ANÁLISE NUMÉRICA I

Lic Matemática

Maria Joana Soares

2013/2104

INTRODUÇÃO

Análise Numérica?

Ramo da Matemática que estuda o desenvolvimento e análise de métodos que permitem a resolução de problemas matemáticos utilizando apenas um número finito de operações elementares da aritmética, isto é, que estuda os chamados métodos numéricos, em oposição aos métodos analíticos, como a derivação, integração, etc.

Exemplo

Calcular o valor de $\exp(0.125)$.
Temos

$$\exp(0.125) = \lim_{n\to\infty} \left(1 + \frac{0.125}{n}\right)^n = ?$$

Uma possível solução *numérica* pode ser obtida usando os primeiros 8 termos da expansão de $\exp(x)$ em série de McLaurin:

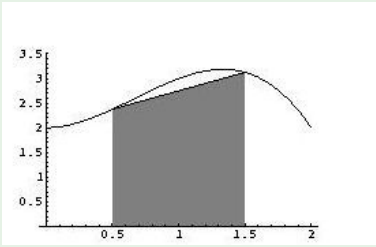
$$\exp(0.125) \approx 1 + 0.125 + 0.125^2/2 + \cdots + 0.125^7/7! \approx 1.13315 \ .$$

Numérico $\neq$ Analítico $\Longrightarrow$ *Erro de truncatura (ou de discretização)*

Erro de truncatura

Exemplo

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



Regra do trapézio

$$E(f) = -\frac{(b-a)^3}{12} f''(\xi), \quad \xi \in [a,b]$$

- Muitos dos métodos numéricos estudados são muito antigos (por exemplo, método de Newton para resolução de equações não lineares, fórmula de Lagrange para o polinómio interpolador, método de Gausss para sistemas lineares), mas vieram a assumir grande importância devido ao aparecimento e acessibilidade de computadores. [1]
- Podemos dizer que a *Análise Numérica* é o estudo dos métodos de resolução de problemas matemáticos usando computadores $\Longrightarrow$ *Erros de arredondamento* (devidos à capacidade limitada de representação de números em computador).
- Ferramenta indispensável em todas as áreas das Ciências Aplicadas e Engenharia.
  Exemplos: dinâmica de fluidos, engenharia de estruturas, engenharia eltrotécnica, meteorologia, astronomia, biologia, medicina, economia...

[1] A própria designação *Análise Numérica* surge em 1946, no início da era do computador.

ARITMÉTICA COMPUTACIONAL

Representação de reais em ponto flutuante

A representação normalizada de um número real $x \neq 0$ na base b (b inteiro, ≥ 2) é a representação desse número na forma

$$x = \pm(.d_1 d_2 d_3 \ldots)_b \times b^e,$$

onde

- $d_i \in \{0, \ldots, b-1\}$, $d_1 \neq 0$
- $e \in \mathbb{Z}$
- $(.d_1 d_2 d_3 \ldots)_b$ designa $d_1\, b^{-1} + d_2\, b^{-2} + d_3 b^{-3} + \cdots$.

Todo o número real admite uma expansão na forma anterior, sendo esta única, se excluirmos representações em que todos os dígitos sejam, a partir de determinada ordem, iguais a $b-1$; por exemplo, na base $b = 10$, excluimos a representação $.499999\ldots$ para o número 0.5.

- $m_x := (.d_1 d_2 \ldots)_b$ $\hookrightarrow$ mantissa de x
- e $\hookrightarrow$ expoente de x.

Note-se que $m_x \geq (.1)_b = b^{-1}$.

Sistema de numeração de máquina

Num computador, o número de dígitos da mantissa é fixo (t) e o expoente é limitado por um valor mínimo (m) e um valor máximo (M).
Assim, um sistema de numeração de ponto flutuante ou sistema de numeração de máquina $F(b, t, m, M)$ é caraterizado por quatro parâmetros:

- b - base;
- t - número de dígitos da mantissa;
- m - valor mínimo dos exponente;
- M - valor máximo do expoente.

Definição

Constituem o sistema $F(b, t, m, M)$, para além do número zero, todos os números que se puderem exprimir na forma

$$\pm(.d_1 d_2 \ldots d_t)_b \times b^e,$$

onde $d_1, d_2, \ldots d_t \in \{0, 1 \ldots, b-1\}$, $d_1 \neq 0$ e $e \in \mathbb{Z}$ com $m \leq e \leq M$.

Overflow e underflow

- Os números acima definidos são os chamados números normais ou normalizados.

  Um sistema $F(b,t,m,M)$ pode ainda admitir os chamados números desnormalizados ou subnormais, que são os números obtidos deixando de impor a condição $d_1 \neq 0$ quando o expoente assume o valor mínimo m.
- O maior número de $F(b,t,m,M)$ é

$$\Omega := (1-b^{-t})b^M,$$

  dito nível de overflow.
- O menor número positivo (normalizado), chamado nível de underflow, é dado por

$$\omega := b^{m-1}.$$

  O menor número positivo de um sistema que admita números desnormalizados é b^{m-t}.

Nota: Se nada for dito em contrário, quando nos referirmos a um sistema $F(b,t,m,M)$, consideramos apenas os números normalizados.

Definição

Ao conjunto

$$R_F := [-\Omega,-\omega]\cup\{0\}\cup[\omega,\Omega]$$

chamamos conjunto dos números representáveis.

Note-se que $F \subsetneq R_F$. Mais precisamente, os números de máquina constituem um subconjunto **finito** do conjunto dos números representáveis.

Exemplo

No sistema $F(10,4,-99,99)$, tem-se:

$$\Omega = (1-10^{-4})\times 10^{99} = 0.9999\times 10^{99},$$
$$\omega = 0.1\times 10^{-99} = 10^{-99-1} = 10^{-100}$$
$$R_F = [-0.9999\times 10^{99}, -10^{-100}]\cup\{0\}\cup[10^{-100}, 0.9999\times 10^{99}]$$

Por exemplo, o número $\pi \in R_F$, mas $\pi \notin F(10,4,-99,99)$, já que $\pi = 3.141592654...$ não pode escrever-se com uma mantissa de apenas quatro dígitos.

Arredondamento

Dado $x \in R_F$, torna-se necessário encontrar um número de máquina que o represente. Será natural pretender que esse número, que designaremos por $fl(x)$, esteja à menor distância possível de x, isto é, satisfaça

$$|fl(x)-x| \leq |g-x|,\ \forall g \in F.$$

Dizemos, então, que $fl(x)$ foi obtido por arredondamento. [2] Caso existam dois números de máquina à mesma distância de x, é frequente, nas máquinas, usar o chamado arredondamento para par, em que o número escolhido para representar x é aquele cujo último dígito é par.
Note-se que, se tivermos um sistema de numeração $F = F(b,t,m,M)$, se $x \in R_F$ tiver um expoente e na notação normalizada, ao usarmos arredondamento, tem-se

$$|fl(x)-x| \leq \frac{1}{2}b\,b^{-(t+1)}\,b^e = \frac{1}{2}b^{-t}b^e.$$

[2] Em oposição à truncatura, em que simplemente se ignoram os dígitos da mantissa para além do t-ésimo dígito.

Arredondamento para infinito

O arredondamento a que estamos mais habituados, em que, no caso de empate, arredondamos "para cima", é chamado arredondamento para infinito.
Por exemplo, se pretendermos apenas números com 3 dígitos, arredondamos 49.65 para 49.7 e 49.55 para 49.6, se usarmos o arredondamento para infinito, mas arredondamos ambos os números para 49.6, no caso de arredondamento para par.

Convenção

Neste curso, se nada for dito em contrário, por uma questão de simplicidade, usaremos o arredondamento para infinito; no entanto, devemos ter presente que, no MATLAB, o arredondamento usado é o arredondamento para par.

Norma IEEE 754

Com o objectivo de uniformizar as operações nos sistemas de ponto flutuante foi publicada, em 1985, a norma IEEE 754. [3] Esta norma especifica dois formatos básicos para representação de números em sistema de ponto flutuante: simples e duplo.

O formato simples corresponde ao sistema $F(2, 24, -125, 128)$ e o duplo corresponde a $F(2, 53, -1021, 1024)$.

Ambos os sistemas admitem números desnormalizados.

O sistema de numeração IEEE admite ainda os "números"especiais $+\infty$ e $-\infty$ (**Inf** e $-$**Inf**, no MATLAB) para representar, por exemplo, o resultado da divisão de um número por zero, bem como o símbolo especial **NaN** (Not a Number), para representar o resultado de operações não definidas matematicamente, tais como $0/0$, $\infty - \infty$, etc.

[3] IEEE- Institute for Electrical and Electronics Engineers.

Norma IEEE (cont.)

A norma IEEE 754 especifica também as regras de arredondamento a utilizar. Por defeito:

- para $x \in R_F$, é uilizado o arredondamento para par;
- se $x > \Omega$, $fl(x) = +\infty$ e se $x < -\Omega$, $fl(x) = -\infty$;
- se $2^{m-t} \leq x < \omega$ (onde $m = -125$, $t = 24$, no formato simples, e $m = -1021$, $t = 53$, no formato duplo), $fl(x)$ é o número desnormalizado mais próximo de x;
- se $x < 2^{m-t}$, $fl(x) = 0$.

É importante saber que

Por defeito, quando trabalhamos no Matlab, é usado o formato duplo IEEE 754, ou seja, o sitema de numeração de máquina é o sistema $F(2, 53, -1021, 1024)$ e o arredondamento usado é o descrito acima.

Definição

Numa máquina com sistema de numerção $F = F(b, t, m, M)$, chama-se epsilon da máquina, e denota-se por ϵ, a diferença entre o número de máquina imediatamente superior a 1 e o número 1, isto é,

$$\epsilon := b^{1-t}.$$

Definição

A unidade de erro de arredondamento de um sistema $F(b, t, m, M)$ é definida como

$$\mu := \frac{1}{2} b^{1-t},$$

ou seja,tem-se $\mu = \frac{1}{2}\epsilon$.

Exemplo

No sistema $F(2, 53, -1021, 1024)$, tem-se:

$$\epsilon = 2^{-52} \approx 2.2204 \times 10^{-16}, \quad \mu = \frac{1}{2} \times 2^{-52} = 2^{-53} \approx 1.1102 \times 10^{-16}$$

Operações de ponto flutuante

Representaremos as operações de ponto flutuante ou operações de máquina pelo símbolo usual rodeado por $\circ$; por exemplo $\oplus$, $\otimes$. Admitimos que o resultado de uma operação de ponto flutuante (envolvendo números de máquina) é obtido por arredondamento do resultado da operação exacta, isto é, se $x, y \in F$, $x \oplus y = fl(x + y)$, $x \otimes y = fl(x \times y)$, etc. De salientar que as operações de ponto flutuante **não** satisfazem todas as propriedades usuais das correspondentes operações em $\mathbb{R}$.

Erro absoluto e erro relativo

Seja x um dado número e $\tilde{x}$ um valor aproximado para x.

Definição

A $E_{\tilde{x}} := x - \tilde{x}$ chamamos erro (absoluto) do valor aproximado $\tilde{x}$ para x.

Definição

Se $x \neq 0$, a $R_{\tilde{x}} := \dfrac{x - \tilde{x}}{x}$ chamamos erro relativo do valor aproximado $\tilde{x}$ para x.

Exemplo

Sejam $x = 1/3, y = 1/3000, \tilde{x} = .3333, \tilde{y} = .0003$. Então

$$E_{\tilde{x}} = E_{\tilde{y}} = 0.00003333\ldots, \; R_{\tilde{x}} = 10^{-4}, \; R_{\tilde{y}} = 10^{-1}.$$

O erro relativo é mais "relevante" do que o erro absoluto.

Notas

1. Na prática, é frequente estarmos apenas interessados no valor absoluto dos erros absoluto e relativo, continuando a designá-los pelos mesmos nomes, desde que tal seja claro pelo contexto.
2. Como na definição de erro relativo o valor de x não é conhecido, é usual considerar-se a estimativa
$$|R_{\tilde{x}}| \approx \frac{|x - \tilde{x}|}{|\tilde{x}|}$$
3. É frequente expressar-se o (valor absoluto) do erro relativo em percentagem; por exemplo, se $|R_{\tilde{x}}| = 0.05$, dizemos que $\tilde{x}$ tem um erro relativo de 5%.
4. Da definição de erro relativo, obtém-se de imediato
$$\tilde{x} = x(1 - R_{\tilde{x}})$$

Erro relativo de arredondamento

Num sistema $F(b, t, m, M)$, dado um número $x \neq 0$ (representável), tem-se

$$|R_{fl(x)}| = \frac{|x - fl(x)|}{|x|} \leq \frac{\frac{1}{2} b^{-t}\, b^{e}}{b^{-1}\, b^{e}} = \frac{1}{2} b^{1-t} = \mu.$$

O erro relativo cometido ao arredondar um número para um número de máquina é majorado pela unidade de erro de arredondamento da máquina.

Exemplo

Por exemplo, quando trabalhamos no MATLAB, tem-se

$$|R_{fl(x)}| \leq 2^{-53} \approx 1.1102 \times 10^{-16}$$

Propagação de erros nas operações usuais

Sejam $\tilde{x}$ e $\tilde{y}$ valores aproximados para x e y, respetivamente $(x, y \neq 0)$, e sejam $S = x + y$, $D = x - y$, $P = x \times y$ e $Q = x/y$. Sejam $\widetilde{S}, \widetilde{P}$ $\widetilde{D}$ e $\widetilde{Q}$ os valores aproximados para S, P, D e Q obtidos usando os valores $\tilde{x}$ e $\tilde{y}$ em vez de x e y e admitindo que as operações são efetuadas exatamente. Designando por $E_{\tilde{u}}$ e $R_{\tilde{u}}$, respectivamente, o erro absoluto e erro relativo no valor aproximado $\widetilde{u}$ para u, podem estabelecer-se os seguintes resultados:

$$E_{\widetilde{S}} = E_{\tilde{x}} + E_{\tilde{y}}$$

$$E_{\widetilde{D}} = E_{\tilde{x}} - E_{\tilde{y}}$$

$$E_{\widetilde{P}} = E_{\tilde{x}}\tilde{y} + E_{\tilde{y}}\tilde{x} + E_{\tilde{x}}E_{\tilde{y}}$$

$$E_{\widetilde{Q}} = \frac{E_{\tilde{x}}\tilde{y} - E_{\tilde{y}}\tilde{x}}{\tilde{y}\,(\tilde{y} + E_{\tilde{y}})}$$

Temos, também

$$R_{\widetilde{S}} = \frac{x}{x+y}R_{\tilde{x}} + \frac{y}{x+y}R_{\tilde{y}}$$

$$R_{\widetilde{D}} = \frac{x}{x-y}R_{\tilde{x}} - \frac{y}{x-y}R_{\tilde{y}}$$

$$R_{\widetilde{P}} = R_{\tilde{x}} + R_{\tilde{y}} - R_{\tilde{x}}R_{\tilde{y}}$$

$$R_{\widetilde{Q}} = \frac{R_{\tilde{x}} - R_{\tilde{y}}}{1 - R_{\tilde{y}}}$$

Supondo $|R_{\tilde{x}}|, |R_{\tilde{y}}| \ll 1$, obtêm-se as seguintes fórmulas simplificadas para o erro relativo do produto e do quociente

$$R_{\widetilde{P}} \approx R_{\tilde{x}} + R_{\tilde{y}}$$

$$R_{\widetilde{Q}} \approx R_{\tilde{x}} - R_{\tilde{y}}$$

Algarismos significativos/casas decimais de precisão

Seja $x = \pm m_x \times 10^e$ um dado número (escrito na notação normalizada no sistema decimal) e seja $\tilde{x}$ uma sua aproximação.

Definição

- Diz-se que $\tilde{x}$ é uma aproximação para x com precisão de p casas decimais (c.d.) ou que $\tilde{x}$ aproxima x com p casas decimais (corretas), se
$$|x - \tilde{x}| \leq 0.5 \times 10^{-p}$$
- Diz-se que $\tilde{x}$ é uma aproximação para x comprecisão de q algarismos signficativos (a.s) ou que $\tilde{x}$ aproxima x com q algarismos significativos (corretos), se
$$|x - \tilde{x}| \leq 0.5 \times 10^{-q} \times 10^e.$$

Exemplo

- $x = 3.127,\ \tilde{x} = 3.123 \rightarrow e = 1$
$|x - \tilde{x}| = 0.4 \times 10^{-2} < 0.5 \times 10^{-2} = 0.5 \times 10^{-3} \times 10^1 \rightarrow$ 2 c.d.; 3 a.s.
- $x = 0.0003127,\ \tilde{x} = 0.0003123 \rightarrow e = -3$
$|x - \tilde{x}| = 0.4 \times 10^{-6} < 0.5 \times 10^{-6} = 0.5 \times 10^{-3} \times 10^{-3} \rightarrow$ 6 c.d.; 3 a.s.

Nota

Se $\tilde{x}$ é uma aproximação para x com p casas decimais corretas, então $\tilde{x}$ tem precisão de

$$q = p + e$$

algarismos significativos.

Algarismos significativos/erro relativo

- O número de casas decimais corretas está ligado ao erro absoluto.
- O número de algarismos significativos corretos está relacionado com o erro relativo. Com efeito:
  - Se $|R_{\tilde{x}}| \leq 0.5 \times 10^{-q}$, então $\tilde{x}$ tem q a.s. corretos.
$$\left|\frac{x - \tilde{x}}{x}\right| \leq 0.5 \times 10^{-q} \Rightarrow |x - \tilde{x}| \leq 0.5 \times 10^{-q} |m_x 10^e| < 0.5 \times 10^{-q} \times 10^e$$
  - Se $\tilde{x}$ é uma aproximação para x com q a.s. corretos, então
$$|R_{\tilde{x}}| = \left|\frac{x - \tilde{x}}{x}\right| \leq \frac{0.5 \times 10^{-q} \times 10^e}{|x|} \leq \frac{0.5 \times 10^{-q} \times 10^e}{0.1 \times 10^e} = 0.5 \times 10^{1-q}$$

No MATLAB, dado $x \in R_F$, tem-se

$$|R_{fl(x)}| \leq 2^{-53} \approx 1.1102 \times 10^{-16} < 0.5 \times 10^{-15}$$

Logo, $fl(x)$ tem (no mínimo) precisão de 15 algarismos significativos.

Cancelamento subtrativo

Sejam

$$x = 7.6545424, \quad y = 7.6544199$$

e consideremos as seguintes aproximações (com precisão de 7 a.s.)

$$\tilde{x} = 7.6545421, \quad \tilde{y} = 7.6544200.$$

Quantos a.s. corretos tem $\tilde{z} = \tilde{x} - \tilde{y}$ como aproximação para $z = x - y$?
Tem-se

$$z = x - y = 0.0001225 = 0.1225 \times 10^{-3}$$
$$\tilde{z} = \tilde{x} - \tilde{y} = 0.0001221 = 0.1221 \times 10^{-3}$$

Assim, $\tilde{z}$ é uma aproximação para z com apenas 3 a.s.corretos.
Isto significa que o erro relativo em $\tilde{z}$ pode ser $10^4 = 10\,000$ superior aos erros relativos em $\tilde{x}$ e $\tilde{y}$.

Cancelamento subtrativo

O efeito da perda de precisão de a.s. na subtração de dois números muito próximos (e consequente aumento do erro relativo) constitui o chamado problema do cancelamento subtrativo.

O cancelamento subtrativo deve ser evitado tanto quanto possível.

Exemplo

Para x grande, tem-se $\sqrt{x+1} \approx \sqrt{x}$, pelo que haverá problemas de cancelamento subtrativo no cálculo de $\sqrt{x+1} - \sqrt{x}$.
Mas,

$$\sqrt{x+1} - \sqrt{x} = \frac{1}{\sqrt{x+1} + \sqrt{x}}$$

A segunda fórmula constitui um processo alternativo de cálculo que não sofre de cancelamento subtrativo.
Por exemplo, se $x = 12000$, entao $\sqrt{x+1} \approx 109.5491$ (7 a.s.) e $\sqrt{x} \approx 109.5445$ (7 a.s.).

$$\sqrt{x+1} - \sqrt{x} \approx 109.5491 - 109.5445 = 0.0046$$
$$\frac{1}{\sqrt{x+1} + \sqrt{x}} \approx \frac{1}{109.5491 + 109.5445} = 0.004564260$$

Note-se que $\sqrt{12001} - \sqrt{12000} = 0.00456425955...$ pelo que a primeira aproximação tem apenas 2 a.s., enquanto a segunda tem 7 a.s.

Condicionamento de um problema

Exemplo1

Considere-se o problema da resolução do seguinte sistema de equações

$$\begin{cases} 1.01x + 0.99y = 2.00 \\ 0.99x + 1.01y = 2.00 \end{cases} \qquad \text{Sol} : x = 1; y = 1.$$

Modifiquemos ligeiramente o lado direito...

$$\begin{cases} 1.01x + 0.99y = 2.02 \\ 0.99x + 1.01y = 1.98 \end{cases} \qquad \text{Sol} : x = 2; y = 0.$$

$$\begin{cases} 1.01x + 0.99y = 1.98 \\ 0.99x + 1.01y = 2.02 \end{cases} \qquad \text{Sol} : x = 0; y = 2.$$

"Pequenas" alterações nos dados $\Longrightarrow$ "grandes alterações" nas soluções
$\longrightarrow$ Problema mal condicionado!

Exemplo 2 (Wilkinson)

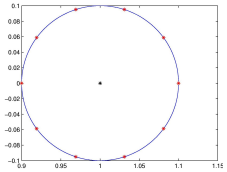
Considere-se o polinómio

$$p(z) = (z-1)^{10} = z^{10} - 10z^9 + 45z^8 - 120z^7 + 210z^6 - 252z^5 + 210z^4 - \\ - 120z^3 + 45z^2 - 10z + 1$$

p tem um único zero, $z = 1$ (de multiplicidade 10).
Seja

$$\tilde{p}(z) = z^{10} - 10z^9 + 45z^8 - 120z^7 + 210z^6 - 252z^5 + 210z^4 - 120z^3 \\ + 45z^2 - 10z + 1 - 10^{-10} = p(z) - 10^{-10}$$

$$\tilde{p}(z) = 0 \Leftrightarrow (z-1)^{10} - 10^{-10} = 0 \Leftrightarrow (z-1)^{10} = 10^{-10}$$
$$\Leftrightarrow z = \tilde{z}_k = 1 + 10^{-1}(\cos(\frac{2k\pi}{10}) + i\,\text{sen}(\frac{2k\pi}{10})); \; k = 0, 1, \ldots, 9.$$



Temos

$$|z - \tilde{z}_k| = \left|10^{-1}(\cos(\frac{2k\pi}{10}) + i\,\mathrm{sen}(\frac{2k\pi}{10}))\right| = 10^{-1} = \mathbf{10^9} \times 10^{-10}$$

$\hookrightarrow$ Problema mal condicionado

Condicionamento de um problema

Um problema diz-se mal condicionado se for muito sensível a pequenas alterações nos seus dados; caso contrário, diz-se bem condicionado.

Número de condição de uma função

Seja $\tilde{x}$ um valor aproximado para x com um erro relativo tal que que

$$|R_{\tilde{x}}| = \frac{|x - \tilde{x}|}{|x|} \ll 1.$$

Seja f uma função continuamente diferenciável numa vizinhança de x (que contém $\tilde{x}$). Pretendemos saber como o erro em x se "propaga" ao cálculo de $f(x)$. Por outras palavras, sendo $y = f(x)$ e $\tilde{y} = f(\tilde{x})$, pretendemos determinar a razão entre $R_{\tilde{y}}$ e $R_{\tilde{x}}$. Pelo teorema do valor médio, temos

$$y - \tilde{y} = f(x) - f(\tilde{x}) = f'(\xi)(x - \tilde{x}), \quad \xi \in (\min\{x, \tilde{x}\}, \max\{x, \tilde{x}\}).$$

Temos, então

$$|R_{\tilde{y}}| = \left|\frac{y - \tilde{y}}{y}\right| = \left|\frac{f'(\xi)(x - \tilde{x})}{f(x)}\right| = \left|\frac{xf'(\xi)}{f(x)}\right| . \left|\frac{x - \tilde{x}}{x}\right| = \left|\frac{xf'(\xi)}{f(x)}\right| |R_{\tilde{x}}| .$$

Número de condição de uma função

Como admitimos que x e $\tilde{x}$ estão próximos, será razoável substituir $f'(\xi)$ por $f'(x)$, tendo-se, então

$$|R_{f(\tilde{x})}| \approx \left|\frac{xf'(x)}{f(x)}\right| |R_{\tilde{x}}|$$

Assim, a quantidade $\left|\frac{xf'(x)}{f(x)}\right|$ é uma medida do condicionamento do cálculo do valor de f em x.

Definição

Chamamos número de condição de f em x e denotamos por $\mathrm{cond} f(x)$ à quantidade dada por

$$\mathrm{cond} f(x) = \left|\frac{xf'(x)}{f(x)}\right|$$

Exemplo

- $f(x) = \sqrt{x}$ $\quad f'(x) = \frac{1}{2\sqrt{x}}$ $\quad \mathrm{cond} f(x) = \frac{1}{2} \longrightarrow$ função bem condicionada para todo o x.
- $f(x) = e^x$ $\quad f'(x) = e^x$ $\quad \mathrm{cond} f(x) = |x| \longrightarrow$ função mal condicionada para valores de x tais que $|x|$ é "grande" e bem condicionda para valores de x tais que $|x|$ é "pequeno".

Estabilidade/instabilidade de um método

Suponhamos que pretendemos calcular os seguintes integrais

$$I_n := \int_0^1 \frac{x^n}{x+10} dx;\ n = 0, 1 \ldots, 10.$$

Método 1:

$$I_n = \int_0^1 \frac{x^{n-1}(x+10-10)}{x+10} dx = \int_0^1 x^{n-1} - 10 \int_0^1 \frac{x^{n-1}}{x+10} dx$$

$$\Downarrow$$

$$\boxed{I_n = \frac{1}{n} - 10 I_{n-1}}$$

- $I_0 = [\ln(x+10)]_0^1 = \ln(\frac{11}{10})$.
- Começando com uma aproximação $\tilde{I}_0 = 0.0953101798$ para I_0 (10 c.d. corretas) e usando a fórmula anterior, obtêm-se as aproximações apresentadas na coluna 2 da tabela dada à frente.

Método 2:

De $I_n = \frac{1}{n} - 10 I_{n-1}$ obtém-se a seguinte fórmula alternativa

$$\boxed{I_{n-1} = \frac{1}{10}\left(\frac{1}{n} - I_n\right)}$$

- Note-se que $I_n \to 0$ quando $n \to \infty$.
- Podemos, por exemplo, tomar $I_{20} \approx \tilde{I}_{20} = 0.00$ (apenas 2 c.d. corretas!) e usar esta nova fórmula para $n = 20, 19, 18, \ldots$. Nesse caso, obtêm-se valores aproximados listados na terceira coluna da tabela.

n	I_n (10 c.d.)	**Método 1**	**Método 2**
1	0.0468982020	0.0468982020	0.046898202
2	0.0310179804	0.0310179800	0.0310179804
3	0.0231535290	0.023153533	0.0231535290
4	0.0184647099	0.0184646667	0.0184647099
5	0.0153529009	0.0153533333	0.0153529009
6	0.0131376582	0.0131333333	0.0131376582
7	0.0114805609	0.0115238095	0.0114805609
8	0.0101943908	0.0097619054	0.0101943908
9	0.0091672034	0.0134920579	0.0091672034
10	0.0083279655	-0.0349205792	0.0083279655

Como explicar a diferença nos resultados ?

Analisemos primeiro o Método 1:

$$\tilde{I}_0 = I_0 + \epsilon$$

$$\tilde{I}_1 = 1 - 10\tilde{I}_0 = \underbrace{1 - 10 I_0}_{I_1} - 10\epsilon \Longrightarrow |\tilde{I}_1 - I_1| = 10|\epsilon|$$

De modo análogo

$$|\tilde{I}_2 - I_2| = 10 \times 10 \times |\epsilon| = 10^2 |\epsilon|$$

$$\vdots$$

$$|\tilde{I}_n - I_n| = 10^n |\epsilon|.$$

A fórmula escolhida faz com que o erro inicial $|\epsilon|$ se amplifique de cada vez que é usada. O processo de cálculo é dito instável.

Quando usamos o Método 2, tem-se

$$\tilde{I}_{20} = I_{20} + \delta$$

$$\tilde{I}_{19} = \frac{1}{10}\left(\frac{1}{20} - (I_{20} + \delta)\right) = \underbrace{\frac{1}{10}\left(\frac{1}{20} - I_{20}\right)}_{I_{19}} - \frac{\delta}{10} \Longrightarrow |\tilde{I}_{19} - I_{19}| = \frac{|\delta|}{10}.$$

De modo análogo se vê que

$$|\tilde{I}_{18} - I_{18}| = \frac{|\delta|}{10^2}$$

$$|\tilde{I}_{17} - I_{17}| = \frac{|\delta|}{10^3} \ \ldots$$

Assim, o erro inicial não tende a aumentar (neste caso, até diminui) à medida que a fórmula é usada. Trata-se de um processo de cálculo estável.

Estabilidade/instabilidade de um método

Dizemos que um método é instável se os erros se amplificam no decurso dos cálculos, de forma inaceitável; caso contrário, o método diz-se estável.

SISTEMAS de EQUAÇÕES LINEARES
Métodos Diretos

Métodos diretos e métodos iterativos

Os métodos numéricos para a resolução de sistemas são, geralmente, agrupados em duas grandes classes:

- Métodos diretos
- Métodos iterativos

Nos métodos diretos, a solução é determinada num número finito de operações aritméticas. Se não houvesse de erros de arbrownondamento, tais métodos produziriam a solução exata. Na prática, os erros de arbrownondamento são inevitáveis, pelo que os métodos conduzem apenas a soluções aproximadas.
Nos métodos iterativos, a partir de uma aproximação inicial para a solução, gera-se uma sequência de aproximações que, sob certas condições, converge para a solução do problema. Na prática, o processo será interrompido ao fim de um certo número de iterações.

Métodos diretos/métodos iterativos

Que tipo de método usar?

Não há regras rígidas. A escolha depende

- do problema em causa
- das facilidades computacionais disponíveis (a nível de *hardware* e *software*)
- da precisão exigida para os resultados, etc.

Em geral:

- se a matriz do sistema é de dimensão razoável e não há problemas com a capacidade de armazenamento de dados, usa-se um método direto;
- se a matriz é muito grande e esparsa (i.e. se for constituída, essencialmente, por zeros) poderá ser mais eficiente recorrer a um método iterativo.

Existem, também, métodos híbridos, os quais conjugam algumas características dos métodos diretos e dos métodos iterativos, mas o seu estudo está fora do âmbito deste curso.

Notações

Considere-se, então, um sistema de n equações lineares em n incógnitas escrito na forma matricial

$$Ax = b$$

onde

- $A = [a_{ij}]$ é uma matriz $n \times n$ (matriz do sistema)
- $b = [b_i]$ é um vetor coluna com n componentes (vetor dos termos independentes)
- $x = [x_i]$ é um vetor cujas componentes pretendemos determinar (vetor das incógnitas).

Seja ainda $[A|b]$ a matriz ampliada do sistema, isto é, a matriz $n \times (n+1)$ formada juntando a A uma última coluna com os elementos de b.

No que se segue, suporemos que o sistema em causa admite solução única.

Sistemas triangulares

Sistema triangular inferior

Suponhamos que a matriz do sistema é triangular inferior, isto é,

$$A = L = [\ell_{ij}], \quad \ell_{ij} = 0 \text{ para } j > i$$

ou seja, que a matriz ampliada do sistema é da forma

$$[L|b] = \left[\begin{array}{ccccc|c} \ell_{11} & 0 & 0 & \cdots & 0 & b_1 \\ \ell_{21} & \ell_{22} & 0 & \cdots & 0 & b_2 \\ \ell_{31} & \ell_{32} & \ell_{33} & \cdots & 0 & b_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & \ell_{nn} & b_n \end{array}\right]$$

Suponhamos, além disso, que $\ell_{ii} \neq 0; i = 1, \ldots, n$.

Substituição direta

A solução de tal sistema pode obter-se facilmente pelo processo da chamada **substituição direta**:

- $x_1 = \frac{b_1}{\ell_{11}}$
- $x_2 = (b_2 - \ell_{21}x_1)/\ell_{22}$
- $x_3 = (b_3 - (\ell_{31}x_1 + \ell_{32}x_2))/\ell_{33}$

  $\cdots$
- $x_n = \left(b_n - \sum_{j=1}^{n-1} \ell_{nj}x_j\right)/\ell_{nn}$

Tem-se:

- $x_1 = \frac{b_1}{\ell_{11}}$
- Para $i = 2, \ldots, n$:

$$x_i = \left(b_i - \sum_{j=1}^{i-1} \ell_{ij}x_j\right)/\ell_{ii}$$

Sistema triangular inferior

Algoritmo de substituição direta

função $x = \mathbf{subDireta}(L, b)$ (1)

% *Resolução de um sistema $Lx = b$, com L triangular inferior*

% ENTRADAS: – matriz L, triangular inferior de ordem n
– vetor b dos termos independentes

% SAÍDA: – vetor x, solução do sistema

$x(1) = b(1)/L(1,1)$ (2)

para $i = 2{:}n$

$$x(i) = \frac{b(i) - L(i, 1{:}i-1) * x(1{:}i-1)}{L(i,i)} \quad (3)$$

fim

Notas

1. No algoritmo anterior, estamos a admitir que o vetor x é inicializado como um vetor coluna. Assim, na fórmula (3), temos:

$$L(i,1:i-1)*x(1:i-1) = \begin{bmatrix} L(i,1) & L(i,2) & \cdots & L(i,i-1) \end{bmatrix} \begin{bmatrix} x(1) \\ x(2) \\ \vdots \\ x(i-1) \end{bmatrix}$$
$$= L(i,1)x(1) + \cdots + L(i,i-1)x(i-1)$$

2. Como, para cada i, $b(i)$ entra apenas no cálculo de $x(i)$, para não aumentar as necessidades de armazenamento, as incógnitas $x(i)$ que vão sendo calculadas podem ir sendo sobrepostas aos valores de $b(i)$, i.e. as linhas (1)–(3) do algoritmo podem ser substituídas por (1')–(3'):

função $b = \mathbf{subDireta}(L, b)$ (1')

$b(1) = b(1)/L(1,1)$ (2')

$$b(i) = \frac{b(i) - L(i,1:i-1)*b(1:i-1)}{L(i,i)} \quad (3')$$

Neste caso, à saída, o vetor b contém a solução do sistema.

Algoritmo da substituição direta

Complexidade

Para ter uma ideia da complexidade computacional de um algoritmo, é usual efetuar-se uma contagem do número de operações de ponto flutuante (*flops*) envolvidas no seu uso. Tradicionalmente, distinguem-se as adições/subtrações das multiplicações/divisões.

Número de operações para o algoritmo de substituição direta

- **Multiplicações/divisões**: i para cada $i = 1, \ldots, n$, ou seja

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2} = \frac{n^2+n}{2}$$

- **Adições/subtrações**: $i-1$ para cada $i = 2, \ldots, n$, ou seja,

$$\sum_{i=2}^{n} (i-1) = \sum_{i=1}^{n-1} i = \frac{(n-1)n}{2} = \frac{n^2-n}{2}$$

O número total de operações é $N_{OP} = n^2$.

Sistema triangular superior

Substituição inversa

Um algoritmo semelhante ao anterior, no caso em que $A = U = [u_{ij}]$ é triangular superior, será o seguinte:

Algoritmo de substituição inversa

função $b = \mathbf{subInversa}(U, b)$

% *Resolução de um sistema $Ux = b$, com U triangular superior*

% ENTRADAS: – matriz U, triangular superior de ordem n
– vetor b dos termos independentes

% SAÍDA: – vetor b, solução do sistema

$b(n) = b(n)/U(n,n)$

para $i = n-1:-1:1$

$$b(i) = \frac{b(i) - U(i,i+1:n)*b(i+1:n)}{U(i,i)}$$

fim

Método de Gauss

Recorde ...

- Operações elementares sobre linhas de uma matriz:
  - troca de linhas
  - multiplicação de uma linha por um escalar diferente de zero
  - adição a uma linha de outra linha multiplicada por um escalar
- Dada a matriz ampliada de um sistema, quaisquer operações elem. sobre as suas linhas transformam-na na matriz de um sistema equivalente (i.e. com as mesmas soluções).

Método de Gauss

1. *Eliminação de Gauss*

$$[A|b] \overset{\text{op. elem.}}{\longrightarrow} [U|\beta], \quad U \text{ triang. superior}$$

2. *Substituição inversa*
Resolver o sistema de matriz $[U|\beta]$ por substituição inversa.

Método de Gauss

Exemplo

Resolver o sistema de matriz ampliada

$$[A|b] = \left[\begin{array}{ccc|c} 4 & -9 & 2 & 2 \\ 2 & -4 & 4 & 3 \\ -1 & 2 & 2 & 1 \end{array}\right]$$

Passo 1 Anular os coeficientes da incógnita x_1 da $2^{\underline{a}}$ e $3^{\underline{a}}$ linhas (subtraindo dessas linhas múltiplos adequados da $1^{\underline{a}}$ linha).

$$L_2 \leftarrow L_2 - m_{21}L_1, \quad m_{21} = \frac{a_{21}}{a_{11}} = 0.5$$
$$L_3 \leftarrow L_3 - m_{31}L_1, \quad m_{31} = \frac{a_{31}}{a_{11}} = -0.25$$

$$\left[\begin{array}{ccc|c} 4 & -9 & 2 & 2 \\ 2 & -4 & 4 & 3 \\ -1 & 2 & 2 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} 4 & -9 & 2 & 2 \\ 0 & 0.5 & 3 & 2 \\ 0 & -0.25 & 2.5 & 1.5 \end{array}\right]$$

Exemplo (cont.)

Passo 2 anular o coeficiente da incógnita x_2 da $3^{\underline{a}}$ linha (subtraindo a essa linha um múltiplo adequado da $2^{\underline{a}}$ linha).

$$L_3 \leftarrow L_3 - m_{32}L_2, \qquad m_{32} = a_{32}/a_{22} = -0.25/0.5 = -0.5$$

$$\left[\begin{array}{ccc|c} 4 & -9 & 2 & 2 \\ 0 & 0.5 & 3 & 2 \\ 0 & -0.25 & 2.5 & 1.5 \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} 4 & -9 & 2 & 2 \\ 0 & 0.5 & 3 & 2 \\ 0 & 0 & 4 & 2.5 \end{array}\right]$$

Temos a matriz ampliada de um sistema, equivalente ao primeiro, na forma $[U|\beta]$, com U triangular superior ($\Rightarrow$ terminou a parte de eliminação Gaussiana).

- Este sistema triangular pode, agora, ser resolvido por substituição inversa:

$$4x_3 = 2.5 \Rightarrow x_3 = 0.625$$
$$0.5x_2 + 3x_3 = 2 \Rightarrow 0.5x_2 + 3 \times 0.625 = 2 \Rightarrow x_2 = (2 - 3 \times 0.625)/0.5 = 0.25$$
$$4x_1 - 9x_2 + 2x_3 = 2 \Rightarrow 4x_1 - 9 \times 0.25 + 2 \times 0.625 = 2 \Rightarrow x_1 = 0.75$$

Método de Gauss

Consideremos, então, a matriz ampliada de um sistema

$$[A|b] = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array}\right]$$

Passo 1 Para $i = 2, \ldots, n$, subtrai-se da linha i a linha 1 multiplicada por

$$\boxed{m_{i1} = a_{i1}/a_{11}}$$

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array}\right] \longrightarrow \left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} & b_n^{(1)} \end{array}\right]$$

$$\boxed{a_{ij}^{(1)} = a_{ij} - m_{i1}a_{1j}, \qquad b_i^{(1)} = b_i - m_{i1}b_1}$$

Método de Gauss (cont.)

Passo 2 Para $i = 3, \ldots, n$, subtrai-se da linha i a linha 2 multiplicada por

$$\boxed{m_{i2} = a_{i2}^{(1)}/a_{22}^{(1)}}$$

$$\left[\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} & b_1 \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & \cdots & a_{3n}^{(1)} & b_3^{(1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & a_{n3}^{(1)} & \cdots & a_{nn}^{(1)} & b_n^{(1)} \end{array}\right] \rightarrow \left[\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} & b_1 \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} & b_3^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} & b_n^{(2)} \end{array}\right]$$

onde

$$\boxed{a_{ij}^{(2)} = a_{ij}^{(1)} - m_{i2}a_{2j}^{(1)}, \qquad b_i^{(2)} = b_i^{(1)} - m_{i2}b_2^{(1)}}$$

Método de Gauss (cont.)

No final do passo $k-1$ (isto é, antes do início do passo k) teremos uma matriz da forma

$$\left[\begin{array}{ccccccc|c} a_{11} & a_{12} & \cdots & a_{1k} & a_{1,k+1} & \cdots & a_{1n} & b_1 \\ 0 & a_{22}^{(1)} & \cdots & a_{2k}^{(1)} & a_{2,k+1}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{kk}^{(k-1)} & a_{k,k+1}^{(k-1)} & \cdots & a_{kn}^{(k-1)} & b_k^{(k-1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{nk}^{(k-1)} & a_{n,k+1}^{(k-1)} & \cdots & a_{nn}^{(k-1)} & b_n^{(k-1)} \end{array}\right]$$

Nota: Quando $k=1$, tem-se, naturalmente, $a_{ij}^{(0)} \equiv a_{ij}$.

Método de Gauss (cont.)

Passo k Para $i = k+1, \ldots, n$, subtrai-se da linha i a linha k multiplicada por

$$m_{ik} = a_{ik}^{(k-1)} / a_{kk}^{(k-1)}$$

obtendo-se a matriz

$$\left[\begin{array}{ccccccc|c} a_{11} & a_{12} & \cdots & a_{1k} & a_{1,k+1} & \cdots & a_{1n} & b_1 \\ 0 & a_{22}^{(1)} & \cdots & a_{2k}^{(1)} & a_{2,k+1}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{kk}^{(k-1)} & a_{k,k+1}^{(k-1)} & \cdots & a_{kn}^{(k-1)} & b_k^{(k-1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & a_{n,k+1}^{(k)} & \cdots & a_{nn}^{(k)} & b_n^{(k)} \end{array}\right]$$

onde

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - m_{ik} a_{kj}^{(k-1)}, \qquad b_i^{(k)} = b_i^{(k-1)} - m_{ik} b_k^{(k-1)}$$

Método de Gauss (cont.)

Ao fim de $n-1$ passos, a matriz do sistema estará reduzida à forma triangular superior, isto é, ter-seá a seguinte matriz ampliada

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{nn}^{(n-1)} & b_n^{(n-1)} \end{array}\right]$$

O sistema correspondente poderá, então, resolver-se pelo método de substituição inversa.

Método de Gauss
Notas

1. Os números $m_{ik} = a_{ik}^{(k-1)} / a_{kk}^{(k-1)}$ são chamados multiplicadores.
2. Os elementos $a_{kk}^{(k-1)}$ são chamados pivôs.
3. No processo que acabámos de descrever, admitimos implicitamente que os pivôs (que são usados como divisores) são não nulos; tal não se verifica necessariamente, mesmo sendo o sistema possível e determinado, devendo o algoritmo ser modificado caso algum dos pivôs seja zero; veremos também, mais à frente que, se um dos pivôs for "pequeno", poderá haver problemas de instabilidade no algoritmo.

Algoritmo do processo de eliminação de Gauss

função $[A, b] = \mathbf{elimGauss}(A, b)$
% *Redução de um sistema $Ax = b$ à forma triangular*
% ENTRADAS: – matriz A de ordem n (matriz simples do sistema)
– vetor b dos termos independentes
% SAÍDAS: – matriz A triangular superior (ver Nota 3)
– vetor b modificado

para $k = 1 : n-1$ % Contagem do passo
% Criação do vetor mult com os multiplicadores (para as linhas $k+1$ a n)
$\ell = k+1 : n$
$\text{mult} = A(\ell, k)/A(k, k)$ (1)
% Modificação da submatriz $A(k+1 : n, k+1 : n)$ e de $b(k+1 : n)$
$j = k+1 : n$
$A(\ell, j) = A(\ell, j) - \text{mult} * A(k, j)$ (2)
$b(\ell) = b(\ell) - \text{mult} * b(k)$ (3)
fim

Notas

1. Como mostram as linhas (2) e (3) do algoritmo, este foi escrito pressupondo que os sucessivos valores $a_{ij}^{(k)}$ e $b_i^{(k)}$ calculados durante o processo são sobrepostos aos valores anteriores, isto é, são armazenados nos "espaços" correspondentes aos valores iniciais a_{ij} e b_i.
2. Ter em conta que o produto $\text{mult} * A(k, j)$ referido na linha (2) do algoritmo corresponde a

$$\begin{bmatrix} m_{k+1,k} \\ \vdots \\ m_{n,k} \end{bmatrix} [a_{k,k+1} \cdots a_{k,n}] = \begin{bmatrix} m_{k+1,k}a_{k,k+1} & \cdots & m_{k+1,k}a_{k,n} \\ \vdots & \vdots & \vdots \\ m_{n,k}a_{k,k+1} & \cdots & m_{n,k}a_{k,n} \end{bmatrix}$$

3. No algoritmo, não tornámos explicitamente iguais a zero os elementos abaixo da diagonal; assim, à saída, A não é, na realidade, uma matriz triangular superior; para usar, de seguida, o aloritmo de substituição inversa, devemos começar por aplicar a função **triu** à matriz A resultante da aplicação deste algoritmo.

Notas

4. Caso estejamos interessados em guardar os diversos multiplicadores usados no processo de redução, poderemos aproveitar a "parte triangular inferior" de A para esse efeito; neste caso, deveremos substituir as linhas (1)–(3) do algoritmo anterior, respetivamente pelas linhas (1')–(3') seguintes:

$$A(\ell, k) = A(\ell, k)/A(k, k) \quad (1')$$

$$A(\ell, j) = A(\ell, j) - A(\ell, k) * A(k, j) \quad (2')$$

$$b(\ell) = b(\ell) - A(\ell, k) * b(k) \quad (3')$$

Complexiade do algorimo de eliminação de Gauss

Número de *flops* para o algoritmo de elim. de Gauss aplicado a $[A|b]$

- **Multiplicações/divisões**: $(n-k) + (n-k)(n-k+1) = (n-k)(n-k+2)$ para cada $k = 1, \ldots, n-1$, ou seja

$$\sum_{k=1}^{n-1} (n-k)(n-k+2) = \frac{2n^3 + 3n^2 - 5n}{6}$$

- **Adições/subtrações**: $(n-k)(n-k+1)$ para cada $k = 1, \ldots, n-1$, ou seja,

$$\sum_{k=1}^{n-1} (n-k)(n-k+1) = \frac{n^3 - n}{3}$$

Podemos, portanto, concluir que o número total de operações é

$$N_{OP} = \frac{4n^3 + 3n^2 - 7n}{6} = \mathcal{O}(n^3)$$

Instabilidade na eliminação de Gauss

Necessidade de escolha de pivô

Exemplo

Considere-se o problema da resolução do seguinte sistema

$$\begin{cases} 10^{-12}x_1 + x_2 = 1 \\ x_1 - x_2 = 0 \end{cases}$$

Efetuando eliminação exata na matriz ampliada, virá

$$\left[\begin{array}{cc|c} 10^{-12} & 1 & 1 \\ 1 & -1 & 0 \end{array}\right] \longrightarrow \left[\begin{array}{cc|c} 10^{-12} & 1 & 1 \\ 0 & -10^{12}-1 & -10^{12} \end{array}\right],$$

ou seja, tem-se a seguinte solução

$$x_2 = \frac{10^{12}}{10^{12}+1}, \qquad x_1 = \frac{1-\frac{10^{12}}{10^{12}+1}}{10^{-12}} = \frac{10^{12}}{10^{12}+1}$$

Note-se que $x_1 = x_2 \approx 1$.

Exemplo (cont.)

Considere-se a resolução do mesmo sistema numa máquina com sistema de numeração $F(10, 12, -99, 99)$.
Note-se que, nesta máquina, temos

$$fl(-10^{12}-1) = fl(-(10^{12}+1)) = -10^{12}.$$

Assim, neste caso, o processo de eliminação conduzirá a:

$$\left[\begin{array}{cc|c} 10^{-12} & 1 & 1 \\ 0 & -10^{12} & -10^{12} \end{array}\right]$$

de onde se obtém a solução aproximada

$$\tilde{x_2} = 1, \quad 10^{-12}\tilde{x_1} = 0 \Longrightarrow \tilde{x_1} = 0 \text{ !!}$$

Eliminação de Gauss/decomposição LU

Vejamos como o processo de eliminção de Gauss aplicado a uma matriz A pode ser descrito matricialmente pela pré-multiplicação de A por matrizes adequadas.

Exemplo (Eliminação de Gauss/decomposição LU)

Seja A a matriz do nosso exemplo inicial da resolução de um sistema pelo método de Gauss:

$$A = \begin{bmatrix} 4 & -9 & 2 \\ 2 & -4 & 4 \\ -1 & 2 & 2 \end{bmatrix}.$$

1. Recorde que os multiplicadores usados no 1º passo de redução foram

$$m_{21} = \frac{2}{4} = 0.5, \quad m_{31} = \frac{-1}{4} = -0.25.$$

Exemplo (Eliminação de Gauss/decomposição LU)

1. (cont.) Seja M_1 a matriz triangular inferior obtida da matriz identidade substituindo os elementos da 1ª coluna, abaixo da diagonal, pelos simétricos dos multiplicadores usados no 1º passo de redução:

$$M_1 = \begin{bmatrix} 1 & 0 & 0 \\ -m_{21} & 1 & 0 \\ -m_{31} & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0.25 & 0 & 1 \end{bmatrix}.$$

Facilmente se verifica que:

- o produto de M_1 por A corresponde a efetuar o primeiro passo de redução (apenas sobre a matriz A):

$$M_1 A = \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0.25 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & -9 & 2 \\ 2 & -4 & 4 \\ -1 & 2 & 2 \end{bmatrix} = \begin{bmatrix} 4 & -9 & 2 \\ 0 & 0.5 & 3 \\ 0 & -0.25 & 2.5 \end{bmatrix}$$

Exemplo (Eliminação de Gauss/decomposição LU)

1. (cont.)
   - a matriz M_1 é invertível, sendo a sua inversa obtida trocando o sinal dos elementos abaixo da diagonal, ou seja,

$$M_1 = \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0.25 & 0 & 1 \end{bmatrix} \Rightarrow M_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ -0.25 & 0 & 1 \end{bmatrix}$$

2. O (único) multiplicador para o passo 2 é $m_{32} = -\frac{0.25}{0.5} = -0.5$. Seja M_2 a matriz obtida da identidade substituindo o elemento na posição (3,2) pelo simétrico de m_{32}, i.e. seja

$$M_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -m_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0.5 & 1 \end{bmatrix}.$$

Exemplo (Eliminação de Gauss/decomposição LU)

2. Facilmente se verifica que:
   - o segundo passo de redução pode ser descrito pela pré-multiplicação da matriz obtida no final do $1^{\underline{o}}$ passo pela matriz M_2, i.e.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0.5 & 1 \end{bmatrix} \begin{bmatrix} 4 & -9 & 2 \\ 0 & 0.5 & 3 \\ 0 & -0.25 & 2.5 \end{bmatrix} = \begin{bmatrix} 4 & -9 & 2 \\ 0 & 0.5 & 3 \\ 0 & 0 & 4 \end{bmatrix}.$$

   - a matriz M_2 é invertível, sendo a sua inversa obtida simplesmente trocando o sinal ao elemento na posição m_{32}:

$$M_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0.5 & 1 \end{bmatrix} \Rightarrow M_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -0.5 & 1 \end{bmatrix}.$$

3. Vemos, então, que o processo de redução de A à forma triangular superior (que, neste caso, envolveu apenas dois passos) pode ser descrito por $M_2 M_1 A = U$, onde U é a matriz triangular superior final.

Exemplo (Eliminação de Gauss/decomposição LU)

Mas,

$$M_2 M_1 A = U \Rightarrow A = \underbrace{M_1^{-1} M_2^{-1}}_{L} U$$

onde

$$L = M_1^{-1} M_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ -0.25 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -0.5 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ -0.25 & -0.5 & 1 \end{bmatrix}$$

Note-se que:

- L é triangular inferior com elementos diagonais iguais a 1 (dizemos que tem diagonal unitária);
- os elementos de L abaixo da diagonal são os multiplicadores usados no processo de redução.

O que vimos neste exemplo verifica-se em geral:

Eliminação de Gauss/decomposição LU

Seja A uma matriz quadrada de ordem n. Se for possível usar eliminação Gaussiana, sem escolha de pivô, para converter A numa matriz U triangular superior [a] e formarmos uma matriz triangular inferior $L = (\ell_{ik})$, com diagonal unitária e tendo, na parte estritamente triangular inferior, os multiplicadores m_{ik} usados no processo de redução, colocados nas posições respetivas i.e.

$$\ell_{ik} = m_{ik}; k = 1, \ldots, n-1; i = k+1, \ldots, n,$$

então estas matrizes L e U dão-nos uma fatorização de A na seguinte forma

$$A = LU$$

[a] Tal será o caso se e só se no processo de eliminação de Gauss não surgirem *pivôs* nulos.

Definição

Chama-se decomposição LU de uma matriz quadrada A à sua fatorização no produto de duas matrizes L e U com as caraterísticas das matrizes acima referidas, ou seja, tais que:

- L é triangular inferior com diagonal unitária;
- U é triangular superior.

Acabámos de ver que, se o processo de eliminação Gaussiana aplicado a uma matriz A puder prosseguir até ao fim (i.e. não surgirem pivôs nulos), então A admitirá uma decomposição LU. Reciprocamente, pode mostrar-se que, se A for invertível e admitir uma decomposição $A = LU$, então podemos converter A na matriz U, por eliminação Gaussiana, usando os elementos de L situados abaixo da diagonal como multiplicadores.

Conclusão:

Eliminação de Gauss/decomposição LU

Seja A uma matriz quadrada invertível. Então, a eliminação Gaussiana sem escolha de pivô aplicada a essa matriz é equivalente à decomposição LU dessa matriz.

Decomposição LU

Unicidade

Vamos considerar agora a decomposição LU de uma matriz A em mais pormenor, de forma independente da eliminção de Gauss, começando por demonstrar um teorema que estabelece a unicidade desta decomposição, caso ela exista.

Teorema

Seja A uma matriz quadrada de ordem n, invertível. Então, se A admite uma decomposição LU, essa decomposição é única.

Dem: Sejam $A = L_1U_1$ e $A = L_2U_2$ duas decomposições LU de A.

1. $$A \text{ invertível} \Longrightarrow L_1, L_2, U_1, U_2 \text{ invertíveis}$$

2. $$A = L_1U_1, \quad A = L_2U_2 \Rightarrow L_1U_1 = L_2U_2 \Rightarrow L_2^{-1}L_1 = U_2U_1^{-1}$$

3. Mas:

$$\left.\begin{array}{l} L_2^{-1}L_1 \text{ é triangular inferior} \\ U_2U_1^{-1} \text{ é triangular superior} \\ L_2^{-1}L_1 = U_2U_1^{-1} \end{array}\right\} \Rightarrow L_2^{-1}L_1 \text{ é uma matriz diagonal}$$

A diagonal de $L_2^{-1}L_1$ é unitária $\Rightarrow L_2^{-1}L_1 = I \Rightarrow L_2 = L_1$.

4. $L_2^{-1}L_1 = U_2U_1^{-1}, \quad L_2^{-1}L_1 = I \Rightarrow U_2U_1^{-1} = I \Rightarrow U_2 = U_1$.

Decomposição LU

Existência

O teorema seguinte estabelece condições de existência da decomposição LU de uma matriz.

Notação

No que se segue, dada uma matriz M, quadrada de ordem n, denotaremos por $M_k (1 \leq k \leq n)$ a submatriz de M contida nas suas primeiras k linhas e k colunas (por vezes designada por submatriz principal liderante de ordem k.)

Teorema

Seja $A = [a_{ij}]$ uma matriz quadrada de ordem n. Se as submatrizes A_k, para $k = 1, \ldots, n-1$, forem invertíveis, então A admite uma decomposição LU. Além disso, se A for inverível, a condição anterior é uma condição necessária para que A tenha uma decomposição LU.

Dem:

- **Condição suficiente** (Por indução sobre n)

1. Caso $n = 1$ Seja $A = [a_{11}]$. Então, A pode decompor-se como

$$A = [1][a_{11}].$$

2. Suponhamos que o resultado se verifica para qualquer matriz de ordem $k-1$ e vejamos que, então, se verifica para uma matriz de ordem k.
Seja A quadrada de ordem k e consideremo-la particionda em blocos como se segue

$$A = \left[\begin{array}{c|c} A_{k-1} & \mathbf{b} \\ \hline \mathbf{a}^{\mathrm{T}} & a_{kk} \end{array}\right],$$

onde $\mathbf{b} = [a_{1k}, \ldots, a_{k-1,k}]^{\mathrm{T}}$ e $\mathbf{a}^{\mathrm{T}} = [a_{k1}, \ldots, a_{k,k-1}]$. Por hipótese de indução, a matriz A_{k-1} admite uma decomposição LU, $A_{k-1} = L_{k-1}U_{k-1}$.

2. Sejam

$$L = \left[\begin{array}{c|c} L_{k-1} & \mathbf{0} \\ \hline \mathbf{l}^{\mathrm{T}} & 1 \end{array}\right] \quad \text{e} \quad U = \left[\begin{array}{c|c} U_{k-1} & \mathbf{u} \\ \hline \mathbf{0}^{\mathrm{T}} & u_{kk} \end{array}\right],$$

com $\mathbf{l}$ e $\mathbf{u}$ vetores com $k-1$ componentes. Então,

$$LU = \left[\begin{array}{c|c} L_{k-1}U_{k-1} & L_{k-1}\mathbf{u} \\ \hline \mathbf{l}^{\mathrm{T}}U_{k-1} & \mathbf{l}^{\mathrm{T}}\mathbf{u} + u_{kk} \end{array}\right]$$

A condição suficiente do teorema ficará demonstrada se mostrarmos que é possível encontrar vetores $\mathbf{l}$ e $\mathbf{u}$ e um escalar u_{kk} tais que

$$\begin{cases} L_{k-1}\mathbf{u} = \mathbf{b} \\ \mathbf{l}^{\mathrm{T}}U_{k-1} = \mathbf{a}^{\mathrm{T}} \\ \mathbf{l}^{\mathrm{T}}\mathbf{u} + u_{kk} = a_{kk} \end{cases}$$

2.

$$\exists\ \mathbf{u}, \mathbf{l}, u_{kk} \ \text{ tais que } \begin{cases} L_{k-1}\mathbf{u} = \mathbf{b} \\ \mathbf{l}^{\mathrm{T}}U_{k-1} = \mathbf{a}^{\mathrm{T}} \\ \mathbf{l}^{\mathrm{T}}\mathbf{u} + u_{kk} = a_{kk}. \end{cases} \ ?$$

Comecemos por notar que $A_{k-1} = L_{k-1}U_{k-1}$ invertível $\Rightarrow L_{k-1}, U_{k-1}$ invertíveis.

- L_{k-1} invertível $\Rightarrow$ sistema $L_{k-1}\mathbf{u} = \mathbf{b}$ tem solução (única) $\mathbf{u}$ dada por

$$\mathbf{u} = L_{k-1}^{-1}\mathbf{b}.$$

- U_{k-1} invertível $\Rightarrow \mathbf{l}^{\mathrm{T}}U_{k-1} = \mathbf{a}^{\mathrm{T}}$ tem solução $\mathbf{l}^{\mathrm{T}}$ dada por

$$\mathbf{l}^{\mathrm{T}} = \mathbf{a}^{\mathrm{T}}U_{k-1}^{-1}$$

- Calculados $\mathbf{l}$ e $\mathbf{u}$, u_{kk} vem dado por

$$u_{kk} = a_{kk} - \mathbf{l}^{\mathrm{T}}\mathbf{u}.$$

Assim, por indução, A admite uma decomposição LU.

- **Demonstração da condição necessária**

Suponhamos que A, invertível, admite uma decomposição LU. Seja

$$A = \left[\begin{array}{cc|cc} & A_k & * & * \\ & & * & * \\ \hline * & * & * & * \\ * & * & * & * \end{array}\right] = \left[\begin{array}{cc|cc} & L_k & \mathbf{O} & \\ \hline * & * & * & * \\ * & * & * & * \end{array}\right] \left[\begin{array}{cc|cc} & U_k & * & * \\ & & * & * \\ \hline & \mathbf{O} & * & * \\ & & * & * \end{array}\right]$$

a decomposição LU de A, com a partição indicada. A matriz L_k é triangular inferior de diagonal unitária, a matriz U_k é triangular superior, $\mathbf{O}$ designa a matriz nula (de ordem apropriada) e os asteriscos indicam submatrizes (não relevantes para a demonstração) de ordens apropriadas. Segue-se, de imediato, que $A_k = L_kU_k$. Mas,

$$\begin{aligned} A_k = L_kU_k &\Rightarrow \det A_k = \det L_k . \det U_k \\ &\Rightarrow \det A_k = \det U_k = u_{11}u_{22}\ldots u_{kk}. \end{aligned}$$

Sendo A invertível, tem-se $\det A = \det A_n = u_{11}\ldots u_{nn} \neq 0$ ou seja, temos que $u_{ii} \neq 0; i = 1, \ldots, n$. Isto garante que $\det A_k \neq 0$ ou seja, garante que A_k é invertível.

Método de Doolitle

Embora a eliminação Gaussiana e a decomposição LU sejam processos equivalentes, se considerarmos diretamente a decomposição LU de A obtemos um procedimento computacional diferente para determinar as matrizes L e U.
Seja

$$A = LU = \begin{bmatrix} 1 & & & & & \\ \vdots & \ddots & & & & \\ \ell_{i1} & \cdots & \ell_{i,i-1} & 1 & & \\ \vdots & & \vdots & & \ddots & \\ \ell_{n,1} & \cdots & \ell_{n,i-1} & \ell_{n,i} & \cdots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & \cdots & u_{1j} & \cdots & u_{1n} \\ & \ddots & \vdots & & \\ & & u_{jj} & \cdots & u_{jn} \\ & & & \ddots & \\ & & & & u_{nn} \end{bmatrix}$$

Então, temos:

- Para $i \leq j$
$$a_{ij} = \ell_{i1}u_{1j} + \ldots + \ell_{i,i-1}u_{i-1,j} + u_{ij}$$
- Para $i > j$
$$a_{ij} = \ell_{i1}u_{1j} + \ldots + \ell_{i,j-1}u_{j-1,j} + \ell_{ij}u_{jj}$$

Método de Doolitle

Para $i \leq j$

$$a_{ij} = \ell_{i1}u_{1j} + \ldots + \ell_{i,i-1}u_{i-1,j} + u_{ij}$$

Para $i > j$

$$a_{ij} = \ell_{i1}u_{1j} + \ldots + \ell_{i,j-1}u_{j-1,j} + \ell_{ij}u_{jj}$$

Estas equações podem ser usadas para obter as incógnitas u_{ij} (para $i \leq j$) e ℓ_{ij} (para $i > j$), desde que se ordenem estas incógnitas convenientemente.
Suponhamos que já determinámos as primeiras $i-1$ linhas de U e as primeiras $i-1$ colunas de L; podemos, então, determinar os elementos que formam a linha i de U e a coluna i de L, do seguinte modo:

- $u_{ij} = a_{ij} - \ell_{i1}u_{1j} - \ell_{i2}u_{2j} - \ldots - \ell_{i,i-1}u_{i-1,j}, \quad i \leq j,$
- $\ell_{ij} = \left[a_{ij} - \ell_{i1}u_{1j} - \ell_{i2}u_{2j} - \ldots - \ell_{i,j-1}u_{j-1,j}\right]/u_{jj}, \quad i > j.$

Esta forma de determinar a decomposição LU de A constitui o chamado Algoritmo de Doolitle.[4]

[4] Existem outras formas de organizar os cálculos que fornecem variantes deste algoritmo.

Exemplo (Algoritmo de Doolitle)

Seja

$$A = \begin{bmatrix} 2 & 2 & -2 \\ 1 & 5 & 7 \\ -1 & 1 & 6 \end{bmatrix}.$$

Usemos as fórmulas anteriores para obter a decomposição LU de A.

- Linha 1 de U:
  $u_{11} = a_{11} = 2; \quad u_{12} = a_{12} = 2; \quad u_{13} = a_{13} = -2$
- Coluna 1 de L:
  $\ell_{21} = a_{21}/a_{11} = \frac{1}{2}; \quad \ell_{31} = a_{31}/a_{11} = -\frac{1}{2}$
- Linha 2 de U:
  $u_{22} = a_{22} - \ell_{21}u_{12} = 5 - \frac{1}{2} \times 2 = 4; \quad u_{23} = a_{23} - \ell_{21}u_{13} = 7 - \frac{1}{2} \times (-2) = 8$
- Coluna 2 de L:
  $\ell_{32} = (a_{32} - \ell_{31}u_{12})/u_{22} = (1 - (-\frac{1}{2}) \times 2)/4 = \frac{1}{2}$
- Linha 3 de U:
  $u_{33} = a_{33} - \ell_{31}u_{13} - \ell_{32}u_{23} = 6 - (-\frac{1}{2}) \times (-2) - \frac{1}{2} \times 8 = 1.$

Exemplo (cont.)

Assim, tem-se a seguinte decomposição LU da matriz A:

$$A = \begin{bmatrix} 2 & 2 & -2 \\ 1 & 5 & 7 \\ -1 & 1 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ -1/2 & 1/2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 2 & -2 \\ 0 & 4 & 8 \\ 0 & 0 & 1 \end{bmatrix}$$

Resolução de sistemas usando a decomposição LU

Suponhamos, agora, que pretendemos resolver um sistema da forma $Ax = b$ e que conhecemos a decomposição LU da matriz A desse sistema. Tem-se

$$\begin{aligned} Ax = b &\iff (LU)x = b \\ &\iff L(\underbrace{Ux}_{y}) = b \\ &\iff \begin{cases} Ly = b \\ Ux = y \end{cases} \end{aligned}$$

Para resolvermos o sistema $Ax = b$, com $A = LU$, podemos:

1. resolver o sistema triangular inferior $Ly = b$ (por substituição direta)
2. resolver, de seguida, o sistema triangular superior $Ux = y$ (por substituição inversa).

Este processo é particularmente útil quando pretendemos resolver vários sistemas de equações que tenham todos a mesma matriz simples; por exemplo, este será o caso quando procurarmos determinar a inversa de uma matriz A.

Decomposição LU/Escolha de pivô

Ao descrevermos a equivalência entre a eliminação de Gauss e decomposição LU de uma matriz A, admitimos não haver necessidade de troca de linhas. Como sabemos, a troca de linhas no algoritmo é indispensável por questões de estabilidade.

Exemplo

Seja

$$A = \begin{bmatrix} 3 & 17 & 10 \\ 2 & 4 & -2 \\ 6 & 18 & -12 \end{bmatrix}$$

e efetuemos a eliminação de Gauss sobre A, com escolha parcial de pivô. Então, o primeiro passo de redução poderá ser descrito do seguinte modo:

$$\begin{bmatrix} 3 & 17 & 10 \\ 2 & 4 & -2 \\ 6 & 18 & -12 \end{bmatrix} \xrightarrow[\text{linha } 1 \leftrightarrow \text{linha } 3]{} \begin{bmatrix} 6 & 18 & -12 \\ 2 & 4 & -2 \\ 3 & 17 & 10 \end{bmatrix} \longrightarrow \begin{bmatrix} 6 & 18 & -12 \\ 0 & -2 & 2 \\ 0 & 8 & 16 \end{bmatrix}$$

sendo os multicadores usados: $m_{21} = \frac{2}{6} = \frac{1}{3}$ e $m_{31} = \frac{3}{6} = \frac{1}{2}$.

Exemplo (cont.)

O segundo passo de redução será então:

$$\begin{bmatrix} 6 & 18 & -12 \\ 0 & -2 & 2 \\ 0 & 8 & 16 \end{bmatrix} \xrightarrow[\text{linha } 2 \leftrightarrow \text{linha } 3]{} \begin{bmatrix} 6 & 18 & -12 \\ 0 & 8 & 16 \\ 0 & -2 & 2 \end{bmatrix} \longrightarrow \underbrace{\begin{bmatrix} 6 & 18 & -12 \\ 0 & 8 & 16 \\ 0 & 0 & 6 \end{bmatrix}}_{U},$$

sendo o multicador usado: $m_{32} = -\frac{2}{8} = -\frac{1}{4}$.
Seja U a matriz triangular superior obtida no final do processo de redução e formemos a matriz L, triangular inferior e de diagonal unitária, do seguinte modo:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ m_{31} & 1 & 0 \\ m_{21} & m_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{3} & -\frac{1}{4} & 1 \end{bmatrix}$$

De notar que os multiplicadores m_{21} e m_{31} estão trocados, em relação ao modo "usual" de obter L ($\Leftarrow$ após o passo 1, procedeu-se à troca das linhas 2 e 3).

Exemplo (cont.)

Formemos o produto LU:

$$\begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{3} & -\frac{1}{4} & 1 \end{bmatrix} \begin{bmatrix} 6 & 18 & -12 \\ 0 & 8 & 16 \\ 0 & 0 & 6 \end{bmatrix} = \begin{bmatrix} 6 & 18 & -12 \\ 3 & 17 & 10 \\ 2 & 4 & -2 \end{bmatrix}.$$

Que relação tem a matriz assim obtida com a matriz inicial

$$A = \begin{bmatrix} 3 & 17 & 10 \\ 2 & 4 & -2 \\ 6 & 18 & -12 \end{bmatrix} ?$$

É imediato concluir que o produto LU resulta de A por troca das suas linhas 1 e 3 seguida da troca das linhas 2 e 3 ($\Leftrightarrow$ trocas de linhas usadas no processo de eliminação).

Matrizes de permutção

Definição

Uma matriz P é uma matriz de permutação se for obtida da matriz identidade I por troca de linhas.

1. A pré-multiplicação de uma matriz A por uma matriz de permutação P, "troca", em A, as mesmas linhas que foram trocadas para obter P a partir de I.
2. Uma matriz de permutação P é invertível, tendo-se $P^{-1} = P$, i.e. $PP = I$.

Exemplo

Seja $P = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ e $A = \begin{bmatrix} 3 & 17 & 10 \\ 2 & 4 & -2 \\ 6 & 18 & -12 \end{bmatrix}$. Então:

$$PA = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 3 & 17 & 10 \\ 2 & 4 & -2 \\ 6 & 18 & -12 \end{bmatrix} = \begin{bmatrix} 6 & 18 & -12 \\ 3 & 17 & 10 \\ 2 & 4 & -2 \end{bmatrix}.$$

Matrizes de permutção

Armazenamento

Note-se que, num computador, uma matriz de permutação de ordem n não necessita de ser armazenada explicitamente, havendo apenas que armazenar um vetor de ordem n, correspondente ao ordenamento das linhas.
Por exemplo, sendo

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

bastaria armazenar o seguinte vetor de permutação $p = [3\ 1\ 2]$.
A matriz produto $B = PA$ seria, então, obtida (em MATLAB) simplesmente do seguinte modo:

$$B = A(p, :).$$

Exemplo (cont.)

Retomando o exemplo que estávamos a examinar, vemos que:

a eliminação de Gauss, com escolha parcial de pivô, aplicada à matriz A, forneceu uma decomposição LU, não da matriz A, mas sim da matriz PA, onde P é a matriz de permutação correspondente às trocas de linhas efetuadas ao longo da eliminação.

O que verificámos neste exemplo, é válido em geral, ou seja, tem-se o seguinte resultado (cuja demonstração pode ser vista, e.g. em G. H. Golub e C. F. van Loan, *Matrix Computations*, pp. 113-114).

Eliminação Gauss com escolha parcial de pivô e decomposição LU

Dada uma matriz não singular A, a redução de A à forma triangular por eliminação de Gauss com escolha parcial de pivô fornece uma decomposição

$$LU = PA$$

onde

1. a matriz U é a matriz triangular superior obtida no final do processo de redução;
2. a matriz L é a matriz triangular inferior, de diagonal unitária, e tendo, na parte estritamente triangular inferior os multiplicadores usados no processo de redução, havendo o cuidado de, na coluna k, trocar a ordem de elementos nas posições $k+1$ a n, sempre que, nos passos $k+1$ a n, houver alguma troca de linhas;
3. P é a matriz de permutação, obtida da matriz identidade efetuando as trocas de linhas correspondente às sescolhas dos pivôs.

Sendo A uma matriz invertível, é sempre possível converter A na forma triangular superior por eliminação de Gauss com escolha parcial de pivô. Assim, podemos concluir que, dada uma matriz invertível A, existe sempre uma decomposição LU de uma matriz obtida de A com permutação adequada das suas linhas, i.e. existem matrizes L, U e P (com L triangular inferior de diagonal untária, U triangular superior e P de permutação) tais que

$$LU = PA.$$

Suponhamos, então, que se pretende resolver um sistema de equações $Ax = b$ e que dispomos da decomposição LU de PA, com P uma matriz de permutação Neste caso, temos

$$\begin{aligned} Ax = b &\Leftrightarrow P(Ax) = Pb \Leftrightarrow (PA)x = Pb \\ &\Leftrightarrow (LU)x = Pb \Leftrightarrow L(\underbrace{Ux}_{y}) = Pb \\ &\Leftrightarrow \begin{cases} Ly = Pb \\ Ux = y \end{cases} \end{aligned}$$

Assim, podemos novamente encontrar a solução do sistema $Ax = b$ resolvendo dois sistemas triangulares, devendo, neste caso, começar por resolver-se o sistema $Ly = Pb$.

Matrizes Definidas Positivas

Definição

Uma matriz real simétrica diz-se definida positiva se

$$\forall x \in \mathbb{R}^n,\ x \neq 0, \qquad x^{\mathrm{T}} A x > 0.$$

Propriedades das matrizes definidas positivas

Seja $A = [a_{ij}]$ uma matriz real, simétrica definida positiva. Então:

1. $a_{ii} > 0; i = 1, \ldots, n.$
2. Os valores próprios de A são positivos. (*)
3. $\det A > 0.$
4. Qualquer submatriz principal A_k contida nas primeiras k linhas e k colunas de A; $k = 1, \ldots, n$, é definida positiva.
5. Qualquer submatriz principal A_k contida nas primeiras k linhas e k colunas de A tem determinante positivo. (*)

Matrizes definidas positivas

Nota: Cada uma das condições assinaladas com (*) (propriedades **2** e **5**) constitui condição suficiente para que A seja definida positiva.

Teorema

Seja A uma matriz real, simétrica, definida positiva. Então, A admite uma decomposição na forma

$$A = LDL^{\mathrm{T}}$$

onde L é uma matriz triangular inferior de diagonal unitária e D é uma matriz diagonal de elementos diagonais positivos.

Decomposição LDL^{T}

Dem:

1. A d.p. $\Rightarrow$ A admite uma única decomposição $A = LU$ onde
   - L triangular inferior de diagonal unitária
   - U triangular superior. (PORQUÊ?)
2. Seja D a matriz diagonal cuja diagonal é a da matriz U, i.e, seja

   $$D = \mathrm{diag}(u_{11}, \ldots, u_{nn}).$$

   Recordadndo qual o efeito da pré-multiplicação de uma matriz por uma matriz diagonal, vemos que a matriz U se pode escrever como $U = DV$ onde $V = [v_{ij}]$ é a matriz cujos elementos são dados por $v_{ij} = \frac{u_{ij}}{u_{ii}}$, ou seja

   $$V = \begin{bmatrix} \frac{u_{11}}{u_{11}} & \frac{u_{12}}{u_{11}} & \cdots & \frac{u_{1n}}{u_{11}} \\ & \frac{u_{22}}{u_{22}} & \cdots & \frac{u_{2n}}{u_{22}} \\ & & \ddots & \\ & & & \frac{u_{nn}}{u_{nn}} \end{bmatrix}$$

   Note-se que V é triangular superior (como U) e de diagonal unitária.

Temos, então, $A = LDV$ com L triangular inferior, U triangular superior (ambas de diagonal unitária) e D diagonal.

3. $$\begin{aligned} A \text{ simétrica} &\Rightarrow LDV = (LDV)^{\mathrm{T}} \\ &\Rightarrow LDV = V^{\mathrm{T}} D^{\mathrm{T}} L^{\mathrm{T}} \\ &\Rightarrow LDV = V^{\mathrm{T}} D L^{\mathrm{T}} \\ &\Rightarrow L(DV) = V^{\mathrm{T}} (DL^{\mathrm{T}}) \\ &\Rightarrow L = V^{\mathrm{T}} \Rightarrow L^{\mathrm{T}} = V \end{aligned}$$
(Porquê?)

$$\therefore \quad A = LDL^{\mathrm{T}}$$

Resta mostrar que os elementos da diagonal de D são positivos.

4. Relemebremos que os elementos da diagonal da matriz D são os da diagonal de U (matriz da dec. LU de A), i.e. são $u_{11}, \ldots, u_{nn}$.
De modo totalmente análogo ao que fizemos na demonstração da condição necessária de existência da decompoisção LU (ver slide nº 45), podemos concluir que, para $k = 1, \ldots, n$,

$$\det A_k = \det U_k = u_{11} u_{22} \ldots u_{kk}.$$

Assim, temos que

$$u_{11} = a_{11}$$

e

$$u_{kk} = \frac{\det A_k}{\det A_{k-1}}; k = 2, \ldots, n.$$

A propriedade **5** das matrizes d.p. garante que $u_{kk} > 0; k = 1, \ldots, n$.

Decomposição de Cholesky

Teorema

Seja A uma matriz real, simétrica, definida positiva. Então, A admite uma decomposição na forma

$$A = R^{\mathrm{T}} R$$

onde R é uma matriz triangular superior de elementos diagonais positivos.

Dem: Seja $A = LDL^{\mathrm{T}}$ a decomposição referida no teorema anterior. Como $D = \mathrm{diag}(u_{11}, \ldots, u_{nn})$ é diagonal com elementos positivos na diagonal, podemos formar a seguinte matriz diagonal, real de elementos diagonais positivos:

$$\tilde{D} := \mathrm{diag}(\sqrt{u_{11}}, \ldots, \sqrt{u_{nn}}),$$

a qual é tal que $\tilde{D}\tilde{D} = D$. Então, tem-se

$$A = LDL^{\mathrm{T}} = L\tilde{D}\tilde{D}L^{\mathrm{T}} = L\tilde{D}\tilde{D}^{\mathrm{T}}L^{\mathrm{T}} = (L\tilde{D})(L\tilde{D})^{\mathrm{T}}.$$

Sendo $R := (L\tilde{D})^{\mathrm{T}}$, tem-se que R é: triangular superior, de elementos positivos na diagonal e tal que $A = R^{\mathrm{T}}R$.

Decomposição de Cholesky

A decomposição $A = R^{\mathrm{T}}R$ com as carcaterísticas referidas no teorema anterior é única e é chamada decomposição de Cholesky da matriz A.
De modo análogo ao que fizemos para a dec. LU, é possível obter a decomposição de Cholesky (de uma matriz simétrica d.p. A) procurando diretamente os elementos $r_{ij}; j \geq i$ da matriz triangular superior $R = [r_{ij}]$. Seja

$$A = R^{\mathrm{T}}R = \begin{bmatrix} r_{11} & & & & \\ \vdots & \ddots & & & \\ r_{1i} & \cdots & r_{ii} & & \\ \vdots & & \vdots & \ddots & \\ r_{1n} & \cdots & r_{i,n} & \cdots & r_{nn} \end{bmatrix} \begin{bmatrix} r_{11} & \cdots & r_{1j} & \cdots & r_{1n} \\ & \ddots & \vdots & & \\ & & r_{jj} & \cdots & r_{jn} \\ & & & \ddots & \\ & & & & r_{nn} \end{bmatrix}$$

Então, tem-se $a_{ij} = r_{1i}r_{1j} + \ldots + r_{i-1,i}r_{i-1,j} + r_{ii}r_{ij}$; $j \geq i$, de onde se obtém:

$$r_{ii} = \sqrt{a_{ii} - r_{1i}^2 - \ldots - r_{i-1,i}^2} \qquad (*)$$

$$r_{ij} = \left(a_{ij} - r_{1i}r_{1j} - \ldots - r_{i-1,i}r_{i-1,j}\right)/r_{ii}; \quad j > i.$$

$(*)$ Sendo A definida positiva, tem-se sempre

$$a_{ii} - r_{1i}^2 - \ldots - r_{i-1,i}^2 > 0.$$

Algoritmo de Cholesky

função $R = \mathbf{decCholesky}(A)$

% *Decomposição de Cholesky de uma matriz A (simétrica, definida positiva)*

% ENTRADAS: matriz A, de ordem n (simétrica, d.p.)

% SAÍDA: matriz R, triangular superior, de elementos
% diagonais positivos e tal que $A = R^{\mathrm{T}} R$

Inicializar R como a matriz nula

para $i = 1{:}n$

$R(i,i) = \sqrt{A(i,i) - R(1{:}i-1,i)^{\mathrm{T}} * R(1{:}i-1,i)}$

$R(i,i+1{:}n) = \left(A(i,i+1{:}n) - R(1{:}i-1,i)^{\mathrm{T}} * R(1{:}i-1,i+1{:}n)\right) / R(i,i)$

fim

Notas

Pode mostrar-se que:

- Se A é uma matriz de diagonal estritamente dominante por linhas, isto é A, é tal que

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|; i = 1, \ldots, n$$

ou se A é de diagonal estritamente dominante por colunas, isto é,

$$|a_{jj}| > \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}|; j = 1, \ldots, n$$

então não há necessidade de escolha de pivô no método de eliminação de Gauss (ou na decomposição LU).

- O algoritmo de decomposição de Cholesky para matrizes definidas positivas é estável, não sendo necessária escolha de pivô.

SISTEMAS de EQUAÇÕES LINEARES
Métodos Iterativos

Normas vetoriais e matriciais
Norma vetorial

Definição

Seja $X = \mathbb{R}^n$. Uma aplicação $\|.\| : X \longrightarrow \mathbb{R}$ diz-se uma norma vetorial se satisfizer as seguintes propriedades:

NV1 $\forall x \in X, \quad \|x\| \geq 0$ e $\|x\| = 0 \iff x = \mathbf{0}$.

NV2 $\forall x \in X, \forall \alpha \in \mathbb{R}, \quad \|\alpha x\| = |\alpha| \|x\|$.

NV3 $\forall x, y \in X, \quad \|x + y\| \leq \|x\| + \|y\|$.

Exemplo

1. $\|x\|_1 = \sum_{i=1}^{n} |x_i|$ (norma 1)
2. $\|x\|_2 = \left\{\sum_{i=1}^{n} |x_i|^2\right\}^{1/2}$ (norma 2 ou norma Euclidiana)
3. $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$ (norma ∞ ou do máximo)

Nota: Se $\|.\|$ é uma norma vetorial, tem-se

$$| \, \|x\| - \|y\| \, | \leq \|x \pm y\|, \ \forall x, y \in X.$$

Norma matricial

Definição

Seja $X = \mathbb{R}^{n\times n}$. Uma aplicação $\|.\| : X \longrightarrow \mathbb{R}$ diz-se uma norma matricial se satisfizer as seguintes propriedades:

NM1 $\forall A \in X, \quad \|A\| \geq 0$ e $\|A\| = 0 \iff A = \mathbf{0}$ (matriz nula).

NM2 $\forall A \in X, \forall \alpha \in \mathbb{R}, \quad \|\alpha A\| = |\alpha|\|A\|$.

NM3 $\forall A, B \in X, \quad \|A + B\| \leq \|A\| + \|B\|$.

NM4 $\forall A, B \in X, \quad \|AB\| \leq \|A\|\|B\|$.

Nota: Também se definem normas para matrizes retangulares e, por vezes, não se exige a condição NM4 na definição de norma; neste curso, estamos apenas interessados em normas para matrizes quadradas que satisfaçam as quatro propriedades acima mencionadas. As definições de norma vetorial e de norma matricial estendem-se, de forma natural, para vetores em $\mathbb{C}^n$ e matrizes em $\mathbb{C}^{n\times n}$.

Norma matricial subordinada a uma norma vetorial

Seja $\|\cdot\|_V$ uma dada norma vetorial. Pode mostrar-se que a função $\|\cdot\|_M : \mathbb{R}^{n\times n} \to \mathbb{R}$ definida por

$$\|A\|_M := \max_{\substack{x\in\mathbb{R}^n \\ x\neq 0}} \frac{\|Ax\|_V}{\|x\|_V}$$

é uma norma matricial.

Definição

Seja $\|\cdot\|_V$ uma dada norma vetorial. A norma matricial $\|\cdot\|_M$ definida por

$$\|A\|_M = \max_{\substack{x\in\mathbb{R}^n \\ x\neq 0}} \frac{\|Ax\|_V}{\|x\|_V}$$

é chamada norma matricial subordinada à norma vetorial $\|\cdot\|_V$ ou norma matricial induzida por $\|\cdot\|_V$.

Normas matriciais subordinadas

Nota: A norma atrás referida pode ser definida, de forma equivalente, por

$$\|A\|_M = \max_{\substack{y\in\mathbb{R}^n \\ \|y\|_v=1}} \|Ay\|_V$$

Pode provar-se que as normas matriciais subordinadas às normas vetoriais $\|\cdot\|_1, \|\cdot\|_2$ e $\|\cdot\|_\infty$ (e que denotamos pelos mesmos símbolos, sendo claro pelo contexto se se trata da norma vetorial ou respetiva norma matricial) são dadas por

1. $\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|$ (máximo de soma por colunas)
2. $\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|$ (máximo de soma por linhas)
3. $\|A\|_2 = \sqrt{\rho(A^{\mathrm{T}}A)}$, onde $\rho(A^{\mathrm{T}}A)$ designa o raio espetral da matriz $B = A^{\mathrm{T}}A$, i.e. a quantidade definida por

$$\rho(B) = \max\{|\lambda| : \lambda \text{ é valor próprio de } B\}.$$

Exemplo

Seja $A = \begin{bmatrix} 1 & -5 & 2 \\ 8 & 1 & -1 \\ 3 & -1 & 1 \end{bmatrix}$. Então:

$$\|A\|_1 = \max\{1+8+3, 5+1+1, 2+1+1\} = \max\{12, 7, 4\} = 12,$$
$$\|A\|_\infty = \max\{1+5+2, 8+1+1, 3+1+1\} = \max\{8, 10, 5\} = 10$$

e (pode calcular-se, com mais algum trabalho) $\|A\|_2 = 8.6104$.

Se $\|.\|_V$ designa uma norma vetorial e $\|.\|_M$ a respetiva norma subordinada, então é imediato concluir que, para qualquer matriz $A \in \mathbb{R}^{n\times n}$ e qualquer vetor $x \in \mathbb{R}^n$ se tem

$$\boxed{\|Ax\|_V \leq \|A\|_M \|x\|_V}$$

Se uma norma matricial $\|.\|_M$ e uma norma vetorial $\|.\|_V$ satisfizerem a condição anterior, dizemos que a norma matricial é compatível com a norma vetorial.

Vemos, portanto, que qualquer norma subordinada é compatível com a respetiva norma vetorial.

Convergência de uma sequência de vetores

Considere-se uma sequência $(x^{(k)})_{k=1}^{\infty}$ de vetores no espaço vetorial $\mathbb{R}^n$.

Definição

Dizemos que a sequência $(x^{(k)})$ converge para o vetor $x \in \mathbb{R}^n$ se, para uma qualquer norma vetorial $\|.\|$, tivermos

$$\lim_{k\to\infty} \|x^{(k)} - x\| = 0.$$

Notas:

1. Se a condição anterior se verifica para uma determinada norma vetorial, então verifica-se também para qualquer outra norma vetorial.
2. A sequência $x^{(k)}$ converge para x se e só se

$$\lim_{k\to\infty} x_i^{(k)} = x_i, \ \forall i \in \{1, 2, \ldots, n\}.$$

Convergência de uma sequência de matrizes

Considere-se uma sequência $(A^{(k)})_{k=1}^{\infty}$ de matrizes no espaço vetorial $\mathbb{R}^{n\times n}$.

Definição

Dizemos que a sequência $(A^{(k)})$ converge para a matriz A se, para uma qualquer norma matricial $\|.\|$, tivermos

$$\lim_{k\to\infty} \|A^{(k)} - A\| = 0$$

Notas:

1. Se a condição anterior se verifica para uma determinada norma matricial, então verifica-se também para qualquer outra norma matricial.
2. A sequência $A^{(k)}$ converge para A se e só se

$$\lim_{k\to\infty} a_{ij}^{(k)} = a_{ij}, \ \forall i, j \in \{1, 2, \ldots, n\}.$$

Métodos iterativos

Como já referimos, a ideia dos métodos iterativos para resolver um sistema $Ax = b$, consiste em, a partir de uma aproximação inicial para a solução, gerar uma sequência de novas aproximações que, sob certas condições, converge para a solução do problema.

- Em teoria, a solução só será obtida no limite.
- Na prática, o processo será interrompido ao fim de um certo número de iterações, quando uma determinada medida do erro for suficientemente pequena.

Considere-se novamente o problema da resolução de um sistema de n equações lineares a n incógnitas $Ax = b$, onde A é não singular.
Vamos descrever três métodos iterativos básicos para a resolução deste sistema: método de Jacobi, método de Gauss-Seidel e método de relaxação sucessiva, geralmente designado por método SR.

Métodos iterativos

Por simplicidade, vamos começar por descrever os métodos para o caso $n = 4$, i.e., consideramos o sistema

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 = b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 = b_3 \\ a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 = b_4 \end{cases}$$

Se $a_{ii} \neq 0$, o sistema pode reescrever-se na forma

$$\begin{cases} x_1 = (b_1 - a_{12}x_2 - a_{13}x_3 - a_{14}x_4)/a_{11} \\ x_2 = (b_2 - a_{21}x_1 - a_{23}x_3 - a_{24}x_4)/a_{22} \\ x_3 = (b_3 - a_{31}x_1 - a_{32}x_2 - a_{34}x_4)/a_{33} \\ x_4 = (b_4 - a_{41}x_1 - a_{42}x_2 - a_{43}x_3)/a_{44} \end{cases}$$

Método de Jacobi

O método de Jacobi é definido por:

$$\begin{aligned} x_1^{(k+1)} &= (b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)} - a_{14}x_4^{(k)})/a_{11} \\ x_2^{(k+1)} &= (b_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k)} - a_{24}x_4^{(k)})/a_{22} \\ x_3^{(k+1)} &= (b_3 - a_{31}x_1^{(k)} - a_{32}x_2^{(k)} - a_{34}x_4^{(k)})/a_{33} \\ x_4^{(k+1)} &= (b_4 - a_{41}x_1^{(k)} - a_{42}x_2^{(k)} - a_{43}x_3^{(k)})/a_{44}; \end{aligned}$$

para $k = 0, 1, 2, \ldots$, com $x_1^{(0)}, x_2^{(0)}, x_3^{(0)}$ e $x_4^{(0)}$ dados.
No caso geral, o método de Jacobi é definido por:

$x_i^{(0)};\ i = 1, \ldots, n,$ dados.
Para $k = 0, 1, 2, \ldots$

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k)} \right); i = 1, \ldots, n.$$

Método de Gauss-Seidel

Neste método, os valores "atualizados"são utilizados, mal estejam disponíveis.
Assim, no caso $n = 4$, o método será definido por:

$$\begin{aligned} x_1^{(k+1)} &= (b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)} - a_{14}x_4^{(k)})/a_{11} \\ x_2^{(k+1)} &= (b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)} - a_{24}x_4^{(k)})/a_{22} \\ x_3^{(k+1)} &= (b_3 - a_{31}x_1^{(k+1)} - a_{32}x_2^{(k+1)} - a_{34}x_4^{(k)})/a_{33} \\ x_4^{(k+1)} &= (b_4 - a_{41}x_1^{(k+1)} - a_{42}x_2^{(k+1)} - a_{43}x_3^{(k+1)})/a_{44}; \end{aligned}$$

para $k = 0, 1, 2, \ldots$, com $x_1^{(0)}, x_2^{(0)}, x_3^{(0)}$ e $x_4^{(0)}$ dados.
No caso geral, o método de Gauss-Seidel é definido por:

$x_i^{(0)};\ i = 1, \ldots, n,$ dados.
Para $k = 0, 1, 2, \ldots$

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right); i = 1, \ldots, n.$$

Método SR(ω)

A equação que define o método de Gauss-Seidel pode escrever-se como

$$x_i^{(k+1)} = x_i^{(k)} + \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij} x_j^{(k)} \right)$$

Deste modo, a componente i da aproximação na iteração $k+1$, $x_i^{(k+1)}$, aparece como resultante da adição à correspondente componente da iteração anterior, $x_i^{(k)}$, de um certo termo de "correção".
No método de relaxação sucessiva, o termo de correção é multiplicado por um certo fator ω, dito parâmetro de relaxação. Isto é, o método SR com parâmetro de relaxação ω , SR(ω), é definido por:

$x_i^{(0)};\ i = 1, \ldots, n,$ dados.
Para $k = 0, 1, 2, \ldots$

$$x_i^{(k+1)} = x_i^{(k)} + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij} x_j^{(k)} \right); i = 1, \ldots, n.$$

Formulação matricial dos métodos

Decomponhamos a matriz $A = [a_{ij}]$ do sistema na soma $A = D + M + N$ onde

1. $D = \text{diag}(a_{11}, \ldots, a_{nn})$
2. M é a matriz triangular inferior

$$M = \begin{bmatrix} 0 & & & & & \\ a_{21} & 0 & & & & \\ a_{31} & a_{32} & 0 & & & \\ \vdots & \vdots & \vdots & & & \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{n,n-1} & 0 \end{bmatrix}$$

3. N é a matriz triangular superior

$$N = \begin{bmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1,n-1} & a_{1n} \\ & 0 & a_{23} & \cdots & a_{2,n-1} & a_{2n} \\ & & & & \vdots & \vdots \\ & & & & 0 & a_{n-1,n} \\ & & & & & 0 \end{bmatrix}$$

Formulação matricial do método de Jacobi

O sistema $Ax = b$ pode, assim, ser escrito como $(D + M + N)x = b$, ou, de forma equivalente, como

$$Dx = -(M + N)x + b,$$

ou ainda, como

$$x = -D^{-1}(M + N)x + D^{-1}b.$$

Do mesmo modo, é imediato concluir que o método de Jacobi é definido pela equação matricial

$$x^{(k+1)} = -D^{-1}(M + N)x^{(k)} + D^{-1}b,$$

ou seja como

$$x^{(k+1)} = B_J x^{(k)} + c_J$$

onde

$$B_J = -D^{-1}(M + N) \quad \text{e} \quad c_J = D^{-1}b$$

A matriz B_J é chamada matriz de iteração de Jacobi e o vetor c_J vetor de iteração de Jacobi.

Nota: Estamos a supor que $a_{ii} \neq 0$, o que garante que D^{-1} existe.

Formulação matricial do método de Gauss-Seidel

O método de Gauss-Seidel pode ser descrito pela equação matricial

$$x^{(k+1)} = D^{-1}(-Mx^{(k+1)} - Nx^{(k)} + b),$$

ou, de forma equivalente, por

$$Dx^{(k+1)} = -Mx^{(k+1)} - Nx^{(k)} + b,$$

ou ainda, por

$$(D + M)x^{(k+1)} = -Nx^{(k)} + b,$$

ou, finalmente, por

$$x^{(k+1)} = B_{GS} x^{(k)} + c_{GS}$$

onde

$$B_{GS} = -(D + M)^{-1}N \quad \text{e} \quad c_{GS} = (D + M)^{-1}b.$$

A matriz B_{GS} é dita matriz de iteração de Gauss-Seidel e o vetor c_{GS} é chamado vetor de iteração de Gauss-Seidel.

Formulação matricial do método SR(ω)

Finalmente, tem-se, para o método SR(ω):

$$x^{(k+1)} = x^{(k)} + \omega D^{-1}\left(-Mx^{(k+1)} - (D + N)x^{(k)} + b\right),$$

ou seja, tem-se

$$Dx^{(k+1)} = Dx^{(k)} + \omega\left(-Mx^{(k+1)} - (D + N)x^{(k)} + b\right),$$

ou ainda,

$$(D + \omega M)x^{(k+1)} = ((1 - \omega)D - \omega N)\, x^{(k)} + \omega b.$$

O método pode, assim, ser descrito pela seguinte equação matricial

$$x^{(k+1)} = B_{SR(\omega)} x^{(k)} + c_{SR(\omega)}$$

onde

$$B_{SR(\omega)} := (D + \omega M)^{-1}\left((1 - \omega)D - \omega N\right), \quad c_{SR(\omega)} := \quad \omega(D + \omega M)^{-1}b.$$

A matriz $B_{SR(\omega)}$ é chamada matriz de iteração do método SR(ω) e $c_{SR(\omega)}$ é o respetivo vetor de iteração.

Nota: Tem-se $B_{SR(1)} = B_{GS}$ e $cSR(1) = c_{GS}$.

Consistência dos métodos

Os três métodos acima referidos são, como vimos, da forma

$$x^{(k+1)} = Bx^{(k)} + c$$

para $k = 0, 1, 2, \ldots$, com $x^{(0)}$ dado, e onde B e c são os respetivos matriz e vetor de iteração. Para cada uma dos métodos considerados, facilmente se constata (verifique!) que o sistema dado $Ax = b$ pode ser reescrito como

$$x = Bx + c$$

Definição (Consistência)

Um método iterativo da forma $x^{(k+1)} = Bx^{(k)} + c$, diz-se consistente com um sistema $Ax = b$, se este sistema puder ser reescrito, de forma equivalente, como $x = Bx + c$.

Os métodos de Jacobi, Gauss-Seidel e SR(ω) são, por isso, métodos consistentes para a resolução de um sistema $Ax = b$.

Convergência dos métodos

Sejam dados um sistema $Ax = b$ (que supomos possível e determinado) e um método iterativo da forma $x^{(k+1)} = Bx^{(k)} + c$, consistente com o sistema.

Definição (Erro na iteração k)

À diferença entre a solução exata do sistema e o vetor obtido na iteração k do método chamamos erro na iteração k, o qual denotamos por $e^{(k)}$, i.e.,

$$e^{(k)} = x - x^{(k)}.$$

Definição (Convergência)

Dizemos que o método iterativo é convergente, se, seja qual for a aproximação inicial $x^{(0)}$, a sucessão dos erros $(e^{(k)})$ convergir para o vetor nulo $\mathbf{0}$, ou seja, se

$$\lim_{k\to\infty} \|e^{(k)}\| = 0$$

para qualquer norma vetorial (ou, de modo equivalente, se $e_i^{(k)} \to 0$, para todo o $i = 1, 2, \ldots, n$).

Convergência dos métodos iterativos

Teorema (Convergência dos métodos consistentes)

Seja dado um método iterativo $x^{(k+1)} = Bx^{(k)} + c$ consistente com o sistema $Ax = b$. Então, o método será convergente se e só se

$$\lim_{k\to\infty} B^k = \mathbf{0}.$$

Dem: Sendo o método consistente, ter-seá $x = Bx + c$. Vemos, então que

$$e^{(k)} = x - x^{(k)} = Bx + c - (Bx^{(k-1)} + c) = B(x - x^{(k-1)}) = Be^{(k-1)};\ k = 1, 2, \ldots,$$

ou seja, tem-se

$$e^{(k)} = Be^{(k-1)} = B^2e^{(k-2)} = \cdots = B^ke^{(0)},$$

onde $e^{(0)} = x - x^{(0)}$ é o erro na aproximação inicial. É imediato concluir que $e^{(k)} \to \mathbf{0}$ para qualquer aproximação inicial $x^{(0)}$, ou seja, para qualquer erro inicial $e^{(0)}$, se e só se a sequência de matrizes B^k convergir para a matriz nula.

Convergência de uma sequência de potências de matrizes

O teorema seguinte dá-nos uma condição necessária e suficiente para que a sequência das potências de uma dada matriz convirja para a matriz nula; a demonstração pode ser vista em e.g. D. M. Young, *Iterative Solutions of Large Linear Systems*, Ac. Press, 1971; veja também o Exercício 7 da folha de exercícios sobre normas matriciais.

Teorema

Dada uma matriz quadrada A, tem-se

$$\lim_{k\to\infty} A^k = \mathbf{0} \iff \rho(A) < 1,$$

onde $\rho(A)$ designa o raio espetral de A.

Como consequência imediata dos dois teorema anteriores, tem-se o seguinte resultado:

Corolário (Condição necessária e suficiente de convergência)

Seja dado um método iterativo $x^{(k+1)} = Bx^{(k)} + c$ consistente com um sistema $Ax = b$. Então, o método será convergente se e só se

$$\rho(B) < 1.$$

Em particular, podemos concluir que:

Convergência dos métodos de Jacobi, Gauss-Seidel e SR(ω)

É condição necessária e suficiente de convergência de qualquer dos três métodos atrás referidos – Jacobi, Gauss-Seidel e SR(ω) – que a respetiva **matriz de iteração** tenha raio espetral inferior a 1.

Teorema sobre convergência

O teorema seguinte estabelece uma condição **suficiente** de convergência dos métodos iterativos e fornece também majorantes para o erro na ieração k.

Teorema

Seja dado um método iterativo $x^{(k+1)} = Bx^{(k)} + c$ consistente com um sistema $Ax = b$. Se $\|B\| < 1$, então:

1. *O método converge;*
2. $\|e^{(k)}\| \leq \frac{\|B\|}{1-\|B\|}\|x^{(k)} - x^{(k-1)}\|$ *(Estimativa a posteriori para o erro)*
3. $\|e^{(k)}\| \leq \frac{\|B\|^k}{1-\|B\|}\|x^{(1)} - x^{(0)}\|$ *(Estimativa a priori para o erro)*

Nota: As normas usadas são quaisquer normas vetorial e matricial compatíveis.

Demosnstração do teorema

Como vimos anteriormente, sendo o método consistente, tem-se

$$e^{(k)} = B^k e^{(0)}.$$

Então, para quaisquer normas matricial e vetorial compatíveis, vem

$$\|e^{(k)}\| = \|B^k e^{(0)}\| \leq \|B^k\|\|e^{(0)}\| \leq \|B\|^k \|e^{(0)}\|.$$

É, então, imediato concluir que, se $\|B\| < 1$, se tem $e^{(k)} \to \mathbf{0}$, ou seja, o método converge, o que estabelece **1.**.
Para estabelecer **2.**, notemos que

$$\begin{aligned}\|e^{(k)}\| = \|x - x^{(k)}\| &= \|Bx + c - (Bx^{(k-1)} + c)\| = \|B(x - x^{(k-1)})\| \\ &\leq \|B\|\|x - x^{(k-1)}\| = \|B\|\|x - x^{(k)} + x^{(k)} - x^{(k-1)}\| \\ &\leq \|B\| \left(\underbrace{\|x - x^{(k)}\|}_{\|e^{(k)}\|} + \|x^{(k)} - x^{(k-1)}\| \right)\end{aligned}$$

Dem (cont.): Segue-se, portanto que

$$(1 - \|B\|)\|e^{(k)}\| \leq \|B\|\|x^{(k)} - x^{(k-1)}\|$$

ou seja (uma vez que $1 - \|B\| > 0$), que

$$\|e^{(k)}\| \leq \frac{\|B\|}{1 - \|B\|}\|x^{(k)} - x^{(k-1)}\|,$$

o que estabelece **2.**
Finalmente, o resultado **3.** ficará demonstrado desde que mostremos que

$$\|x^{(k)} - x^{(k-1)}\| \leq \|B\|^{k-1}\|x^{(1)} - x^{(0)}\|,$$

o que deixamos ao cuidado dos alunos.

Teorema

Se a **matriz** A *do sistema $Ax = b$ for de diagonal estritamente dominante (por linhas), então ambos os métodos de Jacobi e de Gauss-Seidel convergem.*

Dem: Fazemos a demonstração apenas para o caso do método de Jacobi. Para o caso do método de Gauss-Seidel, veja, e.g. K. Atkinson, *An Introduction to Numerical Analysis*, 2ª Ed., John Wiley, 1989, pp. 548-549.
De acordo com o teorema anterior, bastará mostrar que $\|B_J\| < 1$ para uma qualquer norma compatível. Relembremos que A é de diagonal estritamente dominante (por linhas) se e só se vrificar $|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|; i = 1, \ldots, n$. Como a matriz de iteraçao de Jacobi $B_J = -D^{-1}(M + N)$ tem diagonal nula e elmentos fora da diagonal da forma $-\frac{a_{ij}}{a_{ii}}$ é imediato reconhecer que

$$\|B_J\|_\infty = \max_i \left(\sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} \right) = \max_i \left(\frac{\sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|}{|a_{ii}|} \right) < 1.$$

Nota: É também possível mostrar que os métodos convergem se A for de diagonal estritamente dominante por colunas.

Condição necessária de convergência do método SR(ω)

Teorema (de Kahan)

É condição necessária de convergência do método SR(ω) que o parâmetro de relaxação ω satisfaça $0 < \omega < 2$.

Dem: Sejam $\lambda_i; i = 1, \ldots, n$, os valores próprios da matriz $B = B_{SR(\omega)}$ de iteração do método SR(ω). Como sabemos, o determinante de uma matriz é igual ao produto dos seus valores próprios, pelo que teremos

$$\prod_{i=1}^{n} \lambda_i = \det B = \det\left[(D+\omega M)^{-1}\left((1-\omega)D - \omega N\right)\right]$$

$$= \frac{1}{\prod_{i=1}^{n} a_{ii}} \prod_{i=1}^{n}(1-\omega)a_{ii} = \prod_{i=1}^{n}(1-\omega) = (1-\omega)^n.$$

Assim, tem-se $|\prod_{i=1}^{n} \lambda_i| = \prod_{i=1}^{n} |\lambda_i| = |1-\omega|^n$, de onde se conclui de imediato que

$$(\rho(B))^n = (\max_i |\lambda_i|)^n \geq \prod_{i=1}^{n} |\lambda_i| = |1-\omega|^n.$$

Condição necessária de convergência do método SR(ω)

Dem (cont.):
Vimos, assim, que

$$\rho(B) \geq |1-\omega|.$$

Então, tem-se

$$\rho(B) < 1 \Longrightarrow |1-\omega| < 1 \iff -1 < 1-\omega < 1 \iff 0 < \omega < 2.$$

Assim, a condição referida é necessária para que se verifique a condição $\rho(B) < 1$, a qual, como sabemos, é necessária para a convergência do método. Pode também provar-se (Teorema de Ostrowski-Reich) que, se A for definida positiva, a condição $0 < \omega < 2$ será também suficiente para a convergência do método SR(ω). Em particular, tem-se que o método de Gauss-Seidel ($\equiv SR(1)$) converge quando a matriz do sistema é definida positiva.

Critérios de paragem

Ao implementar qualquer um dos método iterativos referidos, há necessidade de decidir quando interromper o processo ierativo. Tipicamente, tal será feito quando se verificar uma (ou por vezes mais do que uma em simultâneo) das condições seguintes:

1. $\|x^{(k)} - x^{(k-1)}\| < \epsilon_1$
2. $\|x^{(k)} - x^{(k-1)}\| < \epsilon_2 \|x^{(k)}\|$
3. $\|r^{(k)}\| < \epsilon_3$, onde $r^{(k)}$ denota o chamado vetor residual definido por $r^{(k)} := Ax^{(k)} - b$
4. $\|r^{(k)}\| < \epsilon_4 \|b\|$,
5. $\|r^{(k)}\| < \epsilon_5 \left(\|A\|\|x^{(k)}\| + \|b\|\right)$,

onde a norma utilizada e os valores das tolerâncias ϵ_i são escolhidos pelo utilizador.

Critérios de paragem

Notas:

1. O facto de termos $\|x^{(k)} - x^{(k-1)}\| < \epsilon_1$ não garante que $\|e^{(k)}\| = \|x - x^{(k)}\| < \epsilon_1$. No entanto, se $\|B\| < 1$, podemos obter facilmente um majorante para $\|e^{(k)}\|$ usando o resultado

$$\|e^{(k)}\| \leq \frac{\|B\|}{1-\|B\|}\|x^{(k)} - x^{(k-1)}\|.$$

   (Note-se que, se $\|B\| < \frac{1}{2}$, então $\frac{\|B\|}{1-\|B\|} < 1$ e o valor $\|x^{(k)} - x^{(k-1)}\|$ dará uma estimativa "pessimista" para o erro $\|e^{(k)}\|$.)

2. Uma discussão mais aprofundada sobre critérios de paragem, em particular uma justificação de alguns dos critérios referidos, pode ser vista em R. Barrett *et al.*, Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, SIAM, Philadelphia (1994).
3. Tendo em atenção que o processo pode divergir ou convergir de forma muito lenta, deverá **sempre** incluir-se um critério de paragem do tipo:

Parar quando $k > kmax$, onde $kmax$ é o número máximo de iterações que o utilizador pretende que sejam efetuadas.

Algumas notas sobre os métodos estudados

1. A rapidez de convergência dos métodos é determinada pelo tamanho do raio espetral da respetiva matriz de iteração B, sendo tanto mais rápida quanto menor for esse valor, tendo-se
$$\|e^{(k)}\| \approx \rho(B)\|e^{(k-1)}\|.$$
2. Quando o método de Jacobi converge, então quase sempre, o método de Gausss-Seidel também converge e sua convergência é mais rápida; por exemplo, para matrizes de diagonal estritamente dominante o método de Gauss-Seidel nunca converge mais lentamente do que o método de Jacobi.
3. O método de Gauss-Seidel tem menores necessidades de armazenamento do que o método de Jacobi, uma vez que a componente $x_i^{(k+1)}$ pode ser sempre sobreposta à componente da iteração anterior $x_i^{(k)}$ que não volta a ser utilizada; por estas razões, o método de Gauss-Seidel é utilizado com maior frequência do que o método de Jacobi.

Algumas notas sobre os métodos estudados

4. No entanto, o método de Jacobi é mais apropriado para ser utilizado numa máquina com arquitetura paralela (o cálculo de $x_i^{(k+1)}$ não ncessita de esperar pelo cálculo de $x_{i-1}^{(k+1)}, x_{i-2}^{(k+1)}, \ldots$) do que o método de Gauss-Seidel.
5. A determinação do parâmetro de relaxação ótimo (isto é, do valor de ω para o qual a convergência do método SR(ω) é o mais rápida possível) é um problema de difícil solução, a não ser para certas classes de matrizes.
6. Existem muitos outros métodos iterativos, mais eficientes do que os métodos que estudámos, mas o seu estudo está fora do âmbito deste curso.

Condicionamento de sistemas lineares

Perturbação no lado direito do sistema

Como sabemos, um problema diz-se mal condicionado se os valores calculados forem muito sensíveis a pequenas alterações nos dados e, consequentemente, muito sensíveis aos inevitáveis erros de arredondamento introduzidos durante o processo de cálculo. Vamos agora debruçar-nos um pouco sobre o problema do condicionamento de um sistema de equações lineares.

Consideremos novamente um sistema

$$Ax = b$$

e suponhamos que $b \neq 0$ e que A é não singular.

Comecemos por estudar o efeito que uma perturbação no lado direito do sistema, definida por um vetor δb, pode produzir na solução.
Seja $x + \delta x$ a solução do sistema perturbado, isto é, suponhamos que

$$A(x + \delta x) = b + \delta b.$$

Nota: No que se segue, admitimos que a norma matricial e vetorial usadas são compatíveis.

Condicionamento de sistemas lineares

Temos

$$\begin{aligned} A(x+\delta x) = b + \delta b &\iff Ax + A\delta x = b + \delta b \iff A\delta x = \delta b \\ &\qquad\qquad\qquad\qquad (Ax = b) \\ &\iff \delta x = A^{-1}\delta b \implies \|\delta x\| = \|A^{-1}\delta b\| \\ &\implies \|\delta x\| \leq \|A^{-1}\|\ \|\delta b\|. \end{aligned} \tag{1}$$

Como $\|b\| = \|Ax\| \leq \|A\|\,\|x\|$, segue-se que

$$\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|} \tag{2}$$

(Note-se que, como $b \neq 0$, também $x \neq 0$).

Condicionamento de sistemas lineares

De (1) e (2), obtém-se, então,

$$\boxed{\frac{\|\delta x\|}{\|x\|} \leq \|A\|\,\|A^{-1}\|\frac{\|\delta b\|}{\|b\|}}$$

A expressão anterior estabelece um majorante para o erro relativo na solução do sistema, $\|\delta x\|/\|x\|$, produzido por uma perturbação em b com erro relativo $\|\delta b\|/\|b\|$. O fator $\|A\|\,\|A^{-1}\|$ mede, assim, a sensibilidade da solução a uma perturbação no lado direito do sistema.

Definição

Seja A uma matriz quadrada invertível. Chamamos número de condição de A e denotamos por $\mathrm{cond}(A)$ o número dado por

$$\mathrm{cond}(A) := \|A\|\,\|A^{-1}\|.$$

Perturbação de A e b

O seguinte teorema (cuja demonstração pode ser vista em, e.g. K.E. Atkinson, *An Introduction to Numerical Analysis*, 2ª ed., John Wiley, 1989, p. 535) estabelece um resultado sobre a sensibilidade de um sistema $Ax = b$ a alterações na matriz A e no lado direito b do sistema.

Teorema

Sejam A e δA matrizes em $\mathbb{R}^{n\times n}$ e b , δb, x e δx vetores de $\mathbb{R}^n$ $(b \neq 0)$ tais que $Ax = b$ e $(A + \delta A)(x + \delta x) = b + \delta b$. Suponhamos, além disso, que A é invertível e que, para uma certa norma subordinada se tem

$$\|A^{-1}\|\,\|\delta A\| < 1.$$

Então, é válido o seguinte resultado

$$\boxed{\frac{\|\delta x\|}{\|x\|} \leq \frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right)}$$

Temos

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right). \qquad (3)$$

Se tivermos

$$\|A^{-1}\|\,\|\delta A\| \ll 1,$$

então

$$1 - \mathrm{cond}(A)\frac{\|\delta A\|}{\|A\|} = 1 - \|A^{-1}\|\,\|\delta A\| \approx 1$$

e, portanto, o majorante para o erro relativo $\frac{\|\delta x\|}{\|x\|}$ dado por (3) poderá ser aproximado por

$$\mathrm{cond}(A)\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right).$$

Costumamos escrever, com esse significado:

$$\boxed{\frac{\|\delta x\|}{\|x\|} \lesssim \mathrm{cond}(A)\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right).}$$

Notas sobre o número de condição de uma matriz

1. O número de condição depende, naturalmente da norma utilizada.
2. Sendo $\|.\|$ uma norma compatível com uma norma vetorial, então, como sabemos, $\|I\| \geq 1$. Então, temos
$$\mathrm{cond}(A) = \|A\|\,\|A^{-1}\| \geq \|AA^{-1}\| = \|I\| \geq 1,$$
ou seja, o número de condição é sempre maior ou igual a 1.
3. Um número de condição $\mathrm{cond}(A)$ pequeno é garantia de que o sistema $Ax = b$ é bem condicionado e um número de condição grande é indicador de que o sistema poderá ser (e geralmente, é) mal condicionado.
4. A determinação do número de condição de uma matriz A requer o cálculo de A^{-1} e envolve, portanto, grande esforço computacional. Além disso, se A é mal condicionada, o cálculo de A^{-1} será um problema de difícil solução e portanto, o cálculo do número de condição também. No entanto, é possível encontrar boas estimativas para o número de condição (relativamente, por exemplo, a $\|\cdot\|_1$) com relativo pouco esforço computacional.

Vetor residual e vetor solução

Seja x a solução extata de um sistema $Ax = b$, com A não singular, e suponhamos que o sistema foi resolvido numericamente e que $\tilde{x}$ é a solução aproximada efetivamente calculada. Seja r o vetor residual, isto é,

$$r = b - A\tilde{x}.$$

Poderíamos ser levados a pensar que um vetor residual de componentes "pequenas" (i.e. com norma pequena) seria garantia de que a aproximação numérica $\tilde{x}$ fosse uma boa aproximação para x. No entanto, temos

$$r = b - A\tilde{x} = Ax - A\tilde{x} = A(x - \tilde{x}) \Rightarrow x - \tilde{x} = A^{-1}r.$$

Assim, podemos apenas garantir que

$$\|x - \tilde{x}\| \leq \|A^{-1}\|\|r\|.$$

Isto mostra que o facto de termos um resíduo pequeno não garante que a solução aproximada seja muito próxima da solução exata. Em particular, se $\mathrm{cond}(A)$ for muito grande (não sendo $\|A\|$ muito grande), o resíduo pode ser pequeno e a solução diferir muito da solução exata.

Exemplo

Considere-se o sistema

$$\begin{cases} 2x_1 + 6x_2 = 8 \\ 2x_1 + 6.00001x_2 = 8.00001 \end{cases},$$

cuja solução exata é $x_1 = x_2 = 1$.
Suponhamos que, numa máquina, o sistema efetivamente resolvido foi

$$\begin{cases} 2x_1 + 6x_2 = 8 \\ 2x_1 + 5.99999x_2 = 8.00002 \end{cases},$$

o qual forneceu a "solução aproximada": $\tilde{x}_1 = 10$ e $\tilde{x}_2 = -2$. Temos

$$\begin{aligned} r = b - A\tilde{x} &= \begin{bmatrix} 8 \\ 8.00001 \end{bmatrix} - \begin{bmatrix} 2 & 6 \\ 2 & 6.00001 \end{bmatrix} \begin{bmatrix} 10 \\ -2 \end{bmatrix} \\ &= \begin{bmatrix} 8 \\ 8.00001 \end{bmatrix} - \begin{bmatrix} 8 \\ 7.99998 \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \times 10^{-5} \end{bmatrix} \end{aligned}$$

No entanto,

$$x - \tilde{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 10 \\ -2 \end{bmatrix} = \begin{bmatrix} -9 \\ 3 \end{bmatrix}.$$

Exemplo (cont.)

O que acontece, neste caso, é que

$$A^{-1} = \frac{1}{2} \times 10^5 \begin{bmatrix} 6.00001 & -6 \\ -2 & 2 \end{bmatrix}$$

tem elementos cuja ordem de grandeza é de 10^5.
Note-se que (relativamente à norma $\| \ \|_\infty$) temos

$$\mathrm{cond}(A) = \|A\|_\infty \|A^{-1}\|_\infty \approx 4.8 \times 10^6.$$

Assim, se pedíssemos uma estimativa para o número de condição da matriz A, ser-nos-ia indicado um número muito grande e, de imediato, deveríamos "desconfiar" da nossa solução aproximada.

EQUAÇÕES NÃO LINEARES

Introdução

O problema sobre o qual nos debruçaremos agora é o da determinação de raízes de equações não lineares (numa variável).Dada uma função real de variável real f, não linear, procuramos $r \in \mathbb{R}$ para o qual se tenha

$$f(r) = 0.$$

Tal valor r é dito uma raiz da equação $f(x) = 0$ ou um zero da função f.

Exemplo

1. $x - \exp(-x) = 0$ (uma única raiz: $r \approx 0.5671$)
2. $x^3 - 4x^2 + x - 6 = 0$ (três raízes: $r_1 = -1, r_2 = 2, r_3 = 3$)
3. $\operatorname{sen} x - 1 = 0$ (uma infinidade de raízes: $r_k = \frac{\pi}{2} + 2k\pi,\ k \in \mathbb{Z}$)
4. $\exp(x) + 1 = 0$ (nenhuma raiz)

Métodos iterativos

Os métodos que iremos estudar são, todos eles, métodos iterativos: são dadas $m+1$ aproximações iniciais $x_0, \ldots, x_m$ para uma raiz r da equação $f(x) = 0$ e determina-se, então, uma nova aproximação x_{m+1}, repetindo-se sucessivamente este processo. Mais precisamente, é gerada uma sequência (x_k) de aproximações para r através do uso de fórmulas do tipo

$$x_{k+1} = g(k, x_k, x_{k-1}, \ldots, x_{k-m}); k = m, m+1, \ldots$$

Se a função g não depender de k, isto é, se a forma da função iterativa se mantiver de iteração para iteração, o método diz-se estacionário.

Definição

Seja $e_k := r - x_k$ o erro na aproximação x_k para r. O método diz-se convergente se

$$\lim_{k\to\infty} x_k = r$$

ou, equivalentemente,

$$\lim_{k\to\infty} e_k = 0.$$

Ordem de convergência

Definição (ordem de convergência)

Seja (x_k) uma sequência de aproximações obtidas por um determinado método iterativo e suponhamos que $\lim_{k\to\infty} x_k = r$. Seja $e_k := r - x_k$. Se existir um número $p \geq 1$ e uma constante $C > 0$ tais que

$$\lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^p} = \lim_{k\to\infty} \frac{|x_{k+1} - r|}{|x_k - r|^p} = C$$

dizemos que o método converge com ordem de convergência p (em r). Se $p = 1$, a convergência diz-se linear, dizendo-se quadrática se $p = 2$, cúbica se $p = 3$, etc; se $1 < p < 2$, a convergência é dita, superlinear. A constante C é chamada constante de convergência assintótica.

É usual escrever a expressão $\lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^p} = C$ de forma assintótica, como

$$|e_{k+1}| \sim C|e_k|^p,$$

a qual nos indica como o erro se comporta quando k é suficientemente grande.

Notas

1. Dados a sucessão (x_k) e $r \in \mathbb{R}$, se $e_k := r - x_k$ satisfaz

$$\lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^p} = C$$

então:

- Se $p = 1$e $C > 1$, não poderá haver convergência de x_k para r.
- Se $p > 1$ ou se $p = 1$ e $C < 1$, há garantia de convergência de x_k para r (desde que x_0 seja escolhido suficientemente próximo de r).

2. Por vezes, não é possível mostrar que $\lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|} = C < 1$, sendo, no entanto possível possível estabelecer uma relação do tipo

$$|e_k| \leq KC^k,\ C < 1.$$

Neste caso, dizemos ainda que o método converge[5] **linearmente** (com razão de convergência C).

[5] Tem-se, naturalmente, $\lim_{k\to\infty} e_k = 0$, ou seja, o método converge, efetivamente.

Notas

3. Quanto maior for a ordem de convergência de um método iterativo e menor a respetiva constante de convergência, maior será a sua rapidez de convergência, i.e., menor será o número de iterações necessárias para se obter uma aproximação razoável para a raiz ; isto não significa que o método seja necessariamente mais eficiente, pois a eficiência dependerá, também, do esforço computacional exigido em cada iteração.

Método da Bisseção

O primeiro método que descrevemos para determinar uma raiz de uma equação não-linear é um processo muito simples e baseia-se exclusivamente no uso sucessivo do Teorema do Valor Intermédio, o qual recordamos:

Teorema (do Valor Intermédio)

Seja $f \in C[a,b]$. Se K é tal que $f(a) < K < f(b)$ ou $f(b) < K < f(a)$, então, existe $\xi \in (a,b)$ tal que $f(\xi) = K$; em particular, se $f(a)f(b) < 0$, então existe pelo menos uma raiz da equação $f(x) = 0$ no interior do intervalo $[a,b]$.

Consideremos, então, uma equação

$$f(x) = 0$$

e suponhamos que encontramos valores $a, b \in \mathbb{R}$, com $a < b$, e tais que $f(a)f(b) < 0$ (por outras palavras, suponhamos que em a e b a função f tem sinais contrários) e que, além disso, f é contínua no intervalo $I = [a,b]$. Então, por aplicação do Teorema do Valor Intermédio, podemos concluir que a equação dada tem (pelo menos) uma raiz no interior do intervalo $I = [a,b]$.

Seja m o ponto médio do intervalo $[a,b]$, isto é, seja

$$m = \frac{a+b}{2}$$

e calculemos o valor de f nesse ponto. Então:

- Se $f(m) = 0$, m é uma raiz da equação $f(x) = 0$;
- Se $f(m) \neq 0$, o sinal de $f(m)$ indica-nos em qual dos dos dois subintervalos (a,m) ou (m,b) se encontra uma raiz. Mais precisamente:
  - Se $f(a)f(m) < 0$, há uma raiz no intervalo (a,m);
  - Se $f(a)f(m) > 0$, há uma raiz no intervalo (m,b).

Método da Bisseção

A amplitude do novo intervalo contendo uma raiz é, assim, metade da do intervalo inicial, podendo tomar-se como aproximação para a raiz o ponto médio desse subintervalo; o processo, pode, então, repetir-se até se encontrar um intervalo, de amplitude suficientemente pequena, onde se localize uma raiz.
Este processo constitui o chamado **método da bisseção** para a determinação de uma raiz de uma equação não linear $f(x) = 0$.
Note-se que o processo anterior, ou termina num número finito de passos, ou determina uma sequência encaixada de intervalos

$$[a,b] = I_0 \supset I_1 \supset I_2 \supset \ldots,$$

existindo, além disso, $r \in I_k$; $k = 0, 1, 2 \ldots$, tal que $f(r) = 0$.
Sendo x_k o ponto médio do intervalo I_{k-1} tem-se

$$|e_k| = |r - x_k| \leq \frac{b-a}{2^k}.$$

Método da Bisseção
Convergência

- A fórmula
$$|e_k| = |r - x_k| \leq \frac{b-a}{2^k} = (b-a)\left(\frac{1}{2}\right)^k$$
mostra-nos (ver Nota 2 da p.5) que o método converge linearmente, com razão de convergência $\frac{1}{2}$. Isto significa que este método converge lentamente, o que constitui uma das suas principais desvantagens.
- A fórmula anterior permite-nos também determinar, "a priori", qual o número de iterações necessárias para garantir que x_k aproxime r com uma determinada precisão. Mais precisamente, supondo que pretendemos ter $|r - x_k| \leq \delta$, deverá ser
$$(b-a)\left(\frac{1}{2}\right)^k \leq \delta \iff 2^k \geq \frac{b-a}{\delta},$$
ou seja, k deverá ser escolhido como o menor um inteiro satisfazendo
$$k \geq \log_2\left(\frac{b-a}{\delta}\right).$$

Método da Bisseção
Convergência global

A principal vantagem do método da bisseção reside no facto de ele ter convergência garantida, seja qual for a amplitude do intervalo inicial $[a, b]$ (supondo, naturalmente,que estão satisfeitas as condições $f \in C[a, b]$ e $f(a)f(b) < 0$).
Assim sendo, não é necessário começar a iterar "suficientemente" próximo da raiz para haver garantia de convergência, o que acontece (como veremos) com outros métodos.
Por essa razão, dizemos que o método da bisseção é um método de convergência global.

Algoritmo do método da bisseção

função $r = \mathbf{metBissec}(f, a, b, tol)$
% Determina uma raiz de uma equação não linear pelo método da bisseção
% ENTRADAS: f função de uma variável
% a, b reais, $a < b$ (t.q. $f(a)f(b) < 0$)
% tol real positivo (tolerância para o erro)
% SAÍDA: r aproximação para uma raiz de $f(x) = 0$
$fa = f(a)$; $fb = f(b)$
% Determinar número de iterações nec. para que erro$\leq tol$.
$n = \text{ceil}((\log(b-a) - \log(tol))/\log(2))$
para $k = 1 : n$
    $r = a + (b-a)/2$; $fr = f(r)$
    se $fa * fr < 0$
        $b = r$; $fb = fr$
    de outro modo se $fa * fr > 0$
        $a = r$; $fa = fr$
    de outro modo
        sair do ciclo para (c/mensagem de que a raiz foi encontrada)
    fim
fim

Notas

1. Embora, teoricamente, o método da bisseção convirja, sendo, portanto possível, iterando um número suficiente de vezes, obter uma aproximação para uma raiz com a precisão desejada, **do ponto de vista computacional** haverá limitações, dependentes da precisão da máquina utilizada.
   Por exemplo, devido a erros de arredondamento, poderá acontecer que, para um determinado valor de k, ao calcular-se $f(x_k)$, o seu sinal não seja determinado corretamente. Nesse caso, ter-seá sempre $|x_p - r| \geq |x_k - r|$, para qualquer $p > k$.
   (Será pouco provável, no entanto, que esta situação ocorra, a não ser que $f(x_k) \approx 0$ e o gráfico de f seja, na vizinhança de r, "quase" paralelo ao eixo dos xx).
2. Na prática, geralmente, o algoritmo da bissecção é utilizado apenas para fornecer uma aproximação "razoável" para uma raiz, a qual servirá de aproximação inicial para outros métodos mais eficientes. Assim, apenas um pequeno número de iterações é, geralmente, efetuado.

Exemplo

Seja $f(x) = x^3 - 2x - 5$.

- $f(2) = -1 < 0$, $f(3) = 16 > 0$
- $f \in C[2,3]$

k	a_k	b_k	Sinal de $f(a_k)$	x_{k+1}	Sinal de $f(x_{k+1})$	Maj. erro
0	2.000000	3.000000	–	2.500000	+	5.00e-001
1	2.000000	2.500000	–	2.250000	+	2.50e-001
2	2.000000	2.250000	–	2.125000	+	1.25e-001
3	2.000000	2.125000	–	2.062500	–	6.25e-002
4	2.062500	2.125000	–	2.093750	–	3.13e-002
5	2.093750	2.125000	–	2.109375	+	1.56e-002
6	2.093750	2.109375	–	2.101563	+	7.81e-003
7	2.093750	2.101563	–	2.097656	+	3.91e-003
8	2.093750	2.097656	–	2.095703	+	1.95e-003
9	2.093750	2.095703	–	2.094727	+	9.77e-004
10	2.093750	2.094727	–	2.094238	n.c.	4.88e-004

Funções contrativas e pontos fixos

Vamos, de seguida, introduzir o chamado Teorema do Ponto Fixo de Banach, o qual que nos servirá de base ao estudo de dois métodos seguintes. Começamos com a seguintes definições.

Definição (Função contrativa)

Seja $\emptyset \neq D \subseteq \mathbb{R}$ e seja $g : D \to \mathbb{R}$ uma certa função. Dizemos que g é contrativa em D, se existe uma constante $0 \leq L < 1$, tal que

$$|g(x) - g(y)| \leq L|x - y|, \ \forall x, y \in D.$$

A constante L é chamada constante de Lipschitz.

Nota: Sendo g contrativa em D, então g é contínua em D (prove!).

Definição (ponto fixo)

Seja $\emptyset \neq D \subseteq \mathbb{R}$ e seja $g : D \to \mathbb{R}$ uma certa função. Diz-se que $\alpha \in D$ é um ponto fixo de g se

$$g(\alpha) = \alpha.$$

Teorema do ponto fixo

Teorema (do ponto fixo de Banach num intervalo fechado de $\mathbb{R}$)

Seja I um intervalo fechado de $\mathbb{R}$ e seja $g : I \to I$ uma aplicação de I em si mesmo, contrativa, com constante de Lipschitz L. Então:

1. *Para qualquer valor inicial $x_0 \in I$, a sequência de iterações (x_k) definida por*

$$x_{k+1} = g(x_k); \ k = 0, 1, 2, \ldots$$

é convergente para um ponto $\alpha \in I$.

2. *O limite α da sequência (x_k) é um ponto fixo de g, sendo, além disso, o único ponto fixo da função g em I.*

3. *O erro $e_k := \alpha - x_k$ satisfaz:*
   - *Estimativa a posteriori:* $|\alpha - x_k| \leq \frac{L}{1-L}|x_k - x_{k-1}|$
   - *Estimativa a priori:* $|\alpha - x_k| \leq \frac{L^k}{1-L}|x_1 - x_0|$

Demonstração

1. Temos

$$\left.\begin{array}{l} g(I) \subseteq I \\ x_0 \in I \\ x_k = g(x_{k-1}); k = 0, 1, 2 \ldots \end{array}\right\} \Rightarrow x_k \in I, \text{para todo o } k$$

Além disso, temos

$$|x_{k+1} - x_k| = |g(x_k) - g(x_{k-1})| \leq L|x_k - x_{k-1}|$$

de onde se deduz facilmente que

$$|x_{k+1} - x_k| \leq L^k|x_1 - x_0|, \ k = 0, 1, 2, \ldots \qquad (4)$$

Dem. (cont.)

Assim, para $p > k$, tem-se

$$\begin{aligned}|x_p - x_k| &\underset{\text{(des. triang.)}}{\leq} |x_p - x_{p-1}| + |x_{p-1} - x_{p-2}| + \cdots + |x_{k+1} - x_k| \\ &\underset{\text{(usando (4))}}{\leq} \left(L^{p-1} + L^{p-2} + \cdots + L^k\right)|x_1 - x_0| \\ &= L^k\left(1 + L + \cdots + L^{p-1-k}\right)|x_1 - x_0| \\ &= \frac{L^k(1 - L^{p-k})}{1-L}|x_1 - x_0| \underset{(L<1)}{\leq} \frac{L^k}{1-L}|x_1 - x_0|.\end{aligned}$$

- $L^k \to 0$ (quando $k \to \infty$) $\Rightarrow (x_k)$ de Cauchy $\Rightarrow (x_k)$ convergente.

Dem (cont.)

Seja $\alpha := \lim_{k\to\infty}(x_k)$.
Como $x_k \in I$ para todo o k e I é fechado, podemos concluir que $\alpha \in I$.
2. Vejamos que α é ponto fixo de g. Tem-se

$$g(\alpha) = g(\lim_k x_k) \underset{①}{=} \lim_k (g(x_k)) = \lim_k x_{k+1} \underset{②}{=} \alpha$$

① Porque g é contínua
② Porque (x_{k+1}) é uma subsucessão de (x_k).
Mostremos, agora, que só há um ponto fixo de g em I. Suponhamos que α, β são ambos pontos fixos de g em I. Então, tem-se

$$\begin{aligned}|\alpha - \beta| = |g(\alpha) - g(\beta)| &\leq L|\alpha - \beta| \\ \Rightarrow (1-L)|\alpha - \beta| \leq 0 &\underset{(L<1)}{\Rightarrow} |\alpha - \beta| = 0 \Rightarrow \alpha = \beta.\end{aligned}$$

Dem (cont.)

Resta-nos, apenas, estabelecer as estimativas para o erro na iteração k.
Tem-se

$$\begin{aligned}|\alpha - x_k| = |g(\alpha) - g(x_{k-1})| &\leq L|\alpha - x_{k-1}| \\ &\leq L\left(|\alpha - x_k| + |x_k - x_{k-1}|\right)\end{aligned}$$

Segue-se, assim, que

$$(1-L)|\alpha - x_k| \leq L|x_k - x_{k-1}|$$

ou seja, uma vez que $(1-L) > 0$, que

$$|\alpha - x_k| \leq \frac{L}{1-L}|x_k - x_{k-1}|.$$

A demosntração da estimativa *a priori* é deixada ao cuidado dos alunos.

Notas

1. A condição
$$|g(x) - g(y)| < |x - y|, \ \forall x, y \in I, \ x \neq y$$
(que é uma condição mais fraca do que a contratividade) não é suficiente para garantir a existência de um ponto fixo em I. Por exemplo, a função $g : [0, \infty) \to [0, \infty)$ dada por
$$g(x) = x + \frac{1}{1+x}$$
satisfaz a condição acima (verifique!), mas não tem nenhum ponto fixo em $I = [0, \infty)$, uma vez que
$$\frac{1}{1+x} > 0, \ \forall x \geq 0.$$

2. No entanto, se, para além de fechado, o intervalo I for também limitado (i.e. se $I = [a, b]$ com $a, b \in \mathbb{R}$) então a condição anterior é suficiente para a exisência de um ponto fixo de g em I (o qual pode ser procurado gerando a sequência $x_{k+1} = g(x_k), x_0 \in I$).

Método das iterações sucessivas (ou do ponto fixo)

Retomemos o problema da determinação de raízes de uma equação não linear

$$f(x) = 0. \tag{5}$$

Suponhamos que, a partir desta equação, obtemos uma equação equivalente, da forma

$$g(x) = x, \tag{6}$$

transformando, assim, o problema da determinação de raízes da equação (5) no da determinação de pontos fixos da função g.

Exemplo

$$x^2 - x - 2 = 0 \Leftrightarrow \underbrace{x^2 - 2}_{g_1(x)} = x$$

$$x^2 - x - 2 = 0 \Leftrightarrow x^2 = 2 + x \Leftrightarrow x = \underbrace{\sqrt{x+2}}_{g_2(x)}, \quad (\text{para } x \geq -2).$$

$$x^2 - x - 2 = \Leftrightarrow x^2 = x + 2 \Leftrightarrow x = \underbrace{1 + \frac{2}{x}}_{g_3(x)}, \quad (\text{para } x \neq 0).$$

Método das iterações sucessivas (ou do ponto fixo)

Se encontrarmos um intervalo fechado I tal que

- $g(I) \subseteq I$
- g seja contractiva em I

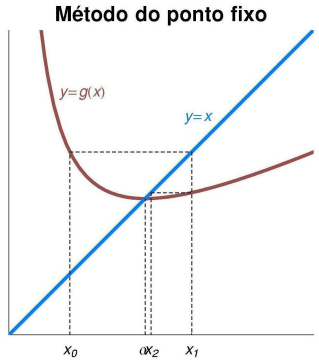
então, tendo em conta o Teorema do Ponto Fixo de Banach, poder-seá procurar α (único ponto fixo de g em I, ou seja, único zero de f em I), usando a fórmula iterativa

$$x_{k+1} = g(x_k);\ k = 0, 1, \ldots,$$

com x_0 escolhido arbitrariamente em I. A este método chamamos **método do ponto fixo** ou **método das iterações sucessivas**.

Método das iterações sucessivas (ou do ponto fixo)

Interpretação geométrica



A verificação de que uma função g é contrativa num intervalo $I = [a, b]$ poderá tornar-se mais simples recorrendo ao seguinte teorema.

Teorema

Seja $I = [a, b]$, com $a, b \in \mathbb{R}$ e seja $g : I \to \mathbb{R}$ uma função de classe C^1 em I e tal que

$$L = \max_{x \in I} |g'(x)| < 1.$$

Então g é contrativa em I, com constante de Lipschitz L.

Dem: Pelo Teorema do valor médio de Lagrange, dados quaisquer $x, y \in I$, sabemos que existe sempre ξ entre x e y tal que $g(x) - g(y) = g'(\xi)(x - y)$. Assim, tem-se

$$|g(x) - g(y)| = |g'(\xi)||x - y| \leq L|x - y|$$

e como, por hipótese, $L < 1$, g é contrativa em I.

Nota: A conclusão do teorema continua válida se I for um intervalo não limitado (i.e.,se I for da forma $I = [a, +\infty)$, $I = (-\infty, b]$ ou mesmo $I = \mathbb{R}$), se f for apenas contínua em I e com derivada no interior de I, desde que substituamos a condição $L = \max_{x \in I} |g'(x)| < 1$ por $L = \sup_{x \in \text{int}(I)} |g'(x)| < 1$.

Exemplo (Método do ponto fixo)

Considere-se o problema da determinação de raízes da equação $f(x) = 0$, com $f(x) = x^3 - x - 1$.
Tem-se

$$x^3 - x - 1 = 0 \Leftrightarrow x^3 = x + 1 \Leftrightarrow x = \sqrt[3]{x+1}.$$

Seja então $g(x) = \sqrt[3]{x+1}$. Considere-se o intervalo $I = [1, 1.5]$ (esta escolha é sugerida por ser fácil de verificar que f tem uma raiz nesse intervalo) e vejamos que g satisfaz as condições de aplicação do Teorema do Ponto Fixo nesse intervalo. Tem-se

$$g(x) = \sqrt[3]{x+1}, \qquad g'(x) = \frac{1}{3\sqrt[3]{(x+1)^2}}, \qquad I = [1, 1.5]$$

Então:

- $g(1) \approx 1.2599 \in I$
- $g(1.5) \approx 1.3572 \in I$
- $g'(x) > 0, \forall x \in I \Rightarrow g$ é crescente em I

$\Rightarrow g(I) \subseteq I.$

Exemplo (cont.)

É fácil de verificar que g é de classe C^1 em $I = [1, 1.5]$ e que

$$L = \max_{x\in[1,1.5]} |g'(x)| = \max_{x\in[1,1.5]} \frac{1}{3\sqrt[3]{(x+1)^2}} = \frac{1}{3\sqrt[3]{4}} \approx 0.22 < 1.$$

Então, g é contrativa em I (com constante de Lipschitz $L \approx 0.22$).
Como $g(I) \subseteq I$ e g é contrativa em I, podemos, então, aplicar o Teorema do Ponto Fixo para procurar o (único) ponto fixo de g em I; esse ponto fixo de g é, naturalmente, a única raiz da equação $f(x) = 0$ em I.

Exemplo (cont.)

Na tabela seguinte registam-se algumas iterações obtidas usando o Método do Ponto Fixo, com aproximação inicial $x_0 = 1.25$, usando a função iterativa g, i.e., usando a fórmula

$$x_{k+1} = \sqrt[3]{x_k + 1}; k = 0, 1, 2, \ldots$$

k	x_k
1	1.3103707
2	1.3219871
3	1.3241990
4	1.3246194
5	1.3246992
6	1.3247144
7	1.3247173
8	1.3247178

Note-se que a única raiz real da equação $x^3 - x - 1 = 0$ é (com precisão de 8 c.d) $r = 1.3247180$.

Convergência local do método do ponto fixo

Teorema

Seja α um ponto fixo de uma função g e suponhamos que g é continuamente diferenciável num certo intervalo centrado em α e que, além disso,

$$|g'(\alpha)| < 1.$$

Então, o método do ponto fixo definido por $x_{k+1} = g(x_k)$ converge localmente para α, i.e., existe um intervalo $I = [\alpha - \delta, \alpha + \delta]$ tal que, se tomarmos a aproximação inicial x_0 em I, as iterações sucessivas convergem para α.

Dem: Vamos mostrar que existe um intervalo $I = [\alpha - \delta, \alpha + \delta]$ tal que

- $L = \max_{x\in I} |g'(x)| < 1$
- $g(I) \subseteq I.$

Se assim for, ter-se-á que g transforma I em si mesmo e é contrativa em I, pelo que estarão satisfeitas as condições do Teorema do Ponto Fixo de Banach, e o resultado do teorema seguir-seá de imediato.

Demonstração (cont.)

- Seja $\epsilon := \frac{1-|g'(\alpha)|}{2}$; como $|g'(\alpha)| < 1$, segue-se que $\epsilon > 0$. Mas, sendo g' contínua em α, existirá então $\delta > 0$ tal que

$$|x-\alpha| \leq \delta \Longrightarrow |g'(x) - g'(\alpha)| \leq \epsilon = \frac{1-|g'(\alpha)|}{2}.$$

Seja, então,

$$I := [\alpha - \delta, \alpha + \delta]$$

e mostremos, primeiramente, que $|g'(x)| < 1$, para todo o $x \in I$.
Se $x \in I$, isto é, se $|x-\alpha| \leq \delta$, então virá:

$$\begin{aligned} |g'(x)| &= |g'(x) - g'(\alpha) + g'(\alpha)| \leq |g'(x) - g'(\alpha)| + |g'(\alpha)| \\ &\leq \frac{1-|g'(\alpha)|}{2} + |g'(\alpha)| = \frac{1+|g'(\alpha)|}{2} < 1. \end{aligned}$$

Demonstração (cont.)

- Mostremos agora que $g(I) \subseteq I$.

(Podemos, naturalmente, assumir que δ foi escolhido suficientemente pequeno de forma a garantir que o intervalo $I = [\alpha - \delta, \alpha + \delta]$ está contido no intervalo onde g é continuamente diferenciável.)

Seja $x \in I$, isto é, seja x tal que $|x-\alpha| \leq \delta$. Pretendemos mostrar que $g(x) \in I$, ou seja que $|g(x) - \alpha| \leq \delta$.

Mas,

$$\begin{aligned} |g(x) - \alpha| = |g(x) - g(\alpha)| &= |g'(\xi)||x-\alpha| \quad (\xi \text{ entre } \alpha \text{ e } x) \\ &< |x-\alpha| \leq \delta, \end{aligned}$$

como queríamos provar.

Notas

1. O teorema anterior mostra que, se α for um ponto fixo de g e $|g'(\alpha)| < 1$ (sendo g' contínua numa vizinhança de α), então podemos aplicar o método das iterações sucessivas para aproximar α, desde que x_0 seja escolhido "suficientemente" próximo de α.
2. Suponhamos, agora, que $|g'(\alpha)| > 1$. Tem-se

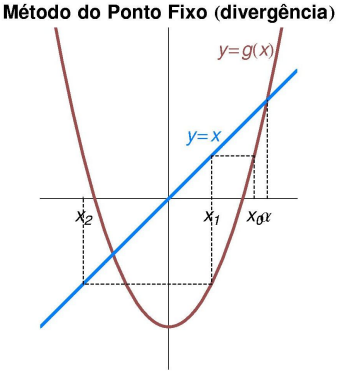
$$|\alpha - x_{k+1}| = |g(\alpha) - g(x_k)| = |g'(\xi_k)||\alpha - x_k|.$$

Mas, para x_k "próximo" de α, será ξ_k "próximo" de α, pelo que $|g'(\xi_k)| > 1$ (relembre-se que g' é, por hipótese, contínua numa vizinhança de α). Isto significa que $|e_{k+1}| > |e_k|$, pelo que, neste caso, o método das iterações sucessivas não poderá convergir[6]; ver exemplo na figura seguinte.

3. Se $|g'(\alpha)| = 1$, não podemos, à partida, tirar qualquer conclusão sobre a convergência/não convergência do método.

[6] Excepto, se para algum k, se tiver $x_k = \alpha$.

Método do ponto fixo
Divergência – ilustração gráfica



Método do ponto fixo

Ordem de convergência (caso $0 < |g'(\alpha)| < 1$)

Teorema

Seja α um ponto fixo de uma função g e suponhamos que g é continuamente diferenciável num certo intervalo centrado em α e que, além disso,

$$0 < |g'(\alpha)| < 1.$$

Seja I o intervalo (que sabemos existir) para o qual se verificam as condições do teorema do ponto fixo, e suponhamos que $x_0 \in I$ e que $x_{k+1} = g(x_k)$; $k = 0, 1, \ldots$. Designando por e_k o erro na iteração k, isto é, sendo $e_k := \alpha - x_k$, (e admitindo que $e_k \neq 0$, para todo o k), tem-se,

$$\lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|} = |g'(\alpha)|.$$

Por outras palavras, nestas condições o método converge linearmente, com constante de convergência assintótica $C = |g'(\alpha)|$.

Demonstração

Tem-se

$$\begin{aligned} |e_{k+1}| &= |\alpha - x_{k+1}| = |g(\alpha) - g(x_k)| \\ &= |g'(\xi_k)||\alpha - x_k| = |g'(\xi_k)||e_k|, \quad \xi_k \text{ entre } \alpha \text{ e } x_k. \end{aligned}$$

Como $x_k \to \alpha$, também $\xi_k \to \alpha$ e, uma vez que g' é contínua em α, ter-seá

$$\lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|} = \lim_{k\to\infty} |g'(\xi_k)| = |g'(\alpha)|,$$

O teorema anterior mostra que a convergência do método do ponto fixo será tanto mais rápida quanto menor for o valor de $C = |g'(\alpha)|$. Pode acontecer que $g'(\alpha) = 0$. Nesse caso, será de esperar que a convergência seja "melhor" que linear. De facto, tem-se o resultado mais geral, contido no teorema seguinte.

Método do ponto fixo

Ordem de convergência (caso $g^{(k)}(\alpha) = 0; k = 0, 1, \ldots, p-1; g^{(p)}(\alpha) = 0$)

Teorema

Seja α um ponto fixo de uma função g e suponhamos que g é p ($p \geq 2$) vezes continuamente diferenciável num certo intervalo centrado em α e que, além disso, $g'(\alpha) = g''(\alpha) = \cdots g^{(p-1)}(\alpha) = 0$ e que $g^{(p)}(\alpha) \neq 0$. Seja (x_k) a sequência de iterações obtida por aplicação do método do ponto fixo, com x_0 escolhido suficientemente próximo de α. Então, se $e_k = \alpha - x_k \neq 0$ para todo o k, tem-se

$$\lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^p} = \frac{1}{p!}|g^{(p)}(\alpha)|.$$

Por outras palavras, nestas condições, a ordem de convergência do método é p e a constante de convergência assintótica é $C = \frac{1}{p!}|g^{(p)}(\alpha)|$.

Dem: Ao cuidado dos alunos.

Convergência do método do ponto fixo (resumo)

Seja α um ponto fixo de uma função g e suponhamos que g é $p \geq 1$ vezes continuamente diferenciável num certo intervalo centrado em α. Então,

- se $|g'(\alpha)| < 1$, existe um intervalo I centrado em α, tal que, com $x_0 \in I$, o método do ponto fixo com função iterativa g converge para α; além disso, sendo $e_k = x_k - \alpha$, tem-se (admitindo que $e_k \neq 0$, paar todo o k)
  - se $g'(\alpha) \neq 0$, então

$$\lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|} = |g'(\alpha)| \to \text{conv. linear}$$

  - se $g'(\alpha) = \cdots = g^{(p-1)}(\alpha) = 0$ e $g^{(p)}(\alpha) \neq 0$ ($p \geq 2$), então

$$\lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^p} = \frac{|g^{(p)}(\alpha)|}{p!} \to \text{ordem de conv. } p$$

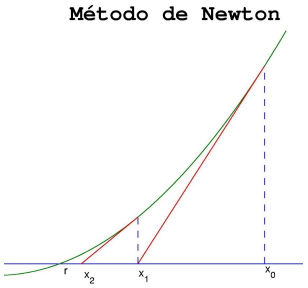
- se $|g'(\alpha)| > 1$, o método do ponto fixo com função iterativa g não converge para α por mais próxima de α que esteja a aproximação inicial x_0 (exceto se, para algum k, for $x_k = \alpha$).

Método de Newton

Se a função f cuja raiz procuramos for continuamente diferenciável e se a derivada de f puder ser calculada sem grande esforço computacional, poderemos procurar a raiz usando o chamado método de Newton.

Em 1669, Newton desenvolveu este método para calcular uma raiz de uma equação polinomial de $3^{\underline{o}}$ grau e em 1690, Joseph Raphson formulou as ideias de Newton para o caso de um polinómio, numa forma mais semelhante à que hoje se utiliza. Por esta razão, este método é também conhecido como método de Newton-Raphson.

Seja x_0 uma aproximação inicial (razoável) para um zero r de f. Consideremos a reta tangente ao gráfico de f no ponto de abcissa x_0.



Na vizinhança de x_0, esta reta deverá "aproximar razoavelmente" o gráfico de f, pelo que a abcissa do ponto de interseção desta tangente com o eixo das abcissas deverá estar próxima da raiz r. Essa abcissa será, então, tomada como uma nova aproximação, x_1, para r.

Método de Newton

Vejamos qual a expressão analítica de x_1. A equação da referida tangente é

$$y = f(x_0) + f'(x_0)(x - x_0).$$

Fazendo $y = 0$ e $x = x_1$ na equação anterior, e resolvendo em ordem a x_1, obtém-se

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

O processo anterior pode ser repetido, obtendo-se uma sequência (x_k) definida por

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}; k = 0, 1, 2, \ldots$$

O método iterativo dado pela fórmula anterior constitui, precisamente, o método de Newton.

Nota: O esquema iterativo anterior pressupõe que, para todo o k, $f'(x_k) \neq 0$, ou seja, que a tangente ao gráfico de f no ponto $(x_k, f(x_k))$ não é paralela ao eixo das abcissas.

Método de Newton como método de ponto fixo

Para estudar a convergência do método de Newton, começamos por notar que ele pode ser interpretado como um método de ponto fixo. Com efeito, para x tal que $f'(x) \neq 0$ tem-se tem-se

$$f(x) = 0 \iff x - \frac{f(x)}{f'(x)} = x \iff g(x) = x,$$

onde g é a função definida por

$$g(x) = x - \frac{f(x)}{f'(x)}. \tag{7}$$

Por outro lado, a fórmula iterativa do método de Newton pode escrever-se, usando a função g definida acima, como

$$x_{k+1} = g(x_k).$$

Em conclusão: o método de Newton não é mais do que é o método das iterações sucessivas com função iterativa g dada pela fórmula (7).

Método de Newton - Convergência

Teorema (convergência do método de Newton para raízes simples)

Seja r uma raiz da equação $f(x) = 0$ e suponhamos que f é três vezes continuamente diferenciável num certo intervalo centrado em r[a] e que $f'(r) \neq 0$, ou seja, que r é um zero simples de f. Então, existe um intervalo $I = [r - \delta, r + \delta]$ tal que, para qualquer aproximação inicial $x_0 \in I$, o método de Newton converge para r, com convergência (pelo menos) quadrática; no caso de convergência quadrática a constante de convergência assintótica é dada por

$$C = \frac{1}{2}\left|\frac{f''(r)}{f'(r)}\right|.$$

[a] Exigimos aqui que f seja três vezes continuamente diferenciável para podermos invocar os resultados de convergência do método do ponto fixo; no entanto, pode mostrar-se que os resultados do teorema se mantêm, se f for apenas duas vezes continuamente diferenciável numa vizinhança de r.

Demonstração

Como $f'(r) \neq 0$ e f' é contínua num intervalo centrado em r, é possível encontrar um intervalo centrado em r no qual f' nunca se anula. Então, nesse intervalo, o método de Newton pode ser visto como um método do ponto fixo, com função iterativa $g(x) = x - \frac{f(x)}{f'(x)}$. Além disso, tem-se

$$g'(x) = 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2}$$

e

$$\begin{aligned} g''(x) &= \frac{(f'(x))^2\left[f'(x)f''(x) + f(x)f'''(x)\right] - 2f(x)f'(x)(f''(x))^2}{(f'(x))^4} \\ &= \frac{(f'(x))^2 f''(x) + f(x)f'(x)f'''(x) - 2f(x)(f''(x))^2}{(f'(x))^3}. \end{aligned}$$

Logo, g é duas vezes continuamente diferenciável num intervalo centrado em r e

$$g'(r) = \frac{f(r)f''(r)}{(f'(r))^2} = 0.$$

Demonstração (cont.)

Tendo em conta os resultados sobre a convergência do método do ponto fixo (teorema da p. 34), podemos concluir que, desde que x_0 seja suficientemente próximo de r (i.e. para x_0 num determinado intervalo $I = [r - \delta, r + \delta]$), o método converge para r (pelo menos) quadraticamente; além disso, se a convergência for quadrática, a constante de convergência assintótica será dada por

$$\frac{1}{2}|g''(r)| = \frac{1}{2}\left|\frac{(f'(r))^2 f''(r)}{(f'(r))^3}\right| = \frac{1}{2}\left|\frac{f''(r)}{f'(r)}\right|,$$

como queríamos demonstrar.

Notas:

1. A convergência será quadrática se $f''(r) \neq 0$, sendo superior a 2 se $f''(r) = 0$.
2. Se r for uma raiz múltipla, então pode provar-se que convergência do método de Newton para essa raiz será apenas linear.

Método de Newton

Exemplo

Retomemos o exemplo da equação

$$f(x) = x^3 - x - 1 = 0,$$

a qual, como vimos, tem uma (única) raiz real no intervalo $I = [1, 1.5]$. Aplicando o método de Newton, com aproximação inicial $x_0 = 1.25$, obtêm-se os resultados constantes da tabela seguinte.

k	x_k
1	1.3305085
2	1.3247490
3	1.3247180
4	1.3247180

Recorde-se que a raiz real da equação $x^3 - x - 1 = 0$ é (com precisão de 8 c.d) $r = 1.3247180$.

Método de Newton

O teorema seguinte estabelece condições suficientes de convergência do método de Newton, para qualquer aproximação inicial num intervalo $I = [a, b]$, que são, por vezes, fáceis de verificar na prática.

Teorema

Seja f uma função definida num certo intervalo $[a, b]$ de $\mathbb{R}$, que verifique as seguintes condições:

1. $f \in C^2[a, b]$;
2. $f(a)f(b) < 0$;
3. $f'(x) \neq 0, \forall x \in [a, b]$;
4. $f''(x) \geq 0, \ \forall x \in [a, b]$ *ou* $f''(x) \leq 0, \ \forall x \in [a, b]$;
5. $$\frac{|f(a)|}{|f'(a)|} < b - a \quad \text{e} \quad \frac{|f(b)|}{|f'(b)|} < b - a.$$

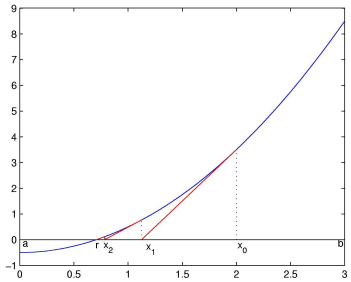
Então, a equação $f(x) = 0$ tem uma única raiz em $[a, b]$ e o método de Newton converge para essa raiz, seja qual for a aproximação inicial $x_0 \in [a, b]$.

Demonstração

Como f é contínua em $[a, b]$ e $f(a)f(b) < 0$, o Teorema do Valor Intermédio garante que f tem um zero em $[a, b]$; como f' é contínua e nunca se anula em $[a, b]$, f é estritamente monótona em $[a, b]$, pelo que esse zero é único; denotemo-o por r. Naturalmente r é o único ponto fixo em $[a, b]$ da função iterativa do método de Newton, $g(x) = x - \frac{f(x)}{f'(x)}$.
Vamos demonstrar apenas o caso (ilustrado na figura seguinte) em que

$$f(a) < 0 \ \text{ e } \ f''(x) \geq 0, \ \forall x \in [a, b],$$

podendo a demonstração dos restantes casos ser feita de modo análogo. Note-se, que neste caso, terá de ser $f'(x) > 0, \forall x \in [a, b]$.



Demonstração (cont.)

Recorde-se que temos $g'(x) = \frac{f(x)f''(x)}{(f'(x))^2}$, pelo que teremos a situação descrita na tabela seguinte:

	a	r	b
f	$-$	0	$+$
f'	$+$	$+$	$+$
f''	$+$	$+$	$+$
g'	$-$	0	$+$
g	$\searrow$	r	$\nearrow$

- Comecemos por considerar o caso em que $x_0 \in [r, b]$.

  Neste caso, será $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \leq x_0$, já que $f(x_0) \geq 0$ e $f'(x_0) > 0$.

  Além disso, como g é crescente em $[r, b]$, teremos $g(x_0) \geq g(r)$, ou seja, teremos $x_1 \geq r$.

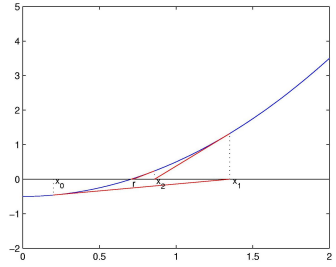
Demonstração (cont.)

O raciocínio repete-se, então, para as iterações sucessivas, isto é, tem-se

$$x_0 \geq x_1 \geq x_2 \geq x_3 \geq \ldots \geq r.$$

A sucessão (x_k) é, portanto, uma sucessão monótona decrescente e limitada inferiormente, pelo que admite limite. A continuidade de g garante que o limite dessa sucessão é o (único) ponto fixo de g em $[a, b]$, ou seja, é a raiz r da equação $f(x) = 0$ em $[a, b]$.

- Consideremos, agora, o caso em que $x_0 \in [a, r]$.



Demonstração (cont.)

Como g é decrescente em $[a, r]$, teremos

$$x_1 = g(x_0) \geq g(r) = r.$$

Se mostrarmos que $x_1 \leq b$ cairemos de novo no caso anterior (a partir de x_1, naturalmente) e nada mais haverá a demonstrar.
Invocando novamente o facto de g ser decrescente em $[a, r]$, concluímos que

$$x_1 = g(x_0) \leq g(a) = a - \frac{f(a)}{f'(a)}.$$

Mas, usando a condição **5.**, podemos garantir que $|\frac{f(a)}{f'(a)}| = -\frac{f(a)}{f'(a)} < b - a$, pelo que virá

$$x_1 < a + (b - a) = b,$$

como pretendíamos verificar.

Nota

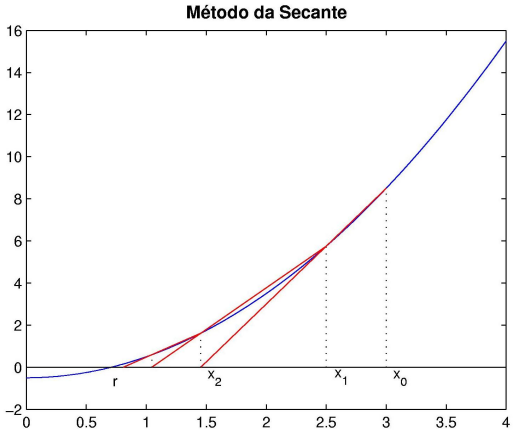
Se f for duas vezes continuamente diferenciável em $\mathbb{R}$, tiver um zero em $\mathbb{R}$ e as condições 3. e 4. do teorema se verificarem para qualquer $x \in \mathbb{R}$, então o método de Newton converge para r para qualquer aproximação inicial $x_0 \in \mathbb{R}$.

Método da secante

O método de Newton exige cálculo de um valor da função e de um valor da sua derivada, por iteração ⇒ caro!
Um método alternativo é o chamado método da secante, em que a tangente ao gráfico de f no ponto de abcissa x_k é substituída pela recta "secante" que une os pontos $(x_{k-1}, f(x_{k-1}))$ e $(x_k, f(x_k))$. Isto é, dadas **duas** aproximações x_{k-1} e x_k para r, a nova aproximação é dada por

$$\begin{aligned} x_{k+1} &= x_k - \frac{f(x_k)}{\frac{f(x_k)-f(x_{k-1})}{x_k - x_{k-1}}} \\ &= x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \end{aligned}$$

Método da secante
Interpretação geométrica



Método da secante – convergência

Contrariamente ao método de Newton, o método da secante não pode ser visto como um método do ponto fixo, pelo que a análise da sua convergência não pode basear-se nos teoremas estudados para esse método. No entanto, pode provar-se que as condições de convergência do método de Newton são também condições de convergência do método da secante. Mais precisamente, tem-se o seguinte teorema, "análogo" ao teorema sobre a convergência do método de Newton.

Teorema

Seja r uma raiz da equação $f(x) = 0$ e suponhamos que f é duas vezes continuamente diferenciável num intervalo centrado em r e que $f'(r) \neq 0$ (ou seja , que r é um zero simples de f). Então, existe um intervalo $I = [r - \delta, r + \delta]$ tal que, para quaisquer aproximações iniciais $x_0, x_1 \in I$, o método da secante converge para r, com ordem de convergência $p = \frac{1+\sqrt{5}}{2}$ e constante de convergência assimptótica dada por $C = \left|\frac{f''(r)}{2f'(r)}\right|^{1/p}$.

Dem: Veja, e.g. Atkinson, K.E., *An Introduction to Numerical Analysis*, John Wiley, 1988.

Exemplo

Retomemos novamente o exmplo da equação

$$f(x) = x^3 - x - 1 = 0.$$

Aplicando o método da secante, com aproximções iniciais $x_0 = 1$ e $x_1 = 1.5$, obtêm-se os resultados constantes da tabela seguinte.

k	x_k
1	1.2666667
2	1.3159617
3	1.3252141
4	1.3247139
5	1.3247180
6	1.3247180

Recorde-se que a raiz real da equação é (com precisão de 8 c.d) $r = 1.3247180$.

Critérios de paragem

Tal como para o caso dos métodos iterativos para a resolução de sistemas lineares, quando se usa qualquer um dos métodos referidos neste capítulo, também há necessidade de definir critérios de paragem. Neste caso, são, geralmente, usados critérios da seguinte forma:

1. $|x_{k+1} - x_k| < \epsilon_1$
2. $|x_{k+1} - x_k| < \epsilon_2 |x_{k+1}|$
3. $|f(x_k)| < \epsilon_3$

Uma vez mais, tendo em atenção que o processo pode divergir ou convergir de forma muito lenta, deverá também ser incluído um critério de paragem do tipo:

4. parar quando $k > kmax$, onde $kmax$ é um inteiro escolhido pelo utilizador, correspondente ao número máximo de iterações a efetuar.

Breve comparação dos métodos

- *Método da bissecção*
  - ↑ Convergência global
  - ↓ Convergência lenta (linear, razão $1/2$)
  - ↑ Apenas um cálculo do valor de f num ponto, por iteração.
- *Método do ponto fixo*
  - ↓ Convergência local
  - ↓ Em geral, convergência não muito rápida (linear)
  - ↑ Apenas um cálculo do valor de g num ponto, por iteração.
- *Método de Newton*
  - ↓ Convergência local
  - ↑ Convergência muito rápida (quadrática)
  - ↓ Cálculo do valor de f e de f' num ponto, por iteração.
- *Método da secante*
  - ↓ Convergência local
  - ↔ Convergência rápida (superlinear)
  - ↑ Cálculo do valor de f num ponto, por iteração (à excepção da primeira).

Métodos híbridos

- Métodos de convergência rápida, tais como o de Newton ou secante, têm convergência local, i.e. podem não convergir se a aproximação inicial não for suficientemente próxima da raiz.
- Métodos seguros, como o da bissecção, são lentos.
- **Métodos híbridos** tentam combinar características dos dois tipos de métodos, de forma a conseguir-se rapidez e segurança.
- Ideia: usar um método de convergência rápida, mas manter sempre um encaixe para a raiz
  - se a próxima iteração do método rápido sair fora de um intervalo que encaixa a raiz, executar um passo de um método seguro (e.g.bissecção);
  - tentar executar o método rápido no novo subintervalo.
- Função **fzero** do MATLAB implementa um método híbrido (método Dekker-Brent, baseado em bissecção, secante e interpolação quadrática inversa).

Sistemas não lineraes

O caso da resolução de um sistema de equações não lineares é bastante mais complexo do que o caso escalar:

- Existe uma muito maior variedade de comportamentos, pelo que saber se existe ou não solução do sistema e, em caso afirmativo, saber quantas soluções existem é difícil;
- determinar uma "boa" aproximação inicial é difícil;
- não existe um resultado tipo do valor intermédio que garanta "intervalo" contendo a raiz;
- esforço computacional dos métodos cresce rapidamente com tamanho do sistema.

Método de Newton para sistemas

Vamos descrever de uma forma muito sucinta como aplicar o método de Newton para sisetmas de equações não lineares. Considere-se um sistema de n equações não lineraes em n incógintas:

Sistemas não lineares

O caso da resolução de um sistema de equações não lineares é bastante mais complexo do que o caso escalar:

- Existe uma muito maior variedade de comportamentos, pelo que saber se existe ou não solução do sistema e, em caso afirmativo, saber quantas soluções existem é difícil;
- determinar uma "boa" aproximação inicial é, geralmente, difícil;
- não existe um resultado do tipo do Teorema do Valor Intermédio que garanta uma região contendo a raiz;
- o esforço computacional dos métodos cresce rapidamente com tamanho do sistema.

Método de Newton para sistemas não lineares

Vamos descrever, de uma forma muito sucinta, como aplicar o método de Newton para sistemas de equações não lineares. Considere-se um sistema de n equações não lineares em n incógintas

$$\begin{cases} f_1(x_1, x_2, \ldots, x_n) = 0 \\ f_2(x_1, x_2, \ldots, x_n) = 0 \\ \vdots \\ f_n(x_1, x_2, \ldots, x_n) = 0 \end{cases}$$

ou, usando notação vetorial,

$$\mathbf{f}(\mathbf{x}) = \mathbf{0},$$

onde $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ e $\mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}$.

Método de Newton para sistemas

O método de Newton para o sistema anterior é definido por

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left(\mathbf{J}(\mathbf{x}^{(k)})\right)^{-1} \mathbf{f}(\mathbf{x}^{(k)}), k = 0, 1, 2 \ldots;$$
$$\mathbf{x}^{(0)} \in \mathbb{R}^n \text{ dado,}$$

onde $\mathbf{J}(\mathbf{x})$ designa a matriz Jacobiana da função $\mathbf{f}$ calculada em $\mathbf{x}$, i.e.

$$\mathbf{J}(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \frac{\partial f_1}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{x}) & \frac{\partial f_2}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_2}{\partial x_n}(\mathbf{x}) \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial f_n}{\partial x_1}(\mathbf{x}) & \frac{\partial f_n}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_n}{\partial x_n}(\mathbf{x}) \end{bmatrix}$$

Método de Newton para sistemas (cont.)

Temos

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} \underbrace{- \left(\mathbf{J}(\mathbf{x}^{(k)})\right)^{-1} \mathbf{f}(\mathbf{x}^{(k)})}_{\mathbf{s}^{(k)}}.$$

Na prática, não invertemos explicitamente a matriz Jacobiana $\mathbf{J}(\mathbf{x}^{(k)})$, mas, em vez disso:

- resolvemos o sistema

  $$\mathbf{J}(\mathbf{x}^{(k)})\mathbf{s}^{(k)} = -\mathbf{f}(\mathbf{x}^{(k)})$$

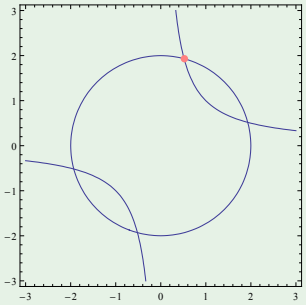
  para determinar o incremento $\mathbf{s}^{(k)}$;
- calculamos a próxima iteração, usando

  $$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{s}^{(k)}.$$

Exemplo

Considere-se o sistema

$$\begin{cases} x^2 + y^2 - 4 = 0 \\ xy - 1 = 0 \end{cases}$$



Uma das suas soluções, $\mathbf{r}$, é o ponto assinalado na figura; vamos determinar essa solução, usando o método de Newton com aproximação inicial $\mathbf{x}^{(0)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

Exemplo (cont.)

Façamos $x_1 = x, x_2 = y$; temos

$$\mathbf{f} = \begin{bmatrix} x_1^2 + x_2^2 - 4 \\ x_1 x_2 - 1 \end{bmatrix}$$

e

$$\mathbf{J}(\mathbf{x}) = \begin{bmatrix} 2x_1 & 2x_2 \\ x_2 & x_1 \end{bmatrix}$$

Vejamos, então, como usar o método no Matlab.

Exemplo (cont.)

```
>> f=@(x) [x(1)^2+x(2)^2-4; x(1)*x(2)-1];
>> J=@(x) [2*x(1), 2*x(2); x(2), x(1)];
>> x0=[0;1];
>> s0=-J(x0)\f(x0); x1=x0+s0
x1 =
    1.0000
    2.5000
>> s1=-J(x1)\f(x1); x2=x1+s1
x2 =
    0.5952
    2.0119
>> s2=-J(x2)\f(x2); x3=x2+s2
x3 =
    0.5200
    1.9342
```


Exemplo (cont.)

```
>> s3=-J(x3)\f(x3); x4=x3+s3
x4 =
    0.5176
    1.9319
>> s4=-J(x4)\f(x4); x5=x4+s4
x5 =
    0.5176
    1.9319
>> format long
>> x5
  x5 =
     0.517638090207228
     1.931851652580324
```


Exemplo (cont.)

```
>> format short
>> norm(s4,inf)
ans =
  2.3144e-006
```

Temos $\|\mathbf{s}^{(4)}\|_\infty = \|\mathbf{x}^{(5)} - \mathbf{x}^{(4)}\|_\infty = 2.3144 \times 10^{-6}$, pelo que $\|\mathbf{x}^{(5)} - \mathbf{r}\| \approx 2.3144 \times 10^{-6}$, ou seja $\begin{bmatrix} 0.51764 \\ 1.93185 \end{bmatrix}$ é uma aproximação para a solução $\mathbf{r}$ cujas componentes têm (em princípio) 5 casas decimais corretas.

Notas

- Se $\mathbf{r}$ for uma raiz de $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, se as funções f_i tiverem derivads parciais (relativamente a cada uma ads variáveis x_i) de primeira e segunda ordem contínuas para todos os pontos numa certa vizinhança de $\mathbf{r}$ e se a matriz Jacobiana $\mathbf{J}'(\mathbf{r})$ for invertível, então o método de Newton converge (localmente) para $\mathbf{r}$, com convergênca quadrática; em geral, é necessária uma boa aproximação inicial $\mathbf{x}_0$ para garantir a convergência.
- O custo do método de Newton para sistemas de ordem n é muito elevado, já que, em cada iteração, há que:
  - calcular a matriz Jacobiana $\mathbf{J}(\mathbf{x}^{(k)})$ ($\Rightarrow n^2$ cálculos de valores de funções);
  - resolver um sistema ($\Rightarrow \mathcal{O}(n^3)$ operações).

Os chamados métodos de atualização da secante (exemplo: Método de Broyden):

- usam valores da função $\mathbf{f}$ em iterações sucessivas para obter uma aproximação para a matriz Jacobiana;
- em cada iteração, não calculam uma nova fatorização dessa matriz (necessária para resolver o sistema), mas simplesmente efetuam uma atualização da fatorização da iteração anterior.