórias de movimentos de robôs. Na informática, as 'splines' são a base da gráfica computacional.

A grande vantagem de se usarem 'splines', comparativamente aos polinómios interpoladores, é evitar o ruído que surge quando se usam muitos pontos, pelo facto de o polinómio obtido ser de grau elevado (Figura 5.1).

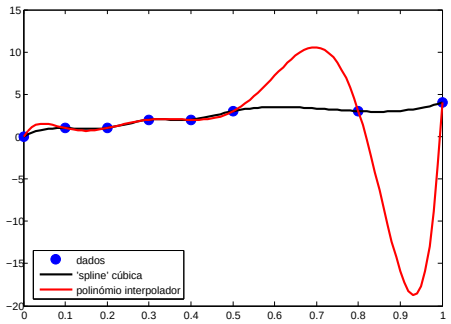


Figura 5.1: 'Spline' cúbica, $s_3(x)$, e polinómio interpolador de grau sete, $p_7(x)$, obtidos para um conjunto de oito pontos.

5.1 Definição

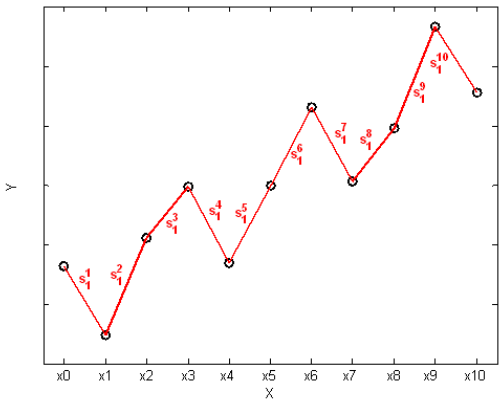
Uma 'spline' é uma função segmentada (definida por segmentos), isto é, é formada por vários polinómios ligados uns aos outros de uma forma contínua e suave.

Dado um conjunto de $n+1$ pontos, $x_0, x_1, \ldots, x_{n-1}, x_n$, chamam-se nós interiores aos pontos $x_1, \ldots, x_{n-1}$, sendo x_0 e x_n os nós exteriores ou fronteiras.

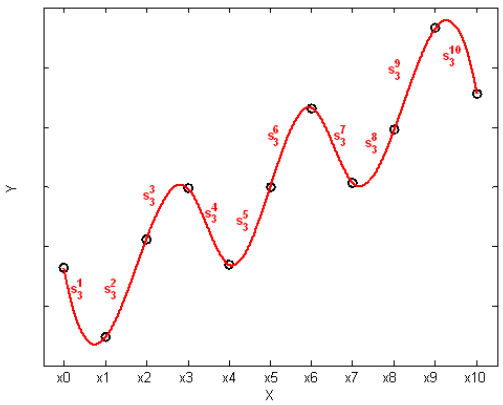
Uma função $s_k(x)$, com k inteiro e não negativo, chama-se 'spline' de grau k se possuir as seguintes propriedades:

- $s_k(x)$ é uma função continuamente diferenciável até à ordem $k-1$;
- $s_k^i(x)$ é um polinómio de grau k, em cada segmento i, para $x \in [x_{i-1}, x_i],\ 1 \leq i \leq n$.

Exemplo 5.1 *'Spline' linear,* $s_1(x)$



Exemplo 5.2 *'Spline' cúbica,* $s_3(x)$



5.2 'Spline' linear

5.2.1 Definição

Uma 'spline' linear é formada pela ligação de polinómios de grau um, em que o segmento i é definido por $[x_{i-1}, x_i]$. Em cada um destes segmentos, o polinómio de grau um obtém-se através de

$$s_1^i(x) = f_{i-1} + \frac{f_i - f_{i-1}}{x_i - x_{i-1}}(x - x_{i-1}),$$

com $i = 1, 2, \ldots, n$ e em que $f_i \equiv f(x_i)$.

5.2.2 Limite superior do erro de truncatura

Seja $f(x)$ contínua, com derivadas contínuas até à segunda ordem. Sejam os pontos do intervalo $[a, b]$ tais que $a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$. Seja ainda $s_1(x)$ a 'spline' linear composta pelos polinómios de grau um $s_1^i(x)$, $i = 1, 2, \ldots, n$, para aproximar $f(x)$ em $[a, b]$. O limite superior do erro de truncatura em valor absoluto cometido por esta aproximação é dado por

$$|f(x) - s_1(x)| \leq \frac{1}{8} h^2 M_2.$$

h é o espaçamento máximo entre os pontos que foram usados para construir a 'spline',

$$h = \max_{0 \leq i \leq n-1} (x_{i+1} - x_i),$$

e M_2 é valor absoluto do majorante da segunda derivada de $f(x)$ em $[a, b]$,

$$\max_{\xi \in [a,b]} \left| f''(\xi) \right| \leq M_2.$$

Se $f(x)$ não for dada por uma expressão matemática, substitui-se M_2 pela diferença dividida de segunda ordem de maior módulo em valor absoluto, multiplicada por 2!.

5.3 'Spline' cúbica

5.3.1 Definição

Uma 'spline' cúbica $s_3(x)$ é formada pela ligação suave dos polinómios de grau três,

$$s_3(x) = \begin{cases} s_3^1(x) & x \in [x_0, x_1] \text{ (para o segmento 1)} \\ s_3^2(x) & x \in [x_1, x_2] \text{ (para o segmento 2)} \\ \vdots & \vdots \\ s_3^n(x) & x \in [x_{n-1}, x_n] \text{ (para o segmento } n). \end{cases}$$

A forma deste polinómio em cada segmento i é dada por

$$\begin{aligned} s_3^i(x) = & \frac{M_{i-1}}{6(x_i - x_{i-1})}(x_i - x)^3 + \frac{M_i}{6(x_i - x_{i-1})}(x - x_{i-1})^3 \\ + & \left[\frac{f_{i-1}}{x_i - x_{i-1}} - \frac{M_{i-1}(x_i - x_{i-1})}{6}\right](x_i - x) \\ + & \left[\frac{f_i}{x_i - x_{i-1}} - \frac{M_i(x_i - x_{i-1})}{6}\right](x - x_{i-1}), \end{aligned} \tag{5.1}$$

em que $i = 1, 2, \ldots, n$, $f_i \equiv f(x_i)$ e $M_i \equiv M(x_i)$ representa a curvatura da 'spline' em x_i, isto é, o valor da segunda derivada da 'spline' no nó x_i.

Para que a ligação entre os vários segmentos seja suave, tem de haver continuidade nos nós e tem de se manter a curvatura. Isto significa que em cada nó interior x_i, $i = 1, 2, \ldots, n-1$ tem de se verificar

- $s_3^i(x_i) = s_3^{i+1}(x_i)$,
- $s_3^{i'}(x_i) = s_3^{i+1'}(x_i)$,
- $s_3^{i''}(x_i) = s_3^{i+1''}(x_i)$.

Pode verificar-se em (5.1) que $s_3^i(x)$ depende de M_{i-1} e M_i, sendo que no primeiro segmento ($i = 1$), é necessário conhecer M_0 e M_1, no segundo segmento ($i = 2$) é necessário conhecer M_1 e M_2, e assim sucessivamente, sendo que para $i = n$ é necessário conhecer M_{n-1} e M_n.

Por se exigir continuidade da primeira derivada nos $n - 1$ nós interiores, para cada um

destes nós resulta a equação

$$\begin{aligned}(x_i - x_{i-1})M_{i-1} + 2(x_{i+1} - x_{i-1})M_i + (x_{i+1} - x_i)M_{i+1} = \\ = \frac{6}{x_{i+1} - x_i}(f_{i+1} - f_i) - \frac{6}{x_i - x_{i-1}}(f_i - f_{i-1}),\end{aligned} \tag{5.2}$$

$i = 1, 2, \ldots, n-1$. Estas $n-1$ equações nas $n+1$ incógnitas M_0, M_1, M_2, M_{n-1}, M_n definem um sistema linear tridiagonal, que deve ser resolvido por EGPP. No entanto, este sistema tem duas incógnitas a mais que o número de equações. Para que o sistema seja possível e determinado, é necessário adicionar mais duas equações ao sistema. Estas vão depender do tipo de 'spline' cúbica.

5.3.2 'Spline' cúbica natural

Se a 'spline' cúbica for natural, a curvatura da 'spline' nos extremos é nula, o que significa que

$$s_3^{1''}(x_0) = 0 \text{ e } s_3^{n''}(x_n) = 0,$$

ou seja

$$M_0 = 0 \text{ e } M_n = 0.$$

Assim, para uma 'spline' cúbica natural, o sistema definido por (5.2) tem $n-1$ equações nas $n-1$ incógnitas $M_1, M_2, \ldots, M_{n-1}$.

5.3.3 'Spline' cúbica completa

Para uma 'spline' cúbica completa, a curvatura nos extremos é não nula, pelo que é necessário acrescentar ao sistema resultante de (5.2) mais duas equações, que dizem respeito ao nó da fronteira inferior x_0 (5.3) e ao nó da fronteira superior x_n (5.4).

$$2(x_1 - x_0)M_0 + (x_1 - x_0)M_1 = \frac{6}{x_1 - x_0}(f_1 - f_0) - 6f'(x_0), \tag{5.3}$$

$$2(x_n - x_{n-1})M_n + (x_n - x_{n-1})M_{n-1} = 6f'(x_n) - \frac{6}{x_n - x_{n-1}}(f_n - f_{n-1}). \tag{5.4}$$

O cálculo destas equações envolve o cálculo das derivadas nos extremos, f'_0 e f'_n. Se $f(x)$ for dada por uma expressão, calcula-se $f'(x)$ e assim $f'_0 \equiv f'(x_0)$ e $f'_n \equiv f'(x_n)$. Se a expressão de

$f(x)$ não for conhecida, as primeiras derivadas $f'(x_0)$ e $f'(x_n)$ têm de ser estimadas recorrendo às diferenças divididas de primeira ordem, ou seja,

$$f'_0 = \frac{f(A) - f_0}{A - x_0} \quad \text{e} \quad f'_n = \frac{f_n - f(B)}{x_n - B}$$

com $x_0 < A$ e $B < x_n$. Quanto mais próximo A estiver de x_0, melhor é a aproximação a $f'(x_0)$. De igual modo, quanto mais próximo B estiver de x_n, melhor é a aproximação a $f'(x_n)$. Em termos práticos, isto significa que A e B são o segundo e o penúltimo pontos, respetivamente, do conjunto de pontos dado. Ao reservar A e B para calcular a aproximação às primeiras derivadas, não é aconselhável incluir os pares $(A, f(A))$ e $(B, F(B))$ na construção da 'spline'.

A 'spline' cúbica completa tem curvatura nos extremos, iniciando-se e terminando de forma suave, ao passo que a 'spline' cúbica natural não tem curvatura nos extremos, iniciando-se e terminando de forma abrupta (Figura 5.2).

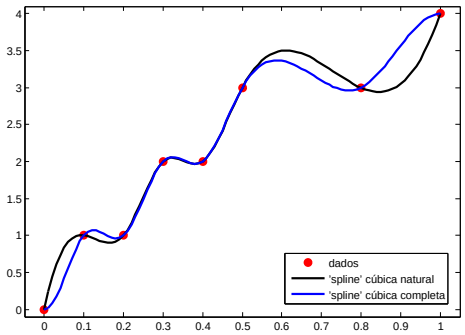


Figura 5.2: 'Spline' cúbica natural e 'spline' cúbica completa para um conjunto de oito pontos.

5.3.4 Limite superior do erro de truncatura

Seja $f(x)$ contínua, com derivadas contínuas até à quarta ordem. Sejam os pontos do intervalo $[a, b]$ tais que $a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$. Seja ainda $s_3(x)$ a 'spline' cúbica composta pelos polinómios de grau três $s_3^i(x)$, $i = 1, 2, \ldots, n$, para aproximar $f(x)$ em $[a, b]$. O limite superior do erro de truncatura em valor absoluto cometido por esta aproximação é dado por

$$|f(x) - s_3(x)| \leq \frac{5}{384} h^4 M_4$$

e

$$|f'(x) - s'_3(x)| \leq \frac{1}{24} h^3 M_4.$$

h é o espaçamento máximo entre os pontos que foram usados para construir a 'spline',

$$h = \max_{0 \leq i \leq n-1} (x_{i+1} - x_i),$$

e M_4 é o valor absoluto do majorante da quarta derivada de $f(x)$ em $[a, b]$,

$$\max_{\xi \in [a,b]} \left| f^{iv}(\xi) \right| \leq M_4.$$

Se $f(x)$ não for dada por uma expressão matemática, substitui-se M_4 pela diferença dividida de quarta ordem de maior módulo em valor absoluto, multiplicada por 4!.

Exemplo 5.3 *Num certo campeonato regional de futebol há 7 equipas. No fim da temporada, o número de pontos ganhos e o número de golos sofridos por 6 das equipas estão representados na tabela*

Equipa	*F.C.Sol*	*F.C.Lá*	*S.C.Gato*	*Nova F.C.*	*Vila F.C.*	*F.C.Chão*
Nº de pontos, x_i	10	12	18	27	30	34
Nº de golos, $f(x_i)$	20	18	15	9	12	10

a) Use uma 'spline' cúbica completa para descrever a relação entre o número de pontos e o número de golos sofridos pelas equipas no campeonato. Sabendo que a 7ª equipa terminou o campeonato com 29 pontos, estime o número de golos que terá sofrido.

b) Calcule uma estimativa do erro de truncatura cometido na alínea anterior.

Resolução:

a) Como a 'spline' cúbica é completa e não é conhecida uma expressão analítica que represente a função, retiram-se o segundo ($A = 12$) e penúltimo ($B = 30$) pontos da tabela, já que serão usados como pontos auxiliares para calcular uma aproximação por diferenças divididas às derivadas nos extremos. Assim,

$$f'_0 = \frac{20 - 18}{10 - 12} = -1 \quad e \quad f'_n = \frac{12 - 10}{30 - 34} = -0.5.$$

Restam 4 pontos $\Rightarrow n = 3 \Rightarrow$ *2 pontos interiores.*

- $i = 0$

$$2(x_1 - x_0)M_0 + (x_1 - x_0)M_1 = \frac{6}{x_1 - x_0}(f_1 - f_0) - 6f'_0 \Leftrightarrow$$
$$\Leftrightarrow 16M_0 + 8M_1 = 2.25$$

- $i = 1$

$$(x_1 - x_0)M_0 + 2(x_2 - x_0)M_1 + (x_2 - x_1)M_2 = \frac{6}{x_2 - x_1}(f_2 - f_1) - \frac{6}{x_1 - x_0}(f_1 - f_0) \Leftrightarrow$$

$$\Leftrightarrow 8M_0 + 34M_1 + 9M_2 = -0.25$$

- $i = 2$

$$(x_2 - x_1)M_1 + 2(x_3 - x_1)M_2 + (x_3 - x_2)M_3 = \frac{6}{x_3 - x_2}(f_3 - f_2) - \frac{6}{x_2 - x_1}(f_2 - f_1) \Leftrightarrow$$

$$\Leftrightarrow 9M_1 + 32M_2 + 7M_3 = 4.8571$$

- $i = 3$

$$2(x_3 - x_2)M_3 + (x_3 - x_2)M_2 = 6f_3' - \frac{6}{x_3 - x_2}(f_3 - f_2) \Leftrightarrow$$

$$\Leftrightarrow 7M_2 + 14M_3 = -3.8571$$

O sistema resultante deve resolver-se por EGPP. A sua solução é

$$\left(\begin{array}{cccc|c} 16 & 8 & 0 & 0 & 2.25 \\ 8 & 34 & 9 & 0 & -0.25 \\ 0 & 9 & 32 & 7 & 4.8571 \\ 0 & 0 & 7 & 14 & -3.8571 \end{array}\right) \longrightarrow \begin{cases} M_0 = 0.2054 \\ M_1 = -0.1295 \\ M_2 = 0.2790 \\ M_3 = -0.4150 \end{cases}$$

O ponto $x = 29$ está no terceiro segmento (entre $x = 27$ e $x = 34$), logo,

$$\begin{aligned} s_3^3(x) &= \frac{M_2}{6(x_3 - x_2)}(x_3 - x)^3 + \frac{M_3}{6(x_3 - x_2)}(x - x_2)^3 + \left(\frac{f_2}{x_3 - x_2} - \frac{M_2(x_3 - x_2)}{6}\right)(x_3 - x) \\ &+ \left(\frac{f_3}{x_3 - x_2} - \frac{M_3(x_3 - x_2)}{6}\right)(x - x_2) \end{aligned}$$

$$s_3^3(x) = 0.0066(34 - x)^3 - 0.0099(x - 27)^3 + 0.9602(34 - x) + 1.9128(x - 27)$$

$$f(29) \approx s_3^3(29) = 9.3779 \approx 9 \text{ golos.}$$

b) Como a função não é conhecida, tem de se construir a tabela das diferenças divididas para calcular uma aproximação ao majorante da quarta derivada.

x_i	f_i	*dd1*	*dd2*	*dd3*	*dd4*
10	*20*				
		-1			
12	*18*		*0.0625*		
		-0.5		*-0.0043*	
18	*15*		*-0.0111*		*0.0006*
		-0.6667		*-0.0083*	
27	*9*		*0.1389*		**-0.0014**
		1		*-0.0221*	
30	*12*		*-0.2143*		
		-0.5			
34	*10*				

$$|f(x) - s_3(x)| \leq \frac{5}{384} h^4 M_4 = \frac{5}{384} \times 9^4 \times 0.0014 \times 4! = 2.8335.$$

Capítulo 6

Integração numérica

6.1 Forma geral do problema

Pretende-se calcular uma aproximação ao integral definido

$$\int_a^b f(x)dx,$$

em que $f(x)$ é a função integranda definida em $[a, b]$ e os limites a e b são finitos.

Este tipo de aproximação é usado quando a primitiva de f não pode vir expressa em termos de funções elementares ou quando a função integranda, ainda que conhecida, é demasiado complicada. Aplica-se ainda quando a função integranda é conhecida apenas para um conjunto discreto de pontos.

Se $p_n(x)$ for uma aproximação polinomial a $f(x)$, então

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

é aproximado por

$$\int_a^b p_n(x)\,dx.$$

Se $e_n(x)$ é o erro da aproximação polinomial, $e_n(x) = f(x) - p_n(x)$, então

$$I = \int_a^b p_n(x)\,dx + \int_a^b e_n(x)\,dx,$$

ou seja,

$$I = \int_a^b p_n(x)\,dx \; + \{\text{erro de integração}\}.$$

6.2 Fórmulas simples de Newton-Cotes

O polinómio de Lagrange de grau menor ou igual a n é dado por

$$p_n(x) = \sum_{j=0}^{n} L_j(x)\, f(x_j)$$

com $f(x_j) = p_n(x_j)$, $j = 0, 1, \ldots, n$, para $n+1$ pontos em $[a, b]$. Se se usar este polinómio $p_n(x)$ para aproximar a função integranda $f(x)$,

$$\int_a^b f(x)dx \approx \int_a^b p_n(x)\, dx = \int_a^b \left(\sum_{j=0}^{n} L_j(x) f(x_j) \right) dx = \sum_{j=0}^{n} \underbrace{\left(\int_a^b L_j(x)dx \right)}_{\omega_j} f(x_j) = \sum_{j=0}^{n} \omega_j\, f(x_j) \tag{6.1}$$

Os coeficientes ω_j dependem da escolha dos pontos x_j, $j = 0, 1, \ldots, n$ através dos polinómios $L_j(x)$ e são independentes dos valores de $f(x)$.

$$\omega_j = \int_a^b L_j(x)\, dx = \int_a^b \frac{(x - x_0) \cdots (x - x_{j-1})(x - x_{j+1}) \cdots (x - x_n)}{(x_j - x_0) \cdots (x_j - x_{j-1})(x_j - x_{j+1}) \cdots (x_j - x_n)} dx.$$

Para um conjunto de pontos igualmente espaçados em $[a, b]$,

$$x_j = a + jh, \quad j = 0, 1, \ldots, n \quad , h = \frac{b - a}{n}.$$

É a partir dos polinómios de Lagrange que são deduzidas as fórmulas ou regras de Newton-Cotes.

6.2.1 Regra do retângulo

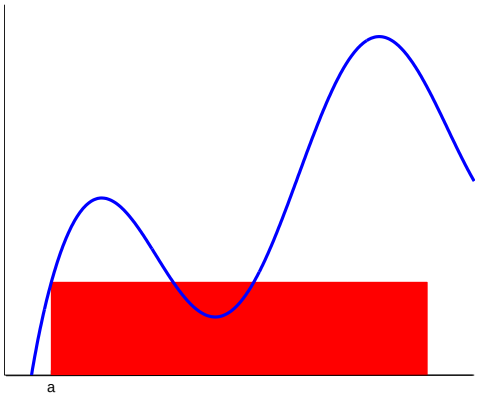


Figura 6.1: Representação gráfica da regra do retângulo.

Na regra do retângulo (Figura 6.1) é usado apenas um ponto dos extremos do intervalo $[a, b]$, logo, o polinómio para aproximar a função integranda é de grau zero.

Por exemplo, usando o ponto a,

$$n = 0,\ x_0 = a,\ L_0 = 1,\ \omega_0 = \int_a^b L_0(x)dx = b - a.$$

De (6.1) vem

$$\int_a^b f(x)dx \approx (b-a)f(a).$$

6.2.2 Regra do ponto médio

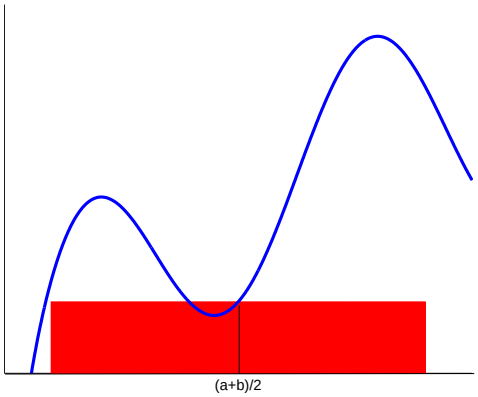


Figura 6.2: Representação gráfica da regra do ponto médio.

Na regra do ponto médio (Figura 6.2) é usado apenas o ponto médio do intervalo $[a, b]$, logo, o polinómio para aproximar a função integranda é também, neste caso, de grau zero.

$$n = 0,\ x_0 = \frac{a+b}{2},\ L_0(x) = 1,\ \omega_0 = \int_a^b L_0(x)dx = b - a.$$

De (6.1) vem

$$\int_a^b f(x)dx \approx (b-a)f\left(\frac{a+b}{2}\right).$$

6.2.3 Regra do trapézio

Na regra do trapézio (Figura 6.3) são usados os dois pontos extremos do intervalo $[a, b]$, logo, o polinómio para aproximar a função integranda é de grau um.

$$n = 1,\ x_0 = a,\ x_1 = b,\ L_0(x) = \frac{x-b}{a-b},\ \omega_0 = \int_a^b L_0(x)dx = \frac{b-a}{2},$$

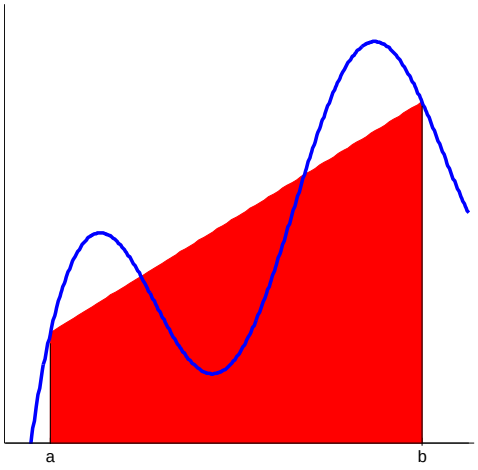


Figura 6.3: Representação gráfica da regra do trapézio.

$$L_1(x) = \frac{x-a}{b-a},\ \omega_1 = \int_a^b L_1(x)dx = \frac{b-a}{2}.$$

De (6.1) vem

$$\int_a^b f(x)dx \approx \frac{b-a}{2}\left[f(a)+f(b)\right].$$

6.2.4 Regra de Simpson

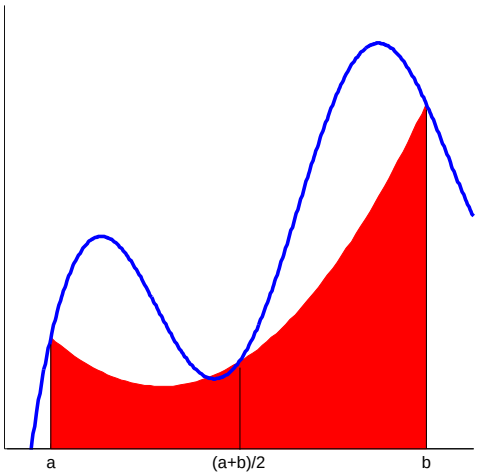


Figura 6.4: Representação gráfica da regra de Simpson.

Na regra de Simpson (Figura 6.4) são usados três pontos do intervalo $[a, b]$ - os dois pontos extremos e o ponto médio -, logo, o polinómio para aproximar a função integranda é de grau dois.

$$n = 2,\ x_0 = a,\ x_1 = \frac{a+b}{2},\ x_2 = b,\ \omega_0 = \omega_2 = \int_a^b \frac{\left(x - \frac{a+b}{2}\right)(x-b)}{\left(a - \frac{a+b}{2}\right)(a-b)}dx = \frac{b-a}{6},$$

$$\omega_1 = \int_a^b \frac{(x-a)(x-b)}{\left(\frac{a+b}{2} - a\right)\left(\frac{a+b}{2} - b\right)}dx = \frac{4(b-a)}{6}.$$

De (6.1) vem

$$\int_a^b f(x)dx \approx \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right].$$

6.2.5 Regra dos três oitavos

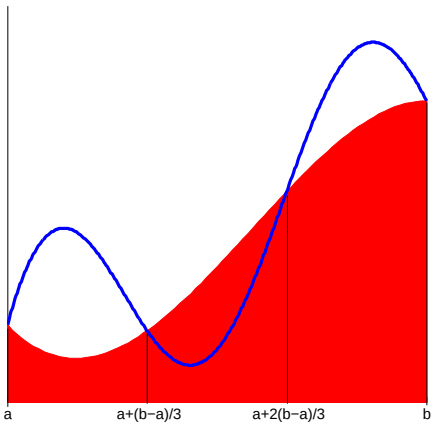


Figura 6.5: Representação gráfica da regra dos três oitavos.

Na regra dos três oitavos (Figura 6.5) são usados os quatro pontos do intervalo $[a, b]$, logo, o polinómio para aproximar a função integranda é de grau três.

$$n = 3,\ x_0 = a,\ x_1 = a + \frac{b-a}{3},\ x_2 = a + 2\frac{b-a}{3},\ x_3 = b,\ \omega_0 = \omega_3 = \frac{b-a}{8},\ \omega_1 = \omega_2 = \frac{3(b-a)}{8}.$$

De (6.1) vem

$$\int_a^b f(x)dx \approx \frac{b-a}{8}\left[f(a) + 3f\left(\frac{2a+b}{3}\right) + 3f\left(\frac{a+2b}{3}\right) + f(b)\right].$$

6.2.6 Erros de truncatura

Como já se pôde perceber anteriormente, as fórmulas de integração numérica, em concreto, as fórmulas de Newton-Cotes, têm um erro de truncatura associado, resultante da discretização do integral.

Se a derivada de ordem $n+1$ de $f(x)$ for contínua em $[a,b]$ e se os x_j, $j = 0, 1, \ldots, n$ pertencem ao intervalo $[a,b]$, então existe um ponto $\eta \in [a,b]$ tal que o erro de aproximar $f(x)$ pelo polinómio $p_n(x)$ é

$$e_n(x) = \prod_{j=0}^{n}(x - x_j)\frac{f^{(n+1)}(\eta)}{(n+1)!}, \quad \eta \in [a,b]$$

e o erro de integração é dado por

$$E(x) = \frac{1}{(n+1)!}\int_a^b \prod_{j=0}^{n}(x - x_j) f^{(n+1)}(\eta)dx, \quad \eta \in [a,b]. \tag{6.2}$$

Esta fórmula pode ser simplificada, usando o teorema do valor médio (Teorema 6.2.1).

Teorema 6.2.1 *Teorema do valor médio*

Se duas funções $f(x)$ e $g(x)$ são contínuas e se, além disso, g não muda de sinal no intervalo $[a,b]$, então existe um ponto $\xi \in [a,b]$ tal que

$$\int_a^b f(x)g(x)dx = f(\xi)\int_a^b g(x)dx.$$

Aplicando o Teorema 6.2.1 a (6.2) vem

$$E(x) = \frac{f^{(n+1)}(\eta)}{(n+1)!}\int_a^b \prod_{j=0}^{n}(x - x_j)dx, \quad \eta \in [a,b].$$

No entanto, nem sempre é possível aplicar o teorema do valor médio. Nas regras do ponto médio e de Simpson, $g(x)$ muda de sinal em $[a,b]$, no ponto $\dfrac{a+b}{2}$. Neste caso, os erros são deduzidos a partir da forma geral do resto (6.2). Assim, tem-se

- Erro de truncatura da regra do retângulo

$$e_R = \frac{(b-a)^2}{2}f'(\eta), \ \eta \in [a,b];$$

- Erro de truncatura da regra do ponto médio

$$e_M = \frac{(b-a)^3}{24}f''(\eta), \ \eta \in [a,b];$$

- Erro de truncatura da regra do trapézio

$$e_T = -\frac{(b-a)^3}{12}f''(\eta), \ \eta \in [a,b];$$

- Erro de truncatura da regra de Simpson

$$e_S = -\frac{(b-a)^5}{32}\frac{1}{90}f^{(iv)}(\eta), \eta \in [a,b];$$

- Erro de truncatura da regra dos três oitavos

$$e_{3/8} = -\frac{(b-a)^5}{6480}f^{(iv)}(\eta), \eta \in [a,b].$$

O comportamento do erro depende da derivada de ordem $n+1$ da função, dependendo também do valor selecionado para n. Se o resultado obtido por uma fórmula simples de Newton-Cotes não é satisfatório, isto é, se o erro de truncatura for muito grande, é possível aumentar o valor de n, o que equivale a aumentar o número de pontos e o grau do polinómio que aproxima $f(x)$. No entanto, nem sempre é verdade que quanto maior for n maior seja a precisão do resultado numérico, uma vez que para polinómios de grau elevado os erros da aproximação podem ser também grandes. A forma mais evidente de reduzir o erro de integração seria diminuir o valor de $(b-a)$, mas este intervalo é fixo! A solução será dividir o intervalo fixo $[a,b]$ em subintervalos, que podem ser tão pequenos quanto o necessário, e em cada um desses subintervalos, $f(x)$ é aproximada por um polinómio de grau baixo. Surgem assim as **fórmulas compostas**.

6.3 Fórmulas compostas

Tome-se como exemplo o intervalo $[a,b]$ subdividido em 6 subintervalos.
Para $a < x_1 < x_2 < \cdots < x_5 < b$:

$$\int_a^b f(x)dx = \int_a^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \int_{x_2}^{x_3} f(x)dx + \int_{x_3}^{x_4} f(x)dx + \int_{x_4}^{x_5} f(x)dx + \int_{x_5}^{b} f(x)dx$$

6.3.1 Fórmula composta do trapézio

A fórmula composta do trapézio pode deduzir-se da correspondente regra simples dividindo o intervalo $[a,b]$ em n subintervalos iguais de comprimento $h = \dfrac{b-a}{n}$ (Figura 6.6),

$$\int_a^b f(x)dx = \sum_{j=0}^{n-1}\left\{\int_{x_j}^{x_{j+1}} f(x)dx\right\}$$

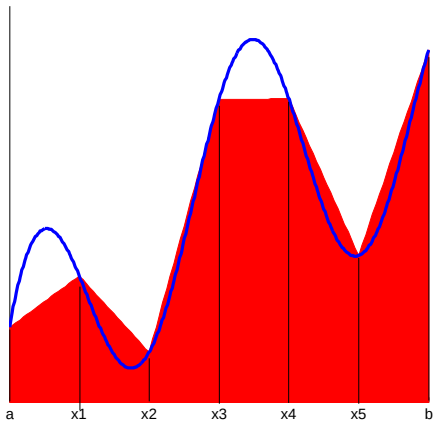


Figura 6.6: Representação gráfica da fórmula composta do trapézio para $n = 6$ subintervalos.

com $x_j = a + jh$ e $j = 0, 1, \ldots, n$. Usando a regra do trapézio em cada subintervalo $[x_j, x_{j+1}]$:

$$\int_{x_j}^{x_{j+1}} f(x)dx \approx \frac{x_{j+1} - x_j}{2} \left[f(x_j) + f(x_{j+1})\right] = \frac{1}{2}h \left[f_j + f_{j+1}\right],$$

com erro

$$e_{T_j} = -\frac{h^3}{12} f''(\eta_j), \text{ com } \eta_j \in [x_j, x_{j+1}].$$

Somando as contribuições dos n subintervalos:

$$\begin{aligned}
\int_a^b f(x)dx &= \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \cdots + \int_{x_{n-2}}^{x_{n-1}} f(x)dx + \int_{x_{n-1}}^{x_n} f(x)dx \\
&\approx \sum_{j=0}^{n-1} \left\{ \frac{1}{2}h \left[f_j + f_{j+1}\right] \right\} = \frac{1}{2}h \left[f_0 + f_1 + f_1 + f_2 \right. \\
&+ \left. f_2 + f_3 + \cdots + f_{n-2} + f_{n-1} + f_{n-1} + f_n \right].
\end{aligned}$$

Assim,

$$T(h) = \frac{h}{2} \left[f_0 + 2f_1 + 2f_2 + \cdots + 2f_{n-1} + f_n\right].$$

Somando também os erros:

$$
\begin{aligned}
e_{CT} &= \sum_{j=0}^{n-1} \left\{ -\frac{h^3}{12} f''(\eta_j) \right\} \\
&= -\frac{h^3}{12} \left\{ f''(\eta_0) + f''(\eta_1) + \cdots + f''(\eta_{n-1}) \right\} \\
&= -\frac{h^2}{12} \frac{(b-a)}{n} \, n \, f''(\eta) \\
&= -\frac{h^2}{12} (b-a) f''(\eta) \\
e_{CT} &= -\frac{h^2}{12} (b-a) f''(\eta), \quad \eta \in [a,b].
\end{aligned}
$$

6.3.2 Fórmula composta de Simpson

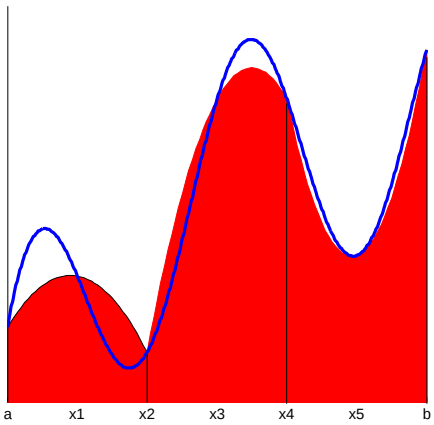


Figura 6.7: Representação gráfica da fórmula composta de Simpson para $n = 6$ subintervalos com $m = 3$ contribuições.

A fórmula composta de Simpson pode deduzir-se da correspondente regra simples. Cada contribuição necessita de dois subintervalos. Assim, $n = 2m$. O espaçamento entre pontos é $h = \dfrac{b-a}{n} = \dfrac{b-a}{2m}$ e existem m contribuições. Em cada par de subintervalos usa-se a regra de Simpson (Figura 6.7):

$$
\int_{x_{j-1}}^{x_{j+1}} f(x)dx \approx \frac{x_{j+1} - x_{j-1}}{6} \left[f_{j-1} + 4f_j + f_{j+1} \right] = \frac{2h}{6} \left[f_{j-1} + 4f_j + f_{j+1} \right]
$$

com erro

$$e_{S_j} = -\frac{(2h)^5}{32 \times 90} f^{(iv)}(\eta_j), \text{ com } \eta_j \in [x_{j-1}, x_{j+1}].$$

Somando as m contribuições:

$$\begin{aligned}\int_a^b f(x)dx &= \sum_{m \text{ termos}} \int_{x_{j-1}}^{x_{j+1}} f(x)dx \approx \sum_{j=1(2 \text{ em } 2)}^{n-1} \left\{ \frac{h}{3} \left[f_{j-1} + 4f_j + f_{j+1} \right] \right\} \\ &= \frac{h}{3} \left[f_0 + 4f_1 + f_2 + f_2 + 4f_3 + f_4 + f_4 + 4f_5 + f_6 + \cdots + f_{n-2} + 4f_{n-1} + f_n \right].\end{aligned}$$

Assim,

$$S(h) = \frac{h}{3} \left[f_0 + 4f_1 + 2f_2 + 4f_3 + 2f_4 + \cdots + 2f_{n-2} + 4f_{n-1} + f_n \right].$$

Somando também os erros:

$$\begin{aligned}e_{CS} &= \sum_{m \text{ termos}} \left\{ -\frac{2^5 h^5}{32 \times 90} f^{(iv)}(\eta_j) \right\} \\ &= -\frac{h^5}{90} \left\{ f^{(iv)}(\eta_1) + f^{(iv)}(\eta_3) + f^{(iv)}(\eta_5) + \cdots + f^{(iv)}(\eta_{n-1}) \right\} \\ &= -\frac{h^4}{90} \frac{(b-a)}{2m} m f^{(iv)}(\eta) \\ &= -\frac{h^4}{180}(b-a) f^{(iv)}(\eta)\end{aligned}$$

$$e_{CS} = -\frac{h^4}{180}(b-a) f^{(iv)}(\eta), \quad \eta \in [a, b].$$

6.3.3 Fórmula composta dos três oitavos

A fórmula composta dos três oitavos pode ser deduzida através da correspondente regra simples, mas, neste caso, cada contribuição necessita de três subintervalos. Assim, $n = 3r$. O espaçamento entre pontos é $h = \dfrac{b-a}{n} = \dfrac{b-a}{3r}$ e existem r contribuições. Em cada conjunto de 3 subintervalos usa-se a regra dos três oitavos (Figura 6.8):

$$\int_{x_{j-1}}^{x_{j+2}} f(x)dx \approx \frac{x_{j+2} - x_{j-1}}{8} \left[f_{j-1} + 3f_j + 3f_{j+1} + f_{j+2} \right] = \frac{3h}{8} \left[f_{j-1} + 3f_j + 3f_{j+1} + f_{j+2} \right],$$

com erro

$$e_{3/8_j} = -\frac{(3h)^5}{6480} f^{(iv)}(\eta_j), \text{ com } \eta_j \in [x_{j-1}, x_{j+2}].$$

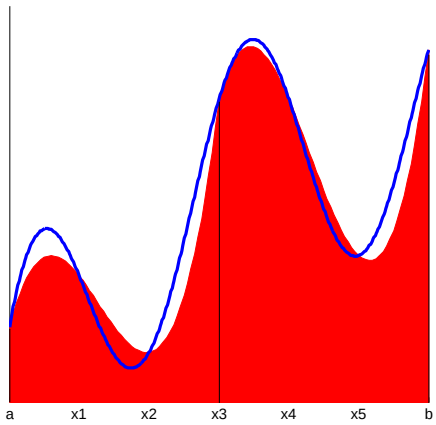


Figura 6.8: Representação gráfica da fórmula composta dos três oitavos para $n = 6$ subintervalos com $r = 2$ contribuições.

Somando todas as r contribuições:

$$\int_a^b f(x)dx = \sum_{r \text{ termos}} \int_{x_{j-1}}^{x_{j+2}} f(x)dx \approx \sum_{j=1(3 \text{ em } 3)}^{n-2} \left\{ \frac{3h}{8} \left[f_{j-1} + 3f_j + 3f_{j+1} + f_{j+2} \right] \right\}$$

$$= \frac{3h}{8} \left[f_0 + 3f_1 + 3f_2 + f_3 + f_3 + 3f_4 + 3f_5 + f_6 + \cdots + f_{n-3} + 3f_{n-2} + 3f_{n-1} + f_n \right].$$

Assim,

$$3/8(h) = \frac{3h}{8} \left[f_0 + 3f_1 + 3f_2 + 2f_3 + 3f_4 + \cdots + 2f_{n-3} + 3f_{n-2} + 3f_{n-1} + f_n \right].$$

Somando os erros:

$$e_{C3/8} = \sum_{r \text{ termos}} \left\{ -\frac{(3h)^5}{6480} f^{(iv)}(\eta_j) \right\}$$

$$= -\frac{(3h)^5}{6480} \left\{ f^{(iv)}(\eta_1) + f^{(iv)}(\eta_4) + f^{(iv)}(\eta_7) + \cdots + f^{(iv)}(\eta_{n-1}) \right\}$$

$$= -\frac{81 \times 3\, h^4}{6480} \frac{(b-a)}{3r} r\, f^{(iv)}(\eta)$$

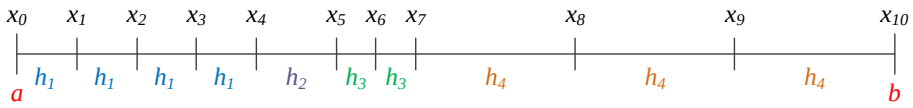
$$= -\frac{h^4}{80}(b-a) f^{(iv)}(\eta)$$

$$e_{C3/8} = -\frac{h^4}{80}(b-a) f^{(iv)}(\eta), \quad \eta \in [a,b].$$

6.4 Aplicação das fórmulas de integração a intervalos de amplitudes diferentes

Como se verificou pelo descrito anteriormente, as fórmulas compostas de integração só podem aplicar-se a intervalos de amplitude constante. Surge então a questão do se deve fazer no caso de haver intervalos cujo espaçamento entre pontos não é constante. A resposta consiste em agrupar subintervalos com amplitudes iguais e aplicar a cada grupo uma fórmula de integração. Atente-se no Exemplo 6.1.

Exemplo 6.1 *Considere-se o seguinte conjunto de 11 pontos no intervalo* $[a, b]$.



$$\int_a^b f(x)dx = \int_{x_0}^{x_{10}} f(x)dx = \int_{x_0}^{x_4} f(x)dx + \int_{x_4}^{x_5} f(x)dx + \int_{x_5}^{x_7} f(x)dx + \int_{x_7}^{x_{10}} f(x)dx$$

por exemplo

$$\approx \underbrace{S(h_1)}_{n=4} + \underbrace{T(h_2)}_{n=1} + \underbrace{S(h_3)}_{n=2} + \underbrace{3/8(h_4)}_{n=3}$$

6.5 Escolha da melhor fórmula

O integral deve ser sempre calculado de forma a que se garanta que o erro de truncatura cometido é o menor possível. Tendo presente que cada caso tem de ser analisado de acordo com as suas particularidades, podem descrever-se algumas regras gerais, que a seguir se apresentam.

- condição de aplicabilidade

  se n (número de subintervalos)

  - é par e múltiplo de 3 (ex. 6, 12, ...) pode usar-se Simpson, 3 oitavos e trapézio;
  - é par e não é múltiplo de 3 (ex. 4, 8, 10, ...) pode usar-se Simpson e trapézio;
  - é múltiplo de 3 e não é par (ex. 9, 15, ...) pode usar-se 3 oitavos e trapézio;
  - não é par nem múltiplo de 3 (ex. 5, 11, ...) pode usar-se trapézio.

- entre Simpson e 3 oitavos, Simpson tem sempre um erro menor;
- entre trapézio e Simpson, ou trapézio e 3 oitavos, deve analisar-se:
  - h^2 (trapézio) *vs* h^4 (Simpson e 3 oitavos);
  - coeficientes em valor absoluto

$$\frac{1}{12} \text{ (trapézio); } \frac{1}{180} \text{ (Simpson); } \frac{1}{80} \text{ (3 oitavos)}$$

  - M_2 *vs* M_4 com

$$\left|f''_{[a,b]}\right| \leq M_2 \text{ (majorante) (trapézio)}$$
$$\left|f^{(iv)}_{[a,b]}\right| \leq M_4 \text{ (majorante) (Simpson e 3 oitavos).}$$

Exemplo 6.2 *Na tabela seguinte são apresentados registos pontuais das vendas de um produto que foi lançado no início do ano de 2009. A variável x representa a semana (de 2009).*

x_i	1	2	3	4	5	7	9	11	13	15	16	17	18	19
$v(x_i)$	10	9	8	8	8	6	5	5	4	4	4	4	3	1

a) Calcule a melhor aproximação ao integral $\int_1^{19} v(x)dx$*, com base em toda a informação fornecida na tabela sobre* $v(x)$*.*

b) Estime o erro de truncatura cometido com a aproximação obtida na alínea anterior no intervalo $[5, 15]$*.*

Resolução:

a)

$$\int_1^{19} v(x)dx = \underbrace{\int_1^5 v(x)dx}_{h=1,\, n=4,\, \boldsymbol{S}} + \underbrace{\int_5^{15} v(x)dx}_{h=2,\, n=5,\, \boldsymbol{T}} + \underbrace{\int_{15}^{19} v(x)dx}_{h=1,\, n=4,\, \boldsymbol{S}}$$

$$\approx \frac{1}{3}(10+4\times 9+2\times 8+4\times 8+8)+\frac{2}{2}(8+2\times 6+2\times 5+2\times 5+2\times 4+4)+$$
$$+\frac{1}{3}(4+4\times 4+2\times 4+4\times 3+1)$$

$$= 32+52+13.666667=99.666667$$

b) **Trapézio - [5,15]**

x_i	v_i	*dd1*	*dd2*
5	*8*		
		-1	
7	*6*		**0.125**
		-0.5	
9	*5*		**0.125**
		0	
11	*5*		**-0.125**
		-0.5	
13	*4*		**0.125**
		0	
15	*4*		

$$|e_T| = \frac{2^2}{12}(15-5)\times 0.125\times 2! = 0.833333.$$

Capítulo 7

Aproximação dos mínimos quadrados

7.1 Forma geral do problema

Pretende-se, com esta aproximação, definir um modelo, $M(x; c_i)$, dado por uma expressão matemática, que se ajuste o melhor possível à função dada $f(x)$, definida no intervalo $[a, b]$. Usando a técnica dos mínimos quadrados, pretende-se minimizar a soma dos quadrados dos erros (Figura 7.1). Podem ter-se dois tipos de problema distintos:

- O problema discreto, em que são dados m pontos $x_1 < x_2 < \ldots < x_m$ no intervalo $[a, b]$ (Figura 7.1(a)) e o objetivo é

$$\text{minimizar} \sum_{j=1}^{m} (f(x_j) - M(x_j; c_i))^2.$$

- O problema contínuo, em que é dada uma função $f(x)$ (Figura 7.1(b)) e o objetivo é

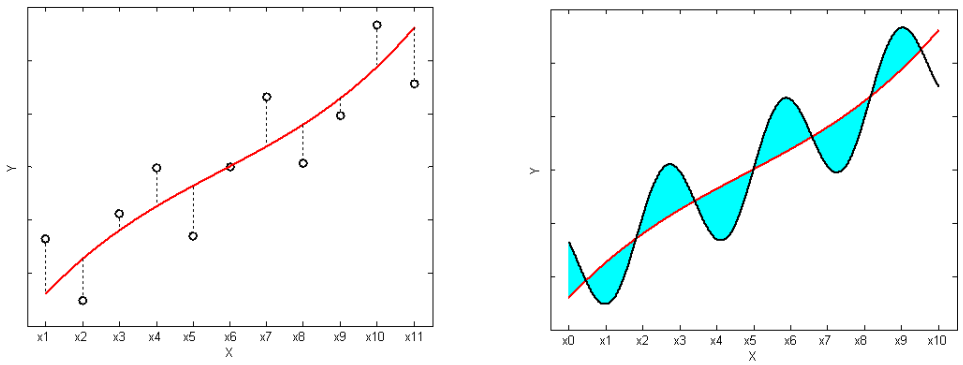
$$\text{minimizar} \int_a^b (f(x) - M(x; c_i))^2 dx.$$

Neste capítulo vai abordar-se apenas o problema discreto, ou seja, a função f é dada por um conjunto discreto de valores $(x_1, f_1), (x_2, f_2), \ldots, (x_m, f_m)$.

Os modelos de mínimos quadrados, dependendo da sua forma, podem classificar-se em

- Modelo linear e polinomial, que define um problema de mínimos quadrados linear e a sua expressão é dada por um polinómio. Por exemplo,

$$M(x; a_0, a_1, a_2) \equiv p_2(x) = a_0 + a_1 x + a_2 x^2.$$



(a) problema discreto (b) problema contínuo

Figura 7.1: Problema dos mínimos quadrados

- Modelo linear e não polinomial, que define um problema de mínimos quadrados linear, mas não é um polinómio. Por exemplo,

$$M(x; c_1, c_2) = c_1 e^{x^2} + c_2 \sin(x).$$

- Modelo não linear, que define um problema de mínimos quadrados não linear. Por exemplo,

$$M(x; c_1, c_2) = \ln(c_1 x^2) + e^{c_2 x}.$$

  O modelo não linear não vai ser abordado neste texto.

7.2 Modelo polinomial - polinómios ortogonais

Pretende-se construir um modelo definido por um polinómio completo de grau n, $p_n(x)$. Para que o problema seja bem definido, deve verificar-se a condição única $m \geq n+1$, sendo m o número de pontos onde a função é conhecida e n o grau do polinómio. No caso particular em que $m = n+1$ está-se perante o polinómio interpolador, já que o polinómio que passa por um dado conjunto de pontos é único e por isso $\sum_{j=1}^{m} (f_j - p_1(x_j))^2 = 0$.

Para que o problema seja bem condicionado, isto é, não seja sensível a erros nos dados ou erros de arredondamento nos cálculos, o polinómio $p_n(x)$ deve ser construído usando uma sequência de polinómios ortogonais, $P_0(x), P_1(x), \ldots, P_n(x)$, na forma

$$p_n(x) = c_0 P_0(x) + c_1 P_1(x) + c_2 P_2(x) + \cdots + c_n P_n(x). \tag{7.1}$$

7.2.1 Polinómios ortogonais

A propriedade que define os polinómios ortogonais é

$$\sum_{i=1}^{m} P_j(x_i)P_k(x_i) \begin{cases} = 0, & \text{se} \quad j \neq k \\ \neq 0, & \text{se} \quad j = k. \end{cases}$$

A partir dos pontos dados (x_j, f_j), $j = 1, 2, \ldots, m$ determina-se a sequência de polinómios ortogonais $P_0(x), \ldots, P_n(x)$ e os coeficientes $c_0, \ldots, c_n$ para construir o polinómio completo $p_n(x)$ através de (7.1).

passo 1 Constroem-se os polinómios ortogonais, $P_0(x), P_1(x), P_2(x), \ldots, P_n(x)$, da sequência de polinómios ortogonais, usando a relação de recorrência (7.2).

$$P_{i+1}(x) = (x - B_i)\, P_i(x) \; - \; \mathbb{C}_i \, P_{i-1}(x), \qquad \text{para} \;\; i = 0, 1, \ldots, n-1 \tag{7.2}$$

em que

$$P_{-1}(x) = 0 \text{ e } P_0(x) = 1,$$

$$B_i = \frac{\displaystyle\sum_{j=1}^{m} x_j \, P_i^2(x_j)}{\displaystyle\sum_{j=1}^{m} P_i^2(x_j)}, \quad \text{para todo o} \quad i$$

$$\mathbb{C}_0 = 0 \quad \text{e} \quad \mathbb{C}_i = \frac{\displaystyle\sum_{j=1}^{m} P_i^2(x_j)}{\displaystyle\sum_{j=1}^{m} P_{i-1}^2(x_j)} \quad \text{para} \quad i > 0.$$

passo 2 Calculam-se os coeficientes do polinómio, $c_0, c_1, c_2, \ldots, c_n$, usando (7.3).

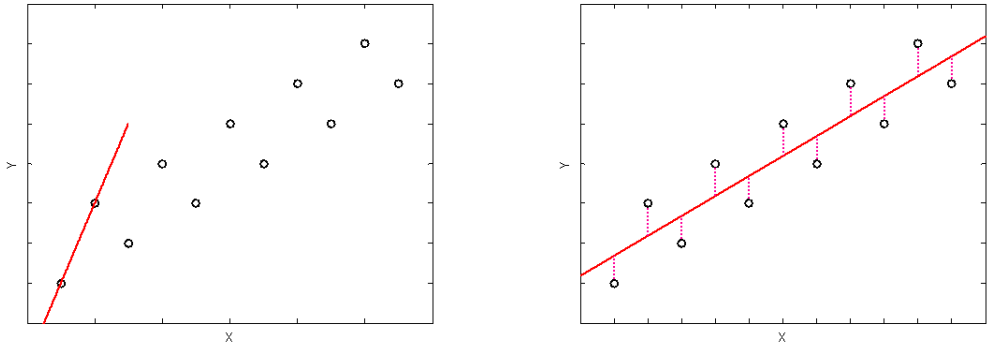
$$c_i = \frac{\displaystyle\sum_{j=1}^{m} f_j \, P_i(x_j)}{\displaystyle\sum_{j=1}^{m} P_i^2(x_j)}, \quad i = 0, 1, \ldots, n. \tag{7.3}$$

passo 3 Forma-se o polinómio pretendido.

$$p_n(x) = c_0 P_0(x) + c_1 P_1(x) + c_2 P_2(x) + \cdots + c_n P_n(x).$$

Exemplo 7.1 *Construir um polinómio de grau 1, $p_1(x)$ (Figura 7.2)*

Se o número de pontos for $m = 2$, o polinómio resultante é o polinómio interpolador, por isso passa nos pontos (Figura 7.2(a)). Se o número de pontos $m > 2$, $p_1(x)$ é o polinómio que melhor se ajusta à "mancha" de pontos (Figura 7.2(b)).

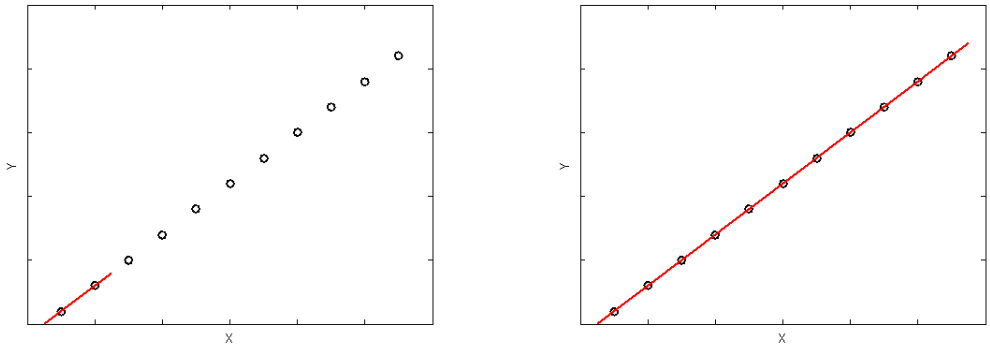


(a) $m = 2$ (b) $m > 2$

Figura 7.2: Polinómio de grau 1, $p_1(x)$

Exemplo 7.2 *Construir um polinómio de grau 1, $p_1(x)$ (Figura 7.3)*

Neste caso todos os polinómios $p_1(x)$ são iguais e passam nos pontos de $f(x)$ porque os pontos pertencem a um polinómio de grau 1, o que significa que $\sum_{j=1}^{m}(f_j - p_1(x_j))^2 = 0$.



(a) $m = 2$ (b) $m > 2$

Figura 7.3: Polinómio de grau 1, $p_1(x)$

7.3 Modelo linear não polinomial

O modelo linear não polinomial tem a forma

$$M(x; c_1, \ldots, c_n) = c_1\Phi_1(x) + c_2\Phi_2(x) + \cdots + c_n\Phi_n(x).$$

Este modelo é linear nos coeficientes, $c_1, c_2, \ldots, c_n$ e $\Phi_1(x), \Phi_2(x), \ldots, \Phi_n(x)$ são funções. No caso deste modelo, a única coisa a determinar são os coeficientes, uma vez que as funções Φ_i, $i = 1, \ldots, n$ são dadas. O número de termos na definição do modelo caracteriza a dimensão do problema, n, que é também o número de coeficientes a determinar.

A condição única para que este problema seja bem definido é que o número de pontos onde a função é definida seja maior ou igual ao número de parâmetros do modelo, isto é, $m \geq n$. Se $m = n$, o modelo passa em todos os pontos da função e por isso $\sum_{j=1}^{m}(f_j - p_1(x_j))^2 = 0$.

O cálculo dos coeficientes $c_1, c_2, \ldots, c_n$ é feito a partir do sistema das equações normais.

7.3.1 Sistema das equações normais

Seja o modelo

$$M(x; c_1, \ldots, c_n) = c_1\Phi_1(x) + c_2\Phi_2(x) + \cdots + c_n\Phi_n(x).$$

No sentido dos mínimos quadrados, o objetivo é encontrar o modelo $M(x; c_1, \ldots, c_n)$ resultante de

$$\underset{c_1,\ldots,c_n}{\text{minimizar}} \;\; S(c_1, c_2, \ldots, c_n) \equiv \sum_{j=1}^{m} \left(f_j - M(x_j; c_1, \ldots, c_n)\right)^2,$$

isto é,

$$\underset{c_1,\ldots,c_n}{\text{minimizar}} \sum_{j=1}^{m} \left(f_j - (c_1\Phi_1(x_j) + c_2\Phi_2(x_j) + \cdots + c_n\Phi_n(x_j))\right)^2.$$

Como se pretende calcular $c_1, c_2, \ldots, c_n$ de forma a que $S(c_1, \ldots, c_n)$ seja mínimo, usa-se cálculo diferencial, ou seja, deriva-se $S(c_1, \ldots, c_n)$ em ordem a cada um dos coeficientes do modelo e igualam-se essas derivadas parciais a zero.

$$\begin{cases} \dfrac{\partial S}{\partial c_1} = -2\displaystyle\sum_{j=1}^{m} \left(f_j - c_1\Phi_1(x_j) - c_2\Phi_2(x_j) - \cdots - c_n\Phi_n(x_j)\right)\Phi_1(x_j) = 0 \\ \dfrac{\partial S}{\partial c_2} = -2\displaystyle\sum_{j=1}^{m} \left(f_j - c_1\Phi_1(x_j) - c_2\Phi_2(x_j) - \cdots - c_n\Phi_n(x_j)\right)\Phi_2(x_j) = 0 \\ \cdots \\ \dfrac{\partial S}{\partial c_n} = -2\displaystyle\sum_{j=1}^{m} \left(f_j - c_1\Phi_1(x_j) - c_2\Phi_2(x_j) - \cdots - c_n\Phi_n(x_j)\right)\Phi_n(x_j) = 0, \end{cases}$$

ou seja

$$\begin{cases} \sum_{j=1}^{m} f_j\Phi_1(x_j) - \sum_{j=1}^{m} c_1\Phi_1(x_j)\Phi_1(x_j) - \cdots - \sum_{j=1}^{m} c_n\Phi_n(x_j)\Phi_1(x_j) = 0 \\ \sum_{j=1}^{m} f_j\Phi_2(x_j) - \sum_{j=1}^{m} c_1\Phi_1(x_j)\Phi_2(x_j) - \cdots - \sum_{j=1}^{m} c_n\Phi_n(x_j)\Phi_2(x_j) = 0 \\ \cdots \\ \sum_{j=1}^{m} f_j\Phi_n(x_j) - \sum_{j=1}^{m} c_1\Phi_1(x_j)\Phi_n(x_j) - \cdots - \sum_{j=1}^{m} c_n\Phi_n(x_j)\Phi_n(x_j) = 0, \end{cases}$$

ou ainda

$$\begin{cases} c_1 \sum_{j=1}^{m} \Phi_1^2(x_j) + c_2 \sum_{j=1}^{m} \Phi_2(x_j)\Phi_1(x_j) + \cdots + c_n \sum_{j=1}^{m} \Phi_n(x_j)\Phi_1(x_j) = \sum_{j=1}^{m} f_j\Phi_1(x_j) \\ c_1 \sum_{j=1}^{m} \Phi_1(x_j)\Phi_2(x_j) + c_2 \sum_{j=1}^{m} \Phi_2^2(x_j) + \cdots + c_n \sum_{j=1}^{m} \Phi_n(x_j)\Phi_2(x_j) = \sum_{j=1}^{m} f_j\Phi_2(x_j) \\ \cdots \\ c_1 \sum_{j=1}^{m} \Phi_1(x_j)\Phi_n(x_j) + c_2 \sum_{j=1}^{m} \Phi_2(x_j)\Phi_n(x_j) + \cdots + c_n \sum_{j=1}^{m} \Phi_n^2(x_j) = \sum_{j=1}^{m} f_j\Phi_n(x_j). \end{cases}$$

Este sistema $n \times n$ é linear nos coeficientes a determinar, $c_1, \ldots, c_n$. Na forma matricial,

$$\begin{pmatrix} \sum_{j=1}^{m} \Phi_1^2(x_j) & \sum_{j=1}^{m} \Phi_2(x_j)\Phi_1(x_j) & \cdots & \sum_{j=1}^{m} \Phi_n(x_j)\Phi_1(x_j) \\ \sum_{j=1}^{m} \Phi_1(x_j)\Phi_2(x_j) & \sum_{j=1}^{m} \Phi_2^2(x_j) & \cdots & \sum_{j=1}^{m} \Phi_n(x_j)\Phi_2(x_j) \\ \cdots & \cdots & \ddots & \cdots \\ \sum_{j=1}^{m} \Phi_1(x_j)\Phi_n(x_j) & \sum_{j=1}^{m} \Phi_2(x_j)\Phi_n(x_j) & \cdots & \sum_{j=1}^{m} \Phi_n^2(x_j) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \cdots \\ c_n \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{m} f_j\Phi_1(x_j) \\ \sum_{j=1}^{m} f_j\Phi_2(x_j) \\ \cdots \\ \sum_{j=1}^{m} f_j\Phi_n(x_j) \end{pmatrix}.$$

A resolução do sistema linear das equações normais fornece os coeficientes pretendidos $c_1, \ldots, c_n$ e deve ser feita por um método direto e estável, por exemplo, a eliminação de Gauss com pivotagem parcial.

Para calcular o modelo na forma

$$M(x; c_1, \ldots, c_n) = c_1\Phi_1(x) + c_2\Phi_2(x) + \cdots + c_n\Phi_n(x)$$

devem seguir-se os passos seguintes.

passo 1 Identificar

a) n, ou seja, o número de termos ou coeficientes, que caracterizam a dimensão do sistema.

b) as n funções $\Phi_1(x)$, $\Phi_2(x)$, $\cdots$, $\Phi_n(x)$.

passo 2 Formar o sistema das equações normais nas n equações e n incógnitas $c_1, \ldots, c_n$, na forma matricial

$$\begin{pmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \cdots \\ c_n \end{pmatrix} = \begin{pmatrix} \cdots \\ \cdots \\ \cdots \\ \cdots \end{pmatrix}$$

passo 3 Resolver este sistema por EGPP.

passo 4 Formar o modelo pretendido

$$M(x; c_1, \ldots, c_n) = c_1\,\Phi_1(x) + c_2\,\Phi_2(x) + \cdots + c_n\,\Phi_n(x).$$

Exemplo 7.3 *A resistência de um certo fio (de uma certa substância), $f(x)$, varia com o diâmetro desse fio, x. A partir de uma experiência registaram-se os seguintes valores:*

x_j	*1.5*	*2.0*	*3.0*	*4.0*
f_j	*4.9*	*3.3*	*2.0*	*1.5*

Foram sugeridos os seguintes modelos para ajustar os valores de $f(x)$, no sentido dos mínimos quadrados:

i. uma reta

ii. o modelo linear $M(x, c_1, c_2) = \dfrac{c_1}{x} + c_2 x$

a) Calcule a reta.

b) Calcule o modelo $M(x)$.

c) Qual dos modelos escolheria? Justifique a sua resposta.

Resolução:

a) $p_1(x) = c_0P_0(x) + c_1P_1(x)$

	x_i	f_i	$P_1(x_i)$	$P_1(x_i)^2$	$f_iP_1(x_i)$
	1.5	*4.9*	*-1.125*	*1.265625*	*-5.5125*
	2	*3.3*	*-0.625*	*0.390625*	*-2.0625*
	3	*2*	*0.375*	*0.140625*	*0.75*
	4	*1.5*	*1.375*	*1.890625*	*2.0625*
$\sum$	**10.5**	**11.7**		**3.6875**	**-4.7625**

- $P_0(x) = 1, \quad C_0 = 0, \quad P_{-1}(x) = 0$
- $P_1(x) = x - B_0$

  $B_0 = \dfrac{\sum xP_0(x)^2}{\sum P_0(x)^2} = \dfrac{10.5}{4} = 2.625$

  $P_1(x) = x - 2.625$
- $c_0 = \dfrac{\sum fP_0(x)}{\sum P_0(x)^2} = \dfrac{11.7}{4} = 2.925$
- $c_1 = \dfrac{\sum fP_1(x)}{\sum P_1(x)^2} = \dfrac{-4.7625}{3.6875} = -1.291525$

$$p_1(x) = 2.925 - 1.291525(x - 2.625)$$

b) $M(x; c_1, c_2) = \dfrac{c_1}{x} + c_2x$

$$\begin{cases} \Phi_1(x) = \dfrac{1}{x} \\ \Phi_2(x) = x \end{cases}$$

	x_i	f_i	$\Phi_1(x_i)$	$\Phi_2(x_i)$	$\Phi_1(x_i)^2$	$\Phi_2(x_i)^2$	$\Phi_1(x_i)\Phi_2(x_i)$	$f_i\Phi_1(x_i)$	$f_i\Phi_2(x_i)$
	1.5	*4.9*	*0.666667*	*1.5*	*0.444444*	*2.25*	*1*	*3.266667*	*7.35*
	2	*3.3*	*0.5*	*2*	*0.25*	*4*	*1*	*1.65*	*6.6*
	3	*2*	*0.333333*	*3*	*0.111111*	*9*	*1*	*0.666667*	*6*
	4	*1.5*	*0.25*	*4*	*0.0625*	*16*	*1*	*0.375*	*6*
$\sum$					**0.868055**	**31.25**	**4**	**5.958334**	**25.95**

$$\begin{pmatrix} \sum_i \Phi_1(x_i)^2 & \sum_i \Phi_1(x_i)\Phi_2(x_i) & | & \sum_i f_i\Phi_1(x_i) \\ \sum_i \Phi_1(x_i)\Phi_2(x_i) & \sum_i \Phi_2(x_i)^2 & | & \sum_i f_i\Phi_2(x_i) \end{pmatrix} \longrightarrow$$

$$\begin{pmatrix} 0.868055 & 4 & | & 5.958334 \\ 4 & 31.25 & | & 25.95 \end{pmatrix} \longrightarrow \begin{cases} c_1 & = 7.405414 \\ c_2 & = -0.117493 \end{cases}$$

$$M(x) = \frac{7.405414}{x} - 0.117493x$$

c) Cálculo da soma dos quadrados dos resíduos

	x_i	f_i	$p_1(x_i)$	$M(x_i)$	$(f_i - p_1(x_i))^2$	$(f_i - M(x_i))^2$
	1.5	*4.9*	*4.377966*	*4.760703*	*0.27252*	*0.019404*
	2	*3.3*	*3.732203*	*3.467721*	*0.1868*	*0.02813*
	3	*2*	*2.440678*	*2.115992*	*0.194197*	*0.013454*
	4	*1.5*	*1.149153*	*1.381382*	*0.123094*	*0.01407*
$\sum$					**0.776611**	**0.075058**

O modelo $M(x)$ ajusta-se melhor no sentido dos mínimos quadrados porque a soma dos quadrados dos resíduos é menor que para o modelo $p_1(x)$ ($0.075058 < 0.776611$).

Capítulo 8

Otimização não linear sem restrições

A otimização surge no processo de tomada de decisão para se atingir o melhor resultado possível. Dificilmente se consegue imaginar uma área de estudo em que os princípios da otimização não estejam presentes. Na realidade, mesmo no dia a dia, tudo se tenta otimizar - pretende-se sempre o mínimo ou o máximo de algo. É, pois, um dos objetivos dos profissionais das áreas das Ciências de Gestão e Engenharia, estando também presente noutras áreas aplicadas, tais como a economia, as finanças, a medicina ou a estatística.

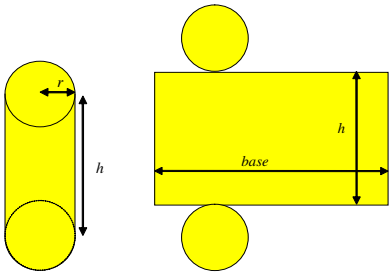
A otimização está relacionada com a maximização ou minimização de modelos matemáticos, e a função que se pretende otimizar é denominada *função objetivo*. No entanto, por vezes as variáveis de decisão estão sujeitas a determinadas condições, designadas por *restrições*.

Os problemas de otimização podem ser classificados de acordo com as características da função objetivo e das restrições, em duas grandes áreas: a Otimização Linear e a Otimização Não Linear (ONL). No primeiro caso, todas as funções (função objetivo e restrições) envolvidas no problema são lineares. Na ONL, pelo menos uma das funções, entre a função objetivo e as restrições, é não linear nas variáveis. Ainda na otimização não linear, há alguns casos particulares mais fáceis de resolver:

- problemas sem restrições nas variáveis;
- problemas quadráticos;
- problemas convexos;
- problemas de mínimos quadrados lineares.

Exemplo 8.1 *Problema com duas variáveis e uma restrição*

Tendo como objetivo fabricar latas cilíndricas com um volume de 1000 cm^3 e tapá-las em ambas as extremidades, qual deverá ser o raio da base e a altura da lata de modo a minimizar a quantidade de placa metálica, em termos de área superficial?



$$\begin{aligned} \textit{Área Total} &= \textit{Área}_{retângulo} + 2 \times \textit{Área}_{círculo} \\ &= base \times h + 2\left(\pi r^2\right) \\ &= \textit{Perímetro}_{círculo} \times h + 2\pi r^2 \\ &= 2\pi rh + 2\pi r^2 \\ \textit{Volume} &= \pi r^2 \times h \\ 1000 &= \pi r^2 \times h \end{aligned}$$

Formulação do problema:

$$\begin{aligned} \text{minimizar} \quad & A(r,h) \equiv 2\pi rh + 2\pi r^2 \\ \text{sujeito a} \quad & \pi r^2 h = 1000 \end{aligned}$$

Exemplo 8.2 *Problema com três variáveis e uma restrição*

O produto de três números positivos é igual a A (dado). Determine esses números por forma que a sua soma seja máxima.

$$\begin{aligned} \text{maximizar} \quad & x_1 + x_2 + x_3 \\ \text{sujeito a} \quad & x_1 x_2 x_3 = A \end{aligned} \tag{8.1}$$

8.1 Forma geral do problema

Um problema de otimização sem restrições, em termos gerais, pode ser definido da seguinte forma:

$$\min_{x \in \mathbb{R}^n} f(x) \tag{8.2}$$

em que $f(x)$ é a função que se pretende minimizar.

Num problema de otimização, pretende-se minimizar ou maximizar um objetivo $f(x)$ - função objetivo do problema.

Num problema de minimização pretende-se calcular um ponto $x^* \in \mathbb{R}^n$, denominado minimizante, que verifique $f(x^*) \leq f(x)$ para todo o $x \in \mathbb{R}^n$.

Num problema de maximização pretende-se calcular $x^* \in \mathbb{R}^n$, denominado maximizante, que verifique $f(x^*) \geq f(x)$ para todo o $x \in \mathbb{R}^n$.

8.2 Classificação de mínimos e máximos

Seja $V(x, \delta)$ uma vizinhança de x^* de raio δ $(\delta > 0)$. x^* é

- minimizante local forte se $\exists \delta > 0$:
  - $f(x)$ é definida em $V(x^*, \delta)$
  - $f(x^*) < \ f(x),\ \forall x \in V(x^*, \delta);\ x \neq x^*$
- minimizante local fraco se $\exists \delta > 0$:
  - $f(x)$ é definida em $V(x^*, \delta)$
  - $f(x^*) \leq \ f(x),\ \forall x \in V(x^*, \delta);\ x \neq x^*$
- maximizante local forte se $\exists \delta > 0$:
  - $f(x)$ é definida em $V(x^*, \delta)$
  - $f(x^*) > \ f(x)\ \ \forall x \in V(x^*, \delta);\ \ x \neq x^*$
- maximizante local fraco se $\exists \delta > 0$:
  - $f(x)$ é definida em $V(x^*, \delta)$
  - $f(x^*) \geq \ f(x)\ \ \forall x \in V(x^*, \delta);\ \ x \neq x^*$
- minimizante global forte se $f(x^*) < f(x)$, para todo o x que pertence ao domínio de $f(x)$, onde a função é definida.

- minimizante global fraco se $f(x^*) \leq f(x)$, para todo o x que pertence ao domínio de $f(x)$, onde a função é definida.
- maximizante global forte se $f(x^*) > f(x)$, para todo o x que pertence ao domínio de $f(x)$, onde a função é definida.
- maximizante global fraco se $f(x^*) \geq f(x)$, para todo o x que pertence ao domínio de $f(x)$, onde a função é definida.

Todo o ótimo global é local. No entanto, um ótimo local pode não ser global.

Exemplo 8.3 *Alguns mínimos e máximos de uma função unidimensional*

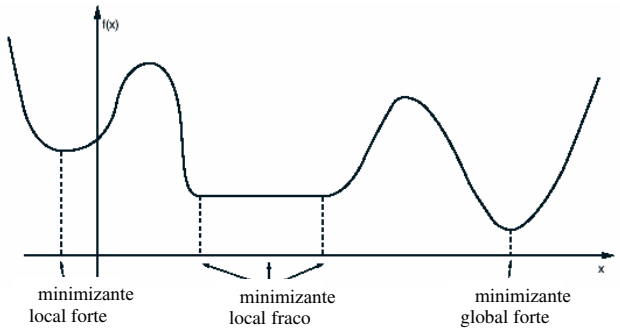


Figura 8.1: Representação gráfica de alguns minimizantes

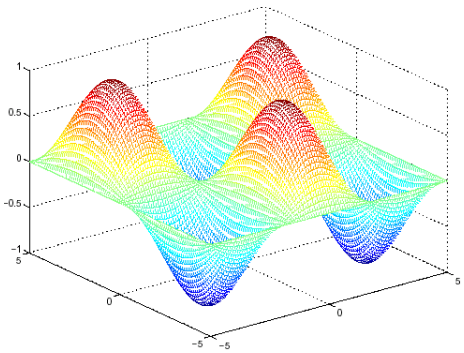
Exemplo 8.4 *Classificação de mínimos e máximos em algumas funções bi-dimensionais*

8.3 Mínimos *versus* máximos

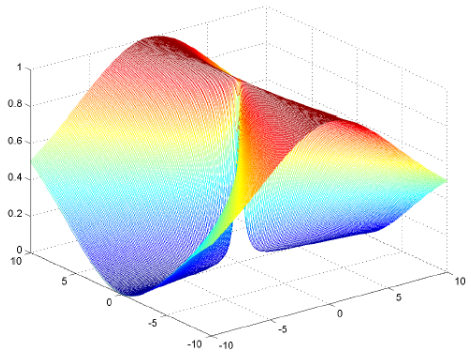
Em geral, os métodos de otimização estão formulados para o cálculo de mínimos. No entanto, podem também ser usados no cálculo de máximos, já que estes podem ser facilmente relacionados com os mínimos da seguinte forma (Figura 8.3):

$$\max f(x) = -\min(-f(x))$$

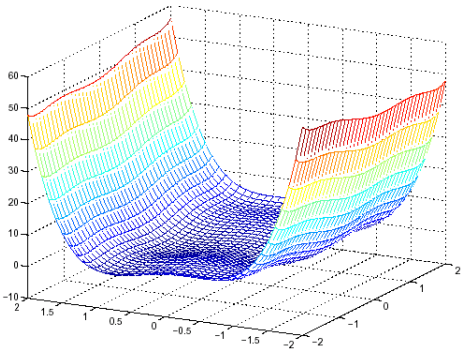
$$x^* = \underbrace{\arg\max\left(f\left(x\right)\right)}_{\text{maximizante}} = \underbrace{\arg\min\left(-f\left(x\right)\right)}_{\text{minimizante}}$$



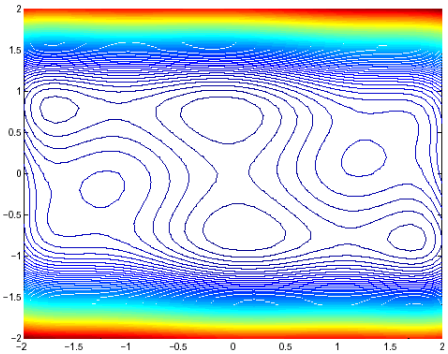
(a) máximos e mínimos fortes



(b) máximos e mínimos fracos



(c) dois mínimos globais e quatro mínimos locais



(d) dois mínimos globais e quatro mínimos locais

Figura 8.2: Classificação de mínimos e máximos

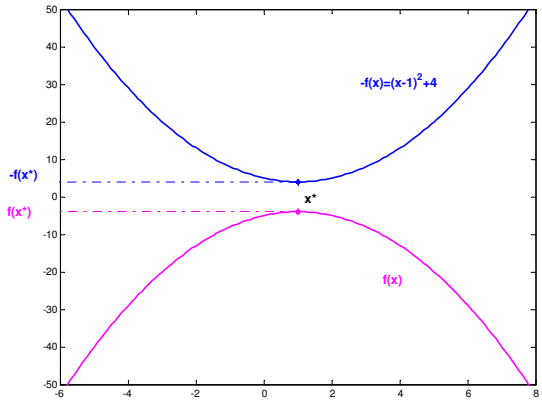


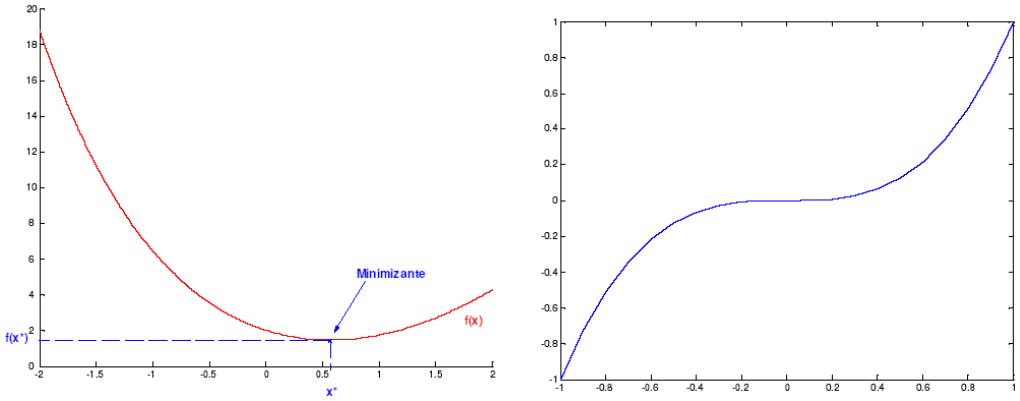
Figura 8.3: Relação entre mínimos e máximos de funções.

Capítulo 9

Otimização unidimensional

Se $n = 1$ (8.2), então está-se perante um problema unidimensional, isto é, com uma só variável, o que significa que x é escalar.

Exemplo 9.5 *Problemas unidimensionais*



(a) $f(x) = x^2 + 2e^{-x}$ (b) $f(x) = x^3$

Figura 9.1: (a) problema com um mínimo; (b) problema sem mínimos (nem máximos)

9.1 Condições de otimalidade

Para aplicar as condições de otimalidade de primeira e segunda ordem, assume-se $f(x)$ continuamente diferenciável até à segunda ordem.

Condição necessária e suficiente de primeira ordem

Se x^* é uma solução do problema (8.2), com $n = 1$, então $f'(x) = 0$. Isto é, a equação $f'(x) = 0$ define os pontos estacionários da função objetivo $f(x)$. Os pontos estacionários podem ser minimizantes, maximizantes ou pontos de inflexão.

Exemplo 9.6 *Determinar os pontos estacionários da função* $f(x) = x^2 + 2e^{-x}$.

$$f'(x) \equiv 2x - 2e^{-x} = 0.$$

A solução desta equação não linear, obtida pelo método iterativo de Newton ou da secante, é única: $x = 0.5671$. *Ou seja,* $x = 0.567143$ *é um ponto estacionário de* $f(x)$.

Condição necessária de segunda ordem

Se x^* é uma solução do problema (8.2) para $n = 1$, que satisfaz a condição de primeira ordem, então a condição necessária para que x^* seja minimizante é $f''(x) \geq 0$ e a condição necessária para que x^* seja maximizante é $f''(x) \leq 0$.

Condição suficiente de segunda ordem

Se x^* é solução do problema (8.2) e se $f''(x) > 0$, então x^* é um minimizante local forte de $f(x)$. Se x^* é solução do problema (8.2) e se $f''(x) < 0$, então x^* é um maximizante local forte de $f(x)$.

Capítulo 10

Método de DSC

10.1 Introdução

Em termos gerais, os métodos de resolução de problemas unidimensionais podem dividir-se em três grandes grupos:

- Métodos de procura ou pesquisa direta
- Métodos de aproximação
- Métodos mistos

Os métodos mistos combinam as técnicas de procura e de aproximação.

O método de Davies, Swann e Campey (DSC) é um método iterativo que só usa informação da função objetivo f e destina-se a problemas de otimização unidimensional. Trata-se de um método misto, isto é, tem uma fase de procura seguida de uma fase de aproximação baseada em interpolação quadrática.

10.2 Fase de procura

Procuram-se, em cada iteração, três pontos igualmente espaçados que definem um intervalo que contém o minimizante da função. Esta procura baseia-se apenas nos valores da função objetivo em diversos pontos.

A procura inicia-se com uma aproximação inicial x_1 e uma perturbação $\delta > 0$. A partir de x_1 e no sentido positivo, calcula-se uma sequência de pontos x_2, x_3, $x_4, \ldots$ distanciados uns

dos outros de δ, 2δ, 4δ, $8\delta, \ldots$. Assim,

$$\begin{aligned} & x_1 \\ & x_2 = x_1 + \delta \\ & x_3 = x_2 + 2\delta \\ & \ldots \\ & x_k = x_{k-1} + 2^{k-2}\delta \end{aligned}$$

até que no ponto x_k se tenha $f(x_k) > f(x_{k-1})$. Nesta altura tem-se $\cdots < x_{k-2} < x_{k-1} < x_k$,

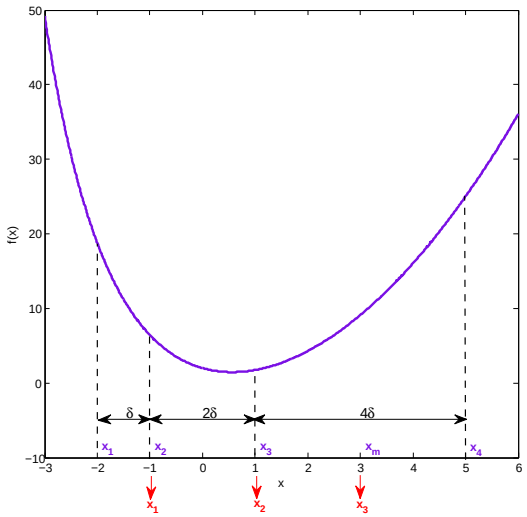


Figura 10.1: Procura do método de DSC para $x_1 = -2$ e $\delta = 1$.

em que $f(x_{k-2}) \geq f(x_{k-1})$ e $f(x_{k-1}) < f(x_k)$. A distância entre x_k e x_{k-1} é duas vezes a distância entre x_{k-1} e x_{k-2}. Para que os pontos estejam igualmente espaçados, calcula-se o ponto médio do último intervalo, $x_m = \dfrac{x_k + x_{k-1}}{2}$. Fica-se, assim, com quatro pontos igualmente espaçados: $x_{k-2} < x_{k-1} < x_m < x_k$. Para a aproximação quadrática, é necessário selecionar três dos quatro pontos encontrados na fase de procura. Para isso, comparam-se os valores de $f(x)$ nos dois pontos interiores do intervalo. Se $f(x_{k-1}) \leq f(x_m)$, então escolhem-se os pontos x_{k-2}, x_{k-1} e x_m, caso contrário escolhem-se os pontos x_{k-1}, x_m e x_k. Ver Figura 10.1.

Quando a partir de x_1 o valor de $f(x_2) > f(x_1)$, com $x_2 = x_1 + \delta$, a procura deve voltar-se para o sentido negativo, a começar novamente por x_1. Neste caso, o próximo ponto na procura é $x_{-1} = x_1 - \delta$. Se $f(x_{-1}) > f(x_1)$, significa que o intervalo definido por $[x_{-1}, x_2]$, com x_1 como ponto médio, contém o minimizante da quadrática que passa pelos três pontos agora

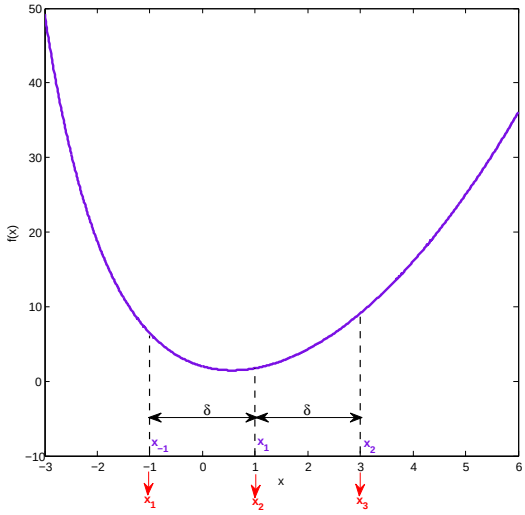


Figura 10.2: Procura do método de DSC para $x_1 = 1$ e $\delta = 2$.

calculados. Ver Figura 10.2.

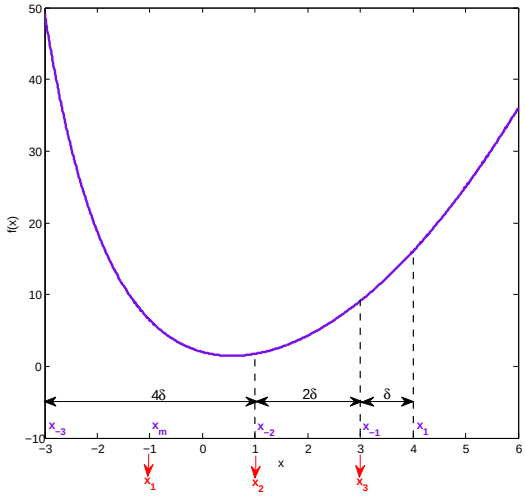


Figura 10.3: Procura do método de DSC para $x_1 = 4$ e $\delta = 1$.

No entanto, se $f(x_{-1}) < f(x_1)$, significa que a procura deve continuar no sentido negativo até que $f(x_{-k}) > f(x_{-(k-1)})$, isto é, procede-se da seguinte forma:

$$x_{-2} = x_{-1} - 2\delta$$
$$\dots$$
$$x_{-k} = x_{-(k-1)} - 2^{k-1}\delta$$

até que no ponto x_{-k} se tenha $f\left(x_{-k}\right) > f\left(x_{-(k-1)}\right)$.

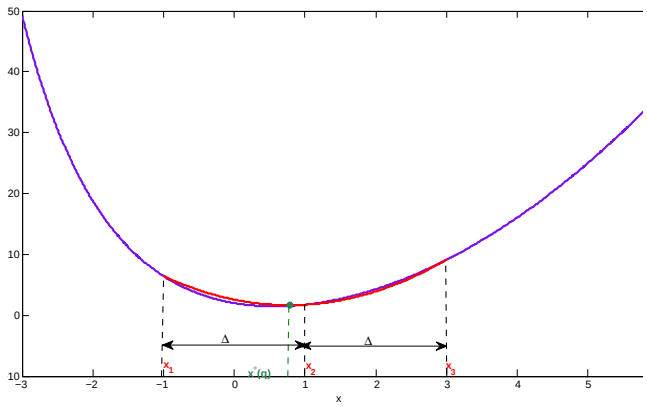


Figura 10.4: Aproximação do método de DSC, com $\Delta = 2$.

Nesta altura, tem-se

$$x_{-k} < x_{-(k-1)} < x_{-(k-2)} < \cdots$$

em que $f\left(x_{-(k-2)}\right) \geq f\left(x_{-(k-1)}\right)$ e $f\left(x_{-(k-1)}\right) < f\left(x_{-k}\right)$, e a distância entre x_{-k} e $x_{-(k-1)}$ é duas vezes a distância entre $x_{-(k-1)}$ e $x_{-(k-2)}$. Para que os pontos estejam igualmente espaçados, calcula-se o ponto médio do último intervalo, $x_m = \dfrac{x_{-k} + x_{-(k-1)}}{2}$. Fica-se, assim, com quatro pontos igualmente espaçados: $x_{-k} < x_m < x_{-(k-1)} < x_{-(k-2)}$. Para a aproximação quadrática, é necessário selecionar três dos quatro pontos encontrados na fase de procura. Para isso, comparam-se os valores de $f(x)$ nos dois pontos interiores do intervalo. Se $f(x_m) \leq f(x_{-(k-1)})$, então escolhem-se os pontos x_{-k}, x_m e $x_{-(k-1)}$, caso contrário escolhem-se os pontos x_m, $x_{-(k-1)}$ e $x_{-(k-2)}$. Ver Figura 10.3.

10.3 Fase de aproximação

Depois de concluída a fase de procura, entra-se na fase de aproximação, em que se aproxima a função no intervalo obtido por uma função quadrática (polinómio de grau dois) e usa-se o seu minimizante como aproximação ao minimizante da função. Esta quadrática passa pelos três pontos selecionados.

O minimizante da quadrática, $x^*(q)$, que passa por estes três pontos, que passam a denominar-se $\mathbf{x}_1 < \mathbf{x}_2 < \mathbf{x}_3$, com $\Delta = (\mathbf{x_2} - \mathbf{x}_1) = (\mathbf{x}_3 - \mathbf{x}_2)$, determina-se por (Figura 10.4)

$$x^*(q) = \mathbf{x}_2 + \Delta \frac{f(\mathbf{x}_1) - f(\mathbf{x}_3)}{2(f(\mathbf{x}_3) - 2f(\mathbf{x}_2) + f(\mathbf{x}_1))}$$

10.4 Paragem do método de DSC

O critério de paragem do método iterativo de DSC consiste em a distância entre os pontos que foram usados para construir a função quadrática não exceder uma certa quantidade positiva e próxima de zero, ou seja

$$(\mathbf{x}_2 - \mathbf{x}_1) = (\mathbf{x}_3 - \mathbf{x}_2) = \Delta \leq \varepsilon, \text{ com } \varepsilon > 0 \text{ e } \varepsilon \approx 0.$$

Se o critério de paragem for verificado, o processo iterativo termina, sendo $x^*(q)$ a melhor aproximação calculada à solução. Se o critério de paragem não se verificar, o processo repete-se e o minimizante da quadrática $x^*(q)$ passa a ser o ponto inicial, x_1, da próxima iteração. A perturbação δ também deve ser atualizada através de $\delta = M\delta$, com $M < 1$.

O Algoritmo 10.1 descreve o método de DSC.

Algoritmo 10.1 Método de Davies, Swann e Campey

ler: x_1, δ, M e ε

repetir

$x_2 \leftarrow x_1 + \delta$

se $f(x_2) \leq f(x_1)$ **então**

$k \leftarrow 2$

repetir

$k \leftarrow k + 1$

$x_k \leftarrow x_{k-1} + 2^{k-2}\delta$

até $f(x_k) > f(x_{k-1})$

$x_m \leftarrow \dfrac{x_k + x_{k-1}}{2}$

se $f(x_{k-1}) \leq f(x_m)$ **então**

$\mathbf{x}_1 \leftarrow x_{k-2}$, $\mathbf{x}_2 \leftarrow x_{k-1}$, $\mathbf{x}_3 \leftarrow x_m$

senão

$\mathbf{x}_1 \leftarrow x_{k-1}$, $\mathbf{x}_2 \leftarrow x_m$, $\mathbf{x}_3 \leftarrow x_k$

fim se

senão

$x_{-1} = x_1 - \delta$

se $f(x_{-1}) < f(x_1)$ **então**

$k \leftarrow 1$

repetir

$k \leftarrow k + 1$

$x_{-k} \leftarrow x_{-(k-1)} - 2^{k-1}\delta$

até $f(x_{-k}) > f(x_{-(k-1)})$

$x_m \leftarrow \dfrac{x_{-k} + x_{-(k-1)}}{2}$

se $f(x_m) < f(x_{-(k-1)})$ **então**

$\mathbf{x}_1 \leftarrow x_{-k}$, $\mathbf{x}_2 \leftarrow x_m$, $\mathbf{x}_3 \leftarrow x_{-(k-1)}$

senão

$\mathbf{x}_1 \leftarrow x_m$, $\mathbf{x}_2 \leftarrow x_{-(k-1)}$, $\mathbf{x}_3 \leftarrow x_{-(k-2)}$

fim se

senão

$\mathbf{x}_3 \leftarrow x_2$, $\mathbf{x}_2 \leftarrow x_1$, $\mathbf{x}_1 \leftarrow x_{-1}$

fim se

fim se

$\Delta \leftarrow (\mathbf{x}_2 - \mathbf{x}_1) = (\mathbf{x}_3 - \mathbf{x}_2)$

$x^*(q) \leftarrow \mathbf{x}_2 + \Delta \dfrac{f(\mathbf{x}_1) - f(\mathbf{x}_3)}{2(f(\mathbf{x}_3) - 2f(\mathbf{x}_2) + f(\mathbf{x}_1))}$

$x_1 \leftarrow x^*(q)$

$\delta = M\delta$

até $\Delta \leq \varepsilon$

$x_{\min} \leftarrow x^*(q)$

Capítulo 11

Otimização multidimensional sem restrições

Um problema multidimensional carateriza-se por envolver mais que uma variável, isto é, $n > 1$ (8.2). Importa, nestes problemas, distinguir os problemas com descontinuidades na função objetivo, uma vez que estes são, em geral, mais difíceis de resolver. Há, no entanto, métodos específicos para esta classe de problemas, conhecida por otimização sem derivadas. Quando os problemas são diferenciáveis, podem ser usados métodos que utilizem informação das derivadas – gradiente e/ou matriz Hessiana – conhecidos por métodos do gradiente.

Exemplo 11.1 *Problemas multidimensionais* $(n = 2)$ *sem restrições*

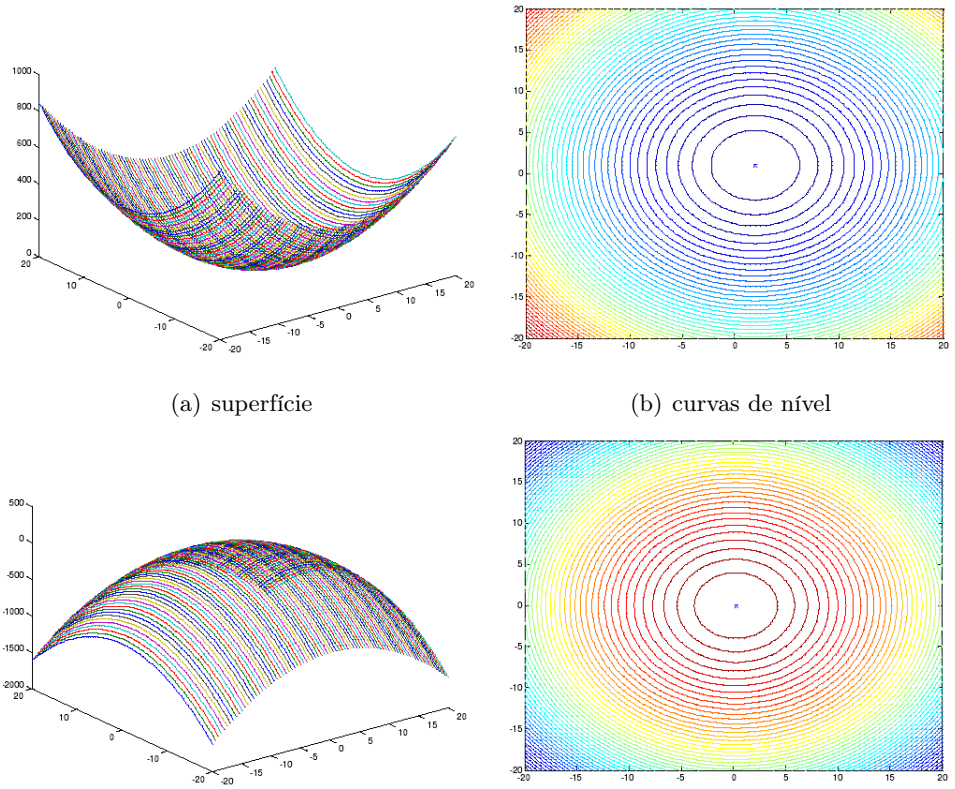
$$\min_{x \in \mathbb{R}^2} \ f(x) \equiv (x_1 - 2)^2 + (x_2 - 1)^2 \quad \text{(Figuras 11.1(a) e 11.1(b))}$$

$$\max_{x \in \mathbb{R}^2} \ f(x) \equiv 2(-x_1^2 - x_2^2 + 1) + x_1 \quad \text{(Figuras 11.1(c) e 11.1(d))}$$

Exemplo 11.2 *Problemas multidimensionais* $(n = 2)$ *sem restrições*

$$\min_{x \in \mathbb{R}^2} \ f(x) \equiv 3x_1^2 - x_2^2 + x_1^3 \quad \text{(Figuras 11.2(a) e 11.2(b))}$$

$$\min_{x \in \mathbb{R}^2} \ f(x) \equiv 3x_1^2 - 4x_1x_2 - 4x_2^2 \quad \text{(Figuras 11.2(c) e 11.2(d))}$$



(a) superfície

(b) curvas de nível

(c) superfície

(d) curvas de nível

Figura 11.1: Problemas com um ótimo

11.1 Notação

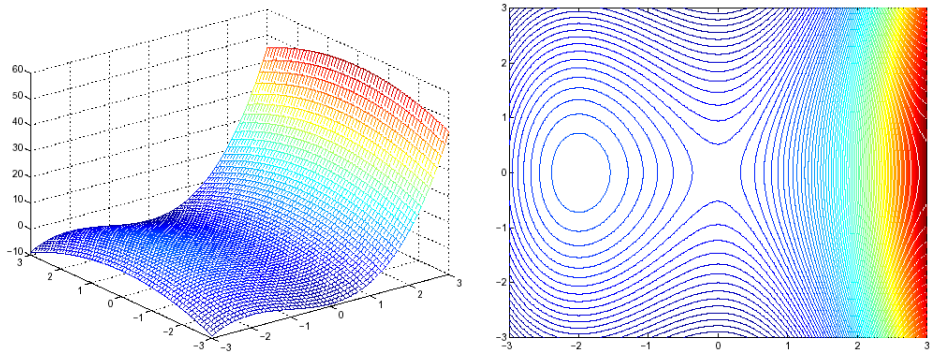
A notação mais usada no âmbito da otimização tem a ver com a definição das primeiras e segundas derivadas da função objetivo.

Seja $f(x)$, $x \in \mathbb{R}^n$ a função objetivo e $x = (x_1, x_2, \ldots, x_n)^T$ um vetor com n componentes.

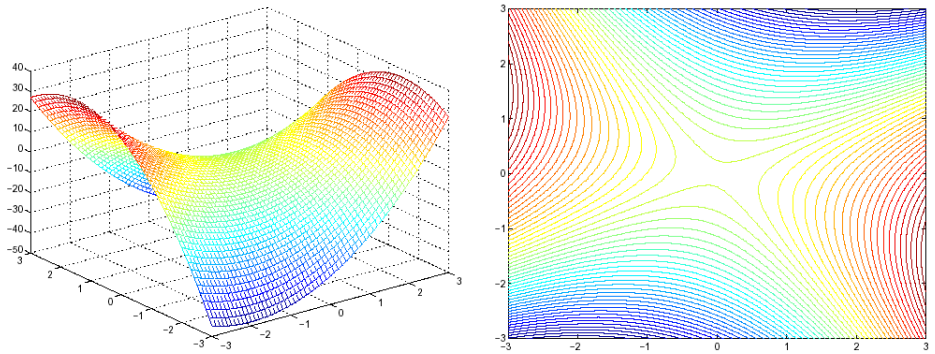
Vetor gradiente de $f(x)$

O vetor gradiente, $\nabla f(x)$, da função $f(x)$, contém as primeiras derivadas parciais de $f(x)$ e é um vetor de dimensão n. Cada i componente do gradiente, $i = 1, \ldots, n$, é dada por $\dfrac{\partial f}{\partial x_i}$. Assim

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}.$$



(a) superfície (b) curvas de nível



(c) superfície (d) curvas de nível

Figura 11.2: Problemas com um ponto sela

Matriz Hessiana de $f(x)$

A matriz Hessiana, $\nabla^2 f(x)$, da função $f(x)$, contém as segundas derivadas parciais de $f(x)$ (ou as primeiras derivadas parciais de cada uma das funções do vetor gradiente) e é uma matriz de dimensão $n \times n$. As ij componentes da Hessiana, $i = 1, \ldots, n$, $j = \ldots, n$ são dadas por $\dfrac{\partial^2 f}{\partial i \partial j}$, para $i \neq j$ e $\dfrac{\partial^2 f}{\partial x_i^2}$, para $i = j$. Assim

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \frac{\partial^2 f}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}.$$

11.2 Condições de otimalidade

Para aplicar as condições de otimalidade de primeira e segunda ordem, assume-se $f(x)$ continuamente diferenciável até à segunda ordem.

Condição necessária e suficiente de primeira ordem

Se x^* é uma solução do problema (8.2), com $n > 1$, então $\nabla f(x^*) = 0$. Isto é, a equação $\nabla f(x^*) = 0$ define os pontos estacionários da função objectivo $f(x)$. Os pontos estacionários podem ser minimizantes (Figuras 11.1(a) e 11.1(a)), maximizantes (Figuras 11.1(c) e 11.1(d)) ou pontos sela (Exemplo 11.2).

Condição necessária de segunda ordem

Se x^* é uma solução do problema (8.2) para $n > 1$, que satisfaz a condição de primeira ordem, então a condição necessária para que x^* seja minimizante é que $\nabla^2 f(x^*)$ seja semi-definida positiva e a condição necessária para que x^* seja maximizante é que $\nabla^2 f(x^*)$ seja semi-definida negativa.

Condição suficiente de segunda ordem

Se x^* é solução do problema (8.2) e $\nabla f(x^*) = 0$, e $\nabla^2 f(x^*) \neq$ matriz nula:

- Se $\nabla^2 f(x^*)$ é definida positiva então x^* é minimizante local forte de $f(x)$;
- Se $\nabla^2 f(x^*)$ é definida negativa então x^* é maximizante local forte de $f(x)$;
- Se $\nabla^2 f(x^*)$ é semi-definida positiva então x^* é minimizante ou ponto sela de $f(x)$;
- Se $\nabla^2 f(x^*)$ é semi-definida negativa então x^* é maximizante ou ponto sela de $f(x)$
- Se $\nabla^2 f(x^*)$ é indefinida então x^* é ponto sela de $f(x)$.

Uma matriz diz-se

- definida positiva se todos os determinantes das submatrizes principais dessa matriz são positivos,

- definida negativa se os determinantes das submatrizes principais dessa matriz são alternadamente negativos e positivos, sendo o determinante da primeira submatriz negativo,
- semi-definida positiva se os determinantes das submatrizes principais dessa matriz são positivos ou iguais a zero,
- semi-definida negativa se os determinantes das submatrizes principais dessa matriz são alternadamente negativos e positivos, sendo o determinante da primeira submatriz negativo, ou iguais a zero,
- indefinida nos restantes casos.

Capítulo 12

Métodos do gradiente

Os métodos do gradiente, tal como o nome indica, usam informação, para além dos valores da função objetivo, das primeiras e/ou segundas derivadas da função (vetor gradiente e matriz Hessiana). Por este motivo, só podem ser usados na resolução de problemas diferenciáveis. Têm a vantagem de convergir mais rapidamente que os métodos de procura direta, que não usam informação sobre as derivadas. Têm, no entanto, a desvantagem do esforço computacional exigido para o cálculo das derivadas.

Estes métodos, que são iterativos, geram uma sucessão de aproximações x^k à solução

$$x^{k+1} = x^k + \alpha^k d^k,$$

em que d^k é a direção de procura ou passo e α^k é o comprimento do passo. d^k é um vetor e α^k é um escalar. A equação iterativa para o cálculo da direção de procura é diferente de acordo com o método que se está a usar.

O Algoritmo 12.2 descreve um método do gradiente.

12.1 Técnicas de globalização

Os métodos do gradiente, quando convergem, convergem para um ponto estacionário, ou seja, um ponto que anula o vetor gradiente de f: $\nabla f(x^*) = 0$. Uma vez que estes métodos exibem convergência local, deve implementar-se uma técnica de globalização de forma a garantir que o método converge, qualquer que seja a aproximação inicial x^1. Significa que x^1 pode estar fora da região de convergência do método. Além disso, garante-se que o ponto estacionário para o qual o método converge é um minimizante.

Há várias técnicas de globalização, entre as quais:

- Procura unidimensional (*line search*)
  - procura unidimensional exata
  - procura unidimensional aproximada
- Região de confiança (*trust region*)
- Filtro

12.2 Procura unidimensional aproximada - critério de Armijo

Na procura unidimensional aproximada pretende-se calcular α^k, o comprimento do passo, dados x^k e d^k, que origina uma redução significativa do valor de f na nova aproximação. Para garantir essa redução pode usar-se a condição de Armijo (12.1).

$$f(x^k + \alpha^k d^k) \leq f(x^k) + \mu \alpha^k \nabla f(x^k)^T d^k, \tag{12.1}$$

com $0 < \mu < \frac{1}{2}$.

Se a direção d^k usada for descendente para f, ou seja, $\nabla f(x^k)^T d^k < 0$, existe um valor de $\alpha^k \in (0, 1]$ que verifica esta condição.

Descreve-se no Algoritmo 12.3 o cálculo α^k usando o critério de Armijo.

12.3 Método de Newton

O método de Newton é um método do gradiente que usa informação sobre as primeiras derivadas da função f – o vetor gradiente – e as segundas derivadas da função f – a matriz Hessiana. Trata-se de um método iterativo, à semelhança dos outros métodos do gradiente, e baseia-se, em cada iteração, numa aproximação local de $f(x)$ a uma função quadrática. Derivando esta função quadrática em ordem a d e igualando o vetor gradiente resultante a zero, define-se a condição de primeira ordem para o mínimo da quadrática, isto é, $\nabla q(d) = 0$. Assim, obtém-se

$$\nabla f(x^k) + \nabla^2 f(x^k) d_N^k = 0 \Leftrightarrow$$

$$\nabla^2 f(x^k) d_N^k = -\nabla f(x^k). \tag{12.2}$$

Algoritmo 12.2 Método do gradiente

ler: x^1 e ε

$k \leftarrow 0$

repetir

$k \leftarrow k + 1$

calcular d^k

calcular α^k

$x^{k+1} \leftarrow x^k + \alpha^k d^k$

até $\left\|\nabla f(x^{k+1})\right\|_2 \leq \varepsilon$

$x^* \leftarrow x^{k+1}$

$f(x^*) \leftarrow f(x^{k+1})$

Algoritmo 12.3 Critério de Armijo

ler: x^k, d^k, $f(x^k)$, $\nabla f(x^k)$ e μ

$\alpha \leftarrow 2$

repetir

$\alpha \leftarrow 0.5 \times \alpha$

$x^{\text{aux}} \leftarrow x^k + \alpha d^k$

até $f(x^{\text{aux}}) \leq f(x^k) + \mu\alpha\nabla f(x^k)^T d^k$

$\alpha^k \leftarrow \alpha$

A solução do sistema linear (12.2), d^k, é a direção de procura. Deve ser usado um método direto e estável, por exemplo, o método de eliminação de Gauss com pivotagem parcial (EGPP), para resolver este sistema.

A nova aproximação $x^k + d_N^k$ não é necessariamente o minimizante de $f(x)$ e por isso o processo deve ser repetido.

12.3.1 Propriedades do método de Newton

O método de Newton tem convergência local, isto é, a convergência para a solução só é garantida se a aproximação inicial, x^1, estiver na vizinhança da solução. Exibe convergência quadrática, ou seja,

$$\left\|x^{k+1} - x^*\right\| \leq \gamma \left\|x^k - x^*\right\|^2, \gamma > 0.$$

O método de Newton possui a propriedade da terminação quadrática, isto é, se $f(x)$, com $x \in \mathbb{R}^n$, for uma função quadrática e convexa, o método de Newton necessita no máximo de n iterações para encontrar a solução exata do problema.

Apesar do método de Newton ter boas propriedades de convergência, tem algumas limitações e desvantagens, que a seguir se descrevem.

12.3.2 Limitações do método de Newton

Em qualquer dos casos a seguir descritos, não é possível obter-se uma direção descendente através do método de Newton.

Direção Newton ascendente

A direção d_N^k, solução do sistema Newton (12.2), pode ser ascendente para f em x^k (Figura 12.1), ou seja,

$$\nabla f(x^k)^T d_N^k > 0.$$

Direção Newton ortogonal ao gradiente

A direção Newton d_N^k, solução do sistema Newton (12.2), pode ser ortogonal ao gradiente em x^k (Figura 12.2), ou seja,

$$\nabla f(x^k)^T d_N^k = 0.$$

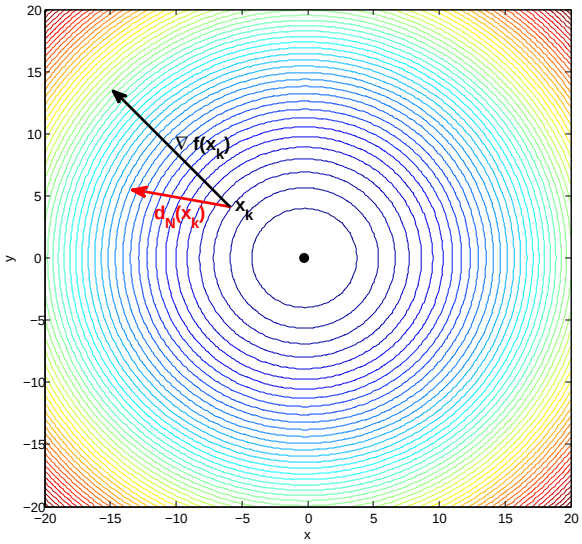


Figura 12.1: Direção ascendente.

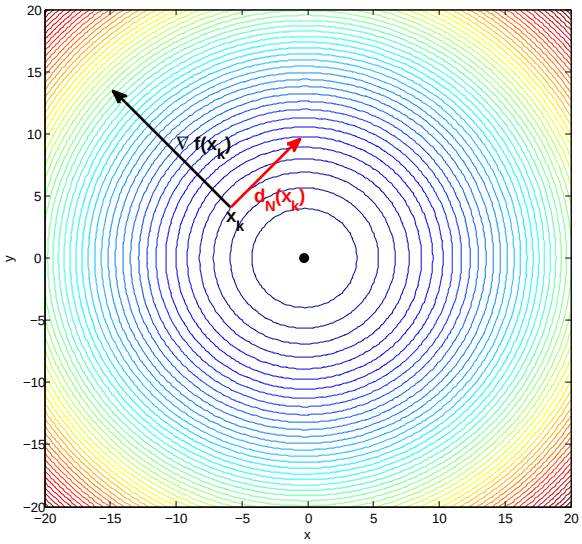


Figura 12.2: Direção ortogonal.

Direção Newton muito grande

A direção d_N^k, solução do sistema Newton (12.2), ainda que seja descendente, isto é, $\nabla f(x^k)^T d_N^k < 0$, pode ser muito grande (Figura 12.3) e por isso não se verifica

$$f(x^k + d_N^k) < f(x^k).$$

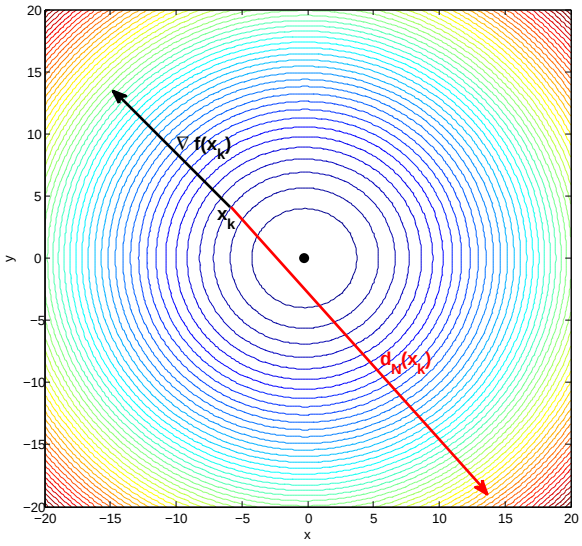


Figura 12.3: Direção muito grande.

Matriz Hessiana em x^k singular

A matriz dos coeficientes do sistema Newton (12.2) $\nabla^2 f(x^k)$ pode ser singular, o que significa que o sistema Newton não tem solução ou tem uma infinidade de soluções, ou seja,

$$\nexists d_N^k \text{ única.}$$

12.3.3 Desvantagens do método de Newton

A maior desvantagem do método de Newton é o facto de exigir o cálculo de segundas derivadas, por vezes muito complexo e dispendioso. Quando a expressão de $f(x)$ é complicada, estas tornam-se difíceis de calcular. Por outro lado, quando a dimensão do problema, n, for muito grande, o esforço de cálculo também será grande. Há ainda a questão de a convergência ser local e, por esse facto, estar condicionada à escolha do valor inicial.

Para estas limitações serem ultrapassadas, podem implementar-se algumas soluções, como a seguir se descreve, dando origem ao método de segurança de Newton.

12.4 Método de segurança de Newton

Para que o método de Newton resulte, é necessário ultrapassar as suas limitações de modo a garantir que a direção obtida é descendente para a função f. Para isso, implementa-se o método de segurança de Newton que consiste em qualquer iteração k:

- quando $\nabla^2 f(x^k)$ é singular, usar-se $d_{SN}^k = -\nabla f(x^k)$, em que $-\nabla f(x^k)$ é a direção de descida máxima e é descendente para f.
- quando d_N^k é ortogonal ao gradiente, isto é, $\left|\nabla f(x^k)^T d_N^k\right| \leq \eta$ (com $\eta > 0$ e $\eta \approx 0$), usar-se $d_{SN}^k = -\nabla f(x^k)$.
- quando d_N^k é ascendente, isto é, $\nabla f(x^k)^T d_N^k > \eta$, usar-se $d_{SN}^k = -d_N^k$.
- quando d_N^k é descendente, isto é, $\nabla f(x^k)^T d_N^k < \eta$, usar-se $d_{SN}^k = d_N^k$.

A direção d_{SN}^k assim obtida, é descendente para todo o k.

No Algoritmo 12.4 encontra-se descrito o método de segurança de Newton.

Algoritmo 12.4 Método de segurança de Newton

ler: x^k e η

resolver o sistema linear Newton $\nabla^2 f(x^k) d_N^k = -\nabla f(x^k)$ por EGPP

se $\exists d_N^k$ (o sistema linear tem solução única) **então**

se $\left|\nabla f(x^k)^T d_N^k\right| \leq \eta$ **então**

$d_{SN}^k = -\nabla f(x^k)$

senão

se $\nabla f(x^k)^T d_N^k > \eta$ **então**

$d_{SN}^k = -d_N^k$

senão

$d_{SN}^k = d_N^k$

fim se

fim se

senão

$d_{SN}^k = -\nabla f(x^k)$

fim se

O método de segurança de Newton mantém as mesmas propriedades de convergência que o método de Newton.

12.5 Método quasi-Newton

Como foi referido anteriormente, uma das grandes desvantagens do método de Newton é o facto de exigir o cálculo das segundas derivadas de f (12.2). Para se evitar o cálculo das segundas derivadas, pode usar-se uma aproximação à matriz Hessiana,

$$B^k \approx \nabla^2 f(x^k)$$

e assim a direção de procura passa a ser calculada pela resolução do sistema linear por EGPP

$$B^k d_{QN}^k = -\nabla f(x^k).$$

No entanto, o sistema Newton (12.2) pode ser escrito de forma equivalente

$$d_N^k = -\left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k),$$

por isso pode usar-se em alternativa, em cada iteração k, uma aproximação à inversa da Hessiana

$$H^k \approx \left(\nabla^2 f(x^k)\right)^{-1}$$

e calcula-se a direção de procura pelo produto desta matriz pelo vetor gradiente, que é um cálculo mais simples que a resolução de um sistema linear. Assim,

$$d_{QN}^k = -H^k \nabla f(x^k). \tag{12.3}$$

Usando este procedimento, evita-se não só o cálculo das segundas derivadas de f para formar a matriz Hessiana, como também a resolução de um sistema linear em cada iteração, que é substituído pelo produto de uma matriz por um vetor.

12.5.1 Características da matriz H

A matriz H deve aproximar o melhor possível a inversa de $\nabla^2 f(x^k)$, isto é, deve verificar-se a condição secante

$$H^k y^{k-1} = s^{k-1},$$

em que y representa a variação no gradiente da iteração $k-1$ para a iteração k e é dado por

$$y^{k-1} = \nabla f(x^k) - \nabla f(x^{k-1}).$$

Por usa vez, s representa a variação verificada em x da iteração $k-1$ para a iteração k e é dado por

$$s^{k-1} = x^k - x^{k-1} = \alpha^{k-1} d_{QN}^{k-1}.$$

A matriz H deve ainda ser, preferencialmente, simétrica, pois $\nabla^2 f(x^k)^{-1}$ também é simétrica e definida positiva e a direção (12.3) é descendente para f em x^k.

As matrizes H^k são geradas através de fórmulas de atualização do tipo

$$H^k = H^{k-1} + E^{k-1}$$

e devem manter-se simétricas e definidas positivas. Para que isso aconteça, a matriz inicial H^1 deve ser também simétrica e definida positiva. A matriz mais usual para iniciar este processo é a matriz identidade, isto é,

$$H^1 = I.$$

A matriz identidade não é necessariamente uma boa aproximação a $\nabla^2 f(x^k)^{-1}$, mas as fórmulas de atualização rapidamente ultrapassam esse problema e melhoram as aproximações.

Existem várias fórmulas de atualização para as matrizes H, no entanto, nem todas conservam a matriz simétrica e definida positiva. As fórmulas (12.4) e (12.5) conservam as aproximações H simétricas e definidas positivas. A condição ${y^{k-1}}^T s^{k-1} > 0$ é necessária e suficiente para que as matrizes se conservem simétricas e definidas positivas e estas fórmulas verificam essa condição.

Fórmula de atualização de Davidon, Fletcher e Powell (DFP)

$$H^k = H^{k-1} - \frac{H^{k-1} y^{k-1} {y^{k-1}}^T H^{k-1}}{{y^{k-1}}^T H^{k-1} y^{k-1}} + \frac{s^{k-1} {s^{k-1}}^T}{{s^{k-1}}^T y^{k-1}} \tag{12.4}$$

Fórmula de atualização de Broyden, Fletcher, Goldfarb e Shanno (BFGS)

$$H^k = \left(I - \frac{s^{k-1} {y^{k-1}}^T}{{s^{k-1}}^T y^{k-1}}\right) H^{k-1} \left(I - \frac{y^{k-1} {s^{k-1}}^T}{{s^{k-1}}^T y^{k-1}}\right) + \frac{s^{k-1} {s^{k-1}}^T}{{s^{k-1}}^T y^{k-1}} \tag{12.5}$$

12.5.2 Propriedades do método quasi-Newton

O método quasi-Newton, tal como o método de Newton, tem convergência local, ou seja, a convergência para a solução só é garantida quando a aproximação inicial x^1 estiver na vizinhança da solução. No entanto, a rapidez de convergência é inferior à do método de Newton, sendo que é superlinear. Significa que se verifica

$$\left\|x^{k+1} - x^*\right\| \leq \gamma_k \left\|x^k - x^*\right\|$$

com a sucessão $\{\gamma_k\} \to 0$ quando $k \to \infty$.

À semelhança do método de Newton, o método quasi-Newton tem a propriedade da terminação quadrática, isto é, o mínimo de uma função quadrática $q(x)$, $x \in \mathbb{R}^n$ obtém-se em n ou menos que n iterações.

Devido aos erros de arredondamento que se cometem nos cálculos ao longo das iterações, a matriz H^k pode deixar de ser definida positiva e assim a direção d^k_{QN} deixa de ser descendente para f em x^k. Neste caso deve fazer-se $H^k = I$, que é simétrica e definida positiva. Quando isto acontece, $d^k_{QN} = -\nabla f(x^k)$ – direção de descida máxima.

Descreve-se no Algoritmo 12.5 o método quasi-Newton.

Algoritmo 12.5 Método quasi-Newton

ler: x^k

se $k = 1$ **então**

$H^k = I$

senão

$s^{k-1} \leftarrow x^k - x^{k-1}$

$y^{k-1} \leftarrow \nabla f(x^k) - \nabla f(x^{k-1})$

atualizar H^k por (12.4) ou (12.5)

fim se

$d^k_{QN} \leftarrow -H^k \nabla f(x^k)$

se $\nabla f(x^k)^T d^k_{QN} \geq 0$ **então**

$d^k_{QN} \leftarrow -\nabla f(x^k)$

fim se

Capítulo 13

Método de Nelder-Mead

O método do simplex de Nelder-Mead é um dos métodos mais conhecidos na classe dos métodos de otimização sem derivadas ("derivative-free optimization"). Estes são os métodos adequados quando se está em presença de uma função objetivo $f(x)$ não suave, não linear, descontínua e não convexa.

Definição 13.0.1 *A função $f(x)$ diz-se suave se é continuamente diferenciável até à segunda ordem no seu domínio, isto é, as primeiras e as segundas derivadas existem em todos os pontos do domínio. Uma função não suave pode ser contínua e não ser diferenciável em alguns pontos do domínio.*

Estas características também podem estar presentes nas restrições do problema, mas este capítulo diz respeito apenas a problemas sem restrições.

Exemplo 13.1 *Considere-se a seguinte função não diferenciável*

$$f(x_1,x_2)=\begin{cases} 5\sqrt{9x_1^2+16x_2^2} & para \quad x_1 \geq |x_2| \\ 9x_1+16|x_2| & para \quad 0<x_1<|x_2| \\ 9x_1+16x_2-x_1^9 & para \quad x_1 \leq 0, x_2 \geq 0 \\ 9x_1-16x_2-x_1^9 & para \quad x_1 \leq 0, x_2 < 0 \end{cases}$$

A função f não é diferenciável em pontos que verificam $x_1 \leq 0$ e $x_2 = 0$. Significa que f não é suave.

Se se usar um método baseado em derivadas para calcular um ponto estacionário a partir de um ponto inicial que verifique $x_1 > |x_2| > (9/16)^2|x_1|$, o processo converge para $(0,0)$.

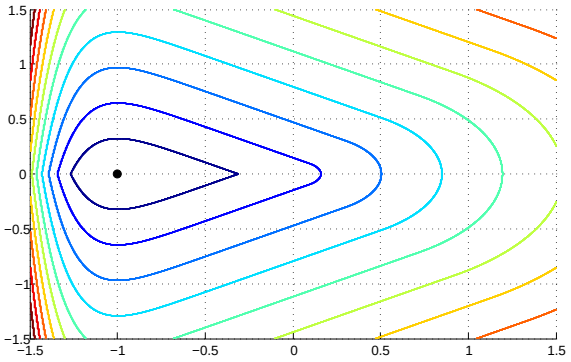


Figura 13.1: Função não diferenciável.

Neste ponto o gradiente pode não ser nulo. Na realidade, a solução para o problema $\min\ f(x)$ *é* $x^* = (-1, 0)^T$.

Se a função é não suave não devem usar-se métodos que recorram a derivadas para calcular a solução, como pode verificar-se no Exemplo 13.1. Mesmo que a derivada exista num certo ponto y e se $y \approx \overline{x}$, em que $\overline{x}$ é um ponto onde a derivada não existe, não deve usar-se para prever o comportamento de f em $\overline{x}$.A informação sobre a derivada também não deve usar-se, neste caso, para terminar o processo iterativo pois não é necessariamente nula perto de $\overline{x}$.

Casos como o descrito no Exemplo 13.1 têm originado um aumento da popularidade dos métodos sem derivadas.

O método de Nelder-Mead é um método baseado num simplex. Um simplex em $\mathbb{R}^n$ pode ser entendido como um poliedro com $n+1$ vértices distintos. Quando os lados do simplex têm todos o mesmo comprimento, este diz-se regular. Assim, em $\mathbb{R}^2$ um simplex regular é um triângulo equilátero e em $\mathbb{R}^3$ um tetraedro. Se os vértices de um simplex em $\mathbb{R}^n$ forem denotados por $X_i,\ i = 1, \ldots, n+1$ e substituirmos um dos vértices por W, obtém-se um novo simplex. Pode assim gerar-se uma sequência de simplex alterando um vértice de cada vez.

Se for estabelecido um conjunto de regras para alterar um dado vértice do simplex e essas regras se basearem apenas no valor da função objetivo f, é possível gerar uma sequência de simplex, cada um deles gerado com base no anterior, avaliando a função objetivo. Se estas regras forem tais que esta sequência contém o minimizante de f, x^*, então deverá existir um método de estimar x^*, sendo que a precisão da estimativa depende do tamanho do simplex final que contém x^*.

O método de Nelder-Mead é iterativo e define em cada iteração um simplex, isto é, um

poliedro. Em $\mathbb{R}^n$, sejam $x_1, x_2, \ldots, x_n, x_{n+1}$ os $n+1$ pontos do simplex de dimensão n. Então

$$S_k = \langle X_1, X_2, \ldots, X_n, X_{n+1} \rangle$$

representa o simplex da iteração k em que os vértices já estão ordenados por ordem crescente dos valores da função objetivo, isto é, $f(X_1) \leq f(X_2) \leq \ldots \leq f(X_n) \leq f(X_{n+1})$. Os vértices mais importantes do simplex são

- X_1 – o melhor vértice;
- X_n – o segundo pior vértice;
- X_{n+1} – o pior vértice.

Por exemplo, em $\mathbb{R}^2$, o simplex formado pelos $n + 1 = 3$ pontos define um triângulo (Figura 13.2).

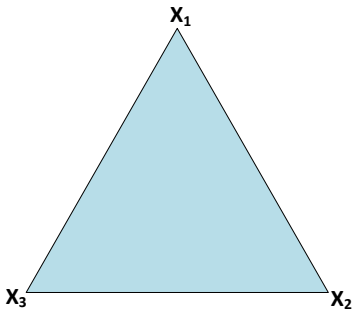


Figura 13.2: Simplex em $\mathbb{R}^2$.

Em cada iteração deste método definem-se pontos auxiliares, que são candidatos a vértices do novo simplex. Serão rejeitados ou aceites de acordo com a comparação do seu valor da função objetivo com os valores da função objetivo nos vértices mais importantes do simplex referidos acima - $f(X_1)$, $f(X_n)$ e $f(X_{n+1})$.

Há um conjunto de operações básicas que permitem construir os pontos auxiliares:

- refletir;
- expandir;
- contrair para o interior;
- contrair para o exterior;

- encolher.

Seja

$$S_k = \langle X_1, \ldots, X_{n+1} \rangle$$

o simplex de uma iteração k já ordenado.

A explicação que se segue é ilustrada para $\mathbb{R}^2$, mas em dimensões superiores os princípios a seguir são os mesmos, aumentando apenas o número de vértices no simplex.

Em cada iteração começa-se por calcular o centróide do simplex (Figura 13.3), que é o ponto médio do hiperplano definido por $X_1, \ldots, X_n$, e é dado por

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

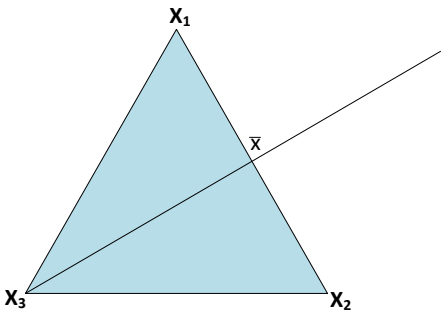


Figura 13.3: Centróide em $\mathbb{R}^2$.

De seguida, calcula-se o vértice refletido (Figura 13.4).

$$x_r = (1 + \alpha)\overline{x} - \alpha X_{n+1}.$$

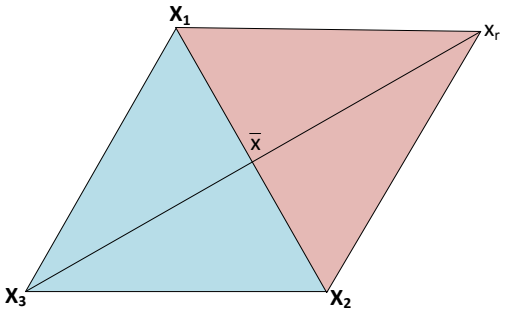


Figura 13.4: Vértice refletido em $\mathbb{R}^2$ com $\alpha = 1$.

Se x_r for bom, isto é, $f(X_1) \leq f(x_r) < f(X_n)$, aceita-se x_r e o simplex da iteração seguinte é

$$S_{k+1} = \langle X_1, \ldots, X_n, x_r \rangle.$$

Se x_r for muito bom, isto é, $f(x_r) < f(X_1)$, faz-se uma expansão do simplex (Figura 13.5),

$$x_e = \gamma x_r + (1 - \gamma)\overline{x}.$$

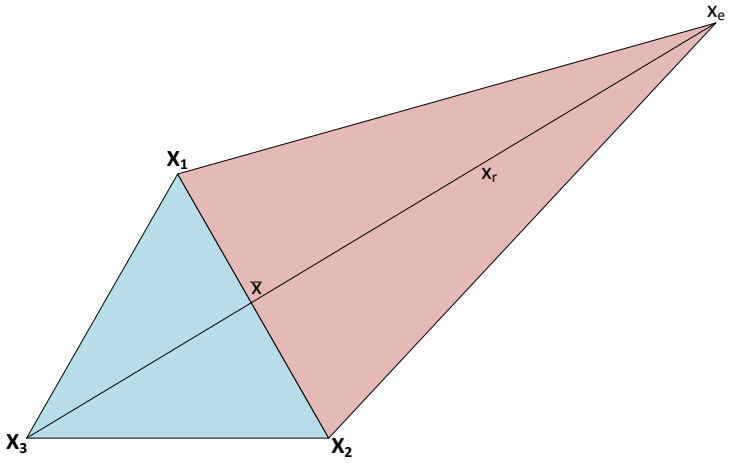


Figura 13.5: Vértice expandido em $\mathbb{R}^2$ com $\gamma = 2$.

Se x_e for muito bom, isto é, $f(x_e) < f(X_1)$, aceita-se x_e e o simplex da iteração seguinte é

$$S_{k+1} = \langle X_1, \ldots, X_n, x_e \rangle,$$

caso contrário aceita-se x_r e o simplex da iteração seguinte é

$$S_{k+1} = \langle X_1, \ldots, X_n, x_r \rangle.$$

Se x_r for fraco, isto é, $f(X_n) \leq f(x_r) < f(X_{n+1})$, faz-se uma contração para o exterior (Figura 13.6),

$$\hat{x}_c = \beta x_r + (1 - \beta)\overline{x}.$$

Se $\hat{x}_c$ for bom, isto é, $f(\hat{x}_c) < f(X_n)$, aceita-se $\hat{x}_c$ e o simplex da próxima iteração é

$$S_{k+1} = \langle X_1, \ldots, X_n, \hat{x}_c \rangle,$$

caso contrário encolhe-se o simplex (Figura 13.8).

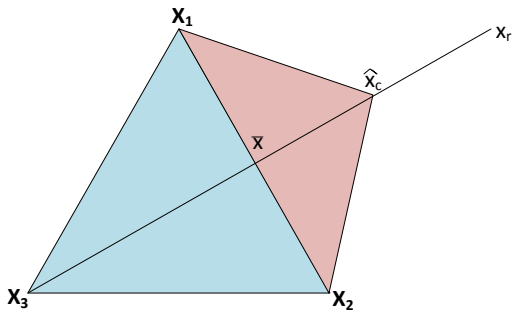


Figura 13.6: Vértice contraído para o exterior em $\mathbb{R}^2$ com $\beta = 1/2$.

Se x_r for muito fraco, isto é, $f(x_r) \geq f(X_{n+1})$, faz-se uma contração para o interior (Figura 13.7),

$$x_c = \beta X_{n+1} + (1 - \beta)\overline{x}.$$

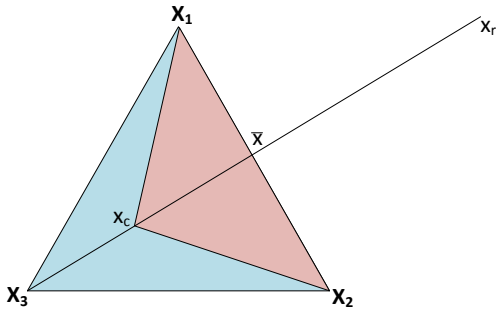


Figura 13.7: Vértice contraído para o interior.

Se x_c for bom, isto é, $f(x_c) < f(X_n)$, aceita-se x_c e o simplex da iteração seguinte é

$$S_{k+1} = \langle X_1, \ldots, X_n, x_c \rangle,$$

caso contrário encolhe-se o simplex (Figura 13.8).

Encolher o simplex consiste em substituir cada um dos vértices X_i, $i = 2, \ldots, n+1$ pelo segmento médio do segmento que une X_i a X_1, isto é

$$x_i = \frac{X_i + X_1}{2},$$

e o simplex da iteração seguinte é

$$S_{k+1} = \langle X_1, x_2 \ldots, x_n, x_{n+1} \rangle.$$

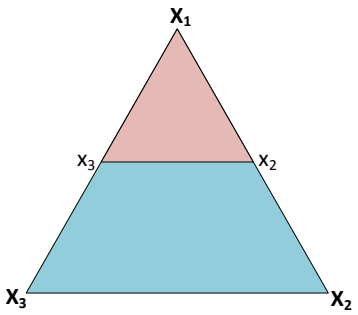


Figura 13.8: Encolher o simplex em $\mathbb{R}^2$.

Em suma,

$$
x_r \begin{cases}
\underset{f(x_r)<f(X_1)}{\text{muito bom}} \Rightarrow x_e \begin{cases} \underset{f(x_e)<f(X_1)}{\text{muito bom}} \Rightarrow \text{aceita-se } x_e \\ \text{caso contrário} \Rightarrow \text{aceita-se } x_r \end{cases} \\
\underset{f(X_1)\leq f(x_r)<f(X_n)}{\text{bom}} \Rightarrow \text{aceita-se } x_r \\
\underset{f(X_n)\leq f(x_r)<f(X_{n+1})}{\text{fraco}} \Rightarrow \hat{x}_c \begin{cases} \underset{f(\hat{x}_c)<f(X_n)}{\text{bom}} \Rightarrow \text{aceita-se } \hat{x}_c \\ \text{caso contrário} \Rightarrow \text{encolhe-se o simplex} \end{cases} \\
\underset{f(x_r)\geq f(X_{n+1})}{\text{muito fraco}} \Rightarrow x_c \begin{cases} \underset{f(x_c)<f(X_n)}{\text{bom}} \Rightarrow \text{aceita-se } x_c \\ \text{caso contrário} \Rightarrow \text{encolhe-se o simplex} \end{cases}
\end{cases}
$$

Critério de paragem

O critério de paragem do método de Nelder-Mead consiste em verificar se o tamanho relativo do simplex é inferior ou igual a uma quantidade positiva pequena. Significa que o processo iterativo para se

$$\frac{1}{\Delta} \max_{2\leq i\leq n+1} \|X_i - X_1\|_2 \leq \varepsilon,$$

com $\Delta = \max(1, \|X_1\|_2)$. Para verificar o critério de paragem é necessário que o simplex se encontre ordenado.

Se o critério de paragem for verificado, o vértice do simplex com menor valor da função objectivo, X_1, é considerado como a melhor aproximação calculada à solução. Caso o critério de paragem não se verifique, o processo iterativo continua. Outras implementações diferentes do método de Nelder-Mead podem surgir com critérios de paragem diferentes.

O método de Nelder-Mead encontra-se descrito no Algoritmo 13.6.

Algoritmo 13.6 Método de Nelder-Mead

ler: $X_1, \ldots, X_{n+1}$ e ε, $\alpha = 1$, $\beta = 0.5$, $\gamma = 2$

ordenar o simplex de acordo com o valor da função ($f(X_1) \leq \ldots \leq f(X_{n+1})$)

$S_1 \leftarrow \langle X_1, \ldots, X_{n+1} \rangle$

$k \leftarrow 1$

repetir

$\overline{x} \leftarrow \frac{1}{n} \sum_{i=1}^{n} X_i$

$x_r \leftarrow (1 + \alpha)\overline{x} - \alpha X_{n+1}$

se $f(x_r) < f(X_n)$ **então**

se $f(x_r) \geq f(X_1)$ **então**

$S_{k+1} \leftarrow \langle X_1, \ldots, X_n, x_r \rangle$

senão

$x_e \leftarrow \gamma x_r + (1 - \gamma)\overline{x}$

se $f(x_e) < f(X_1)$ **então**

$S_{k+1} \leftarrow \langle X_1, \ldots, X_n, x_e \rangle$

senão

$S_{k+1} \leftarrow \langle X_1, \ldots, X_n, x_r \rangle$

fim se

fim se

senão

se $f(x_r) \geq f(X_{n+1})$ **então**

$x_c \leftarrow \beta X_{n+1} + (1 - \beta)\overline{x}$

se $f(x_c) < f(X_n)$ **então**

$S_{k+1} \leftarrow \langle X_1, \ldots, X_n, x_c \rangle$

senão

$x_i \leftarrow \frac{X_i + X_1}{2}, \quad i = 2, \ldots, n+1$

$S_{k+1} \leftarrow \langle X_1, x_2, \ldots, x_{n+1} \rangle$

fim se

senão

$\hat{x}_c \leftarrow \beta x_r + (1 - \beta)\overline{x}$

se $f(\hat{x}_c) < f(X_n)$ **então**

$S_{k+1} \leftarrow \langle X_1, \ldots, X_n, \hat{x}_c \rangle$

senão

$x_i \leftarrow \frac{X_i + X_1}{2}, \quad i = 2, \ldots, n+1$

$S_{k+1} \leftarrow \langle X_1, x_2, \ldots, x_{n+1} \rangle$

fim se

fim se

fim se

ordenar o simplex de acordo com o valor da função ($f(X_1) \leq \ldots \leq f(X_{n+1})$)

$S_{k+1} \leftarrow \langle X_1, \ldots, X_{n+1} \rangle$

$k \leftarrow k + 1$

$\Delta = \max(1, \|X_1\|_2)$

até $\frac{1}{\Delta} \times \max_{2 \leq i \leq n+1} \|X_i - X_1\|_2 \leq \varepsilon$

$x_{\min} \leftarrow X_1$