Métodos Numéricos (M2039)

Maria João Rodrigues
Teresa Feio Mendonça

DM/FCUP – 2023/2024

Departamento de Matemática
Faculdade de Ciências da Universidade do Porto

Bibliografia

- Conte S., de Boor C. – Elementary Numerical Analysis, 1981
- Gautshi W. – Numerical analysis 2nd ed, 2012
- Mathews J., Fink K. – Numerical methods using Matlab, 2004
- Pina H. – Métodos Numéricos, 2010
- Quarteroni A., Sacco R., Saleri F. – Numerical mathematics, 2000
- Quarteroni A., Sacco R., Saleri F. – Scientific Computing with Matlab and Octave, 2006
- Redivo Zaglia M. – Calcolo Numerico metodi ed algoritmi, 2005

Temas

- Capítulo 1: Erros. Propagação dos erros. Erros na resolução numérica de sistemas de equações lineares.
- Capítulo 2: Resolução numérica de equações não lineares.
- Capítulo 3: Aproximação numérica. Interpolação. Aproximação no sentido dos mínimos quadrados.
- Capítulo 4: Derivação numérica e integração numérica.
- Capítulo 5: Resolução numérica de equações diferenciais.

Objetivos da Análise Numérica

Dado um problema matemático a Análise Numérica

- estuda condições suficientes para a existência e unicidade de solução;
- define um método e fornece um algoritmo para o cálculo da solução;
- calcula um valor aproximado da solução e controla o erro cometido, intervindo de forma a minimizar erros
  - de aproximação
  - de arredondamento
  - de propagação de erros anteriores

Fontes de erro

- **erros de modelação**: são usados modelos matemáticos para representar a realidade física - modelação matemática
- **erros de observação**: a maior parte dos dados resultam de observação (medição), ficando afetados de erros que vão ser transmitidos aos cálculos efetuados sobre eles. Os métodos numéricos não podem remover esses erros de observação mas podem intervir de forma a encontrar a melhor forma de minimizar os efeitos de propagação
- **erros de programação**: devem ser feitos testes de forma a garantir que os programas usados dão a solução correta num exemplo conhecido, para evitar fornecer uma solução disparatada
- **erros de aproximação**: erros que resultam de se utilizarem fórmulas aproximadas para o cálculo de determinadas funções, por exemplo:

$$e^x \approx 1 + x + \frac{x^2}{2!}$$

- **erros de arredondamento**: erros de representação dos números em aritmética em vírgula flutuante

Capítulo 1

Teoria de erros

- Representação dos números reais em computador.
- Tipos de erros.
- Erro absoluto e erro relativo. Erro de arredondamento. Casas decimais corretas e algarismos significativos.
- Fórmula fundamental do cálculo de erros. Erro na avaliação de funções.
- Cálculo aproximado da soma de série numérica convergente.
- Erros na resolução numérica de sistemas de equações lineares usando eliminação gaussiana.

Representação dos números reais em computador

Sistemas de vírgula flutuante

- Consideremos um número real a escrito normalizado numa base $\mathbf{b}$

$$a = \pm(0.a_1 a_2 a_3 \ldots)_{\mathbf{b}} \times \mathbf{b}^e = \pm(0.m)_{\mathbf{b}} \times \mathbf{b}^e$$

onde os a_i são dígitos na base $\mathbf{b}$, isto é $0 \leq a_i \leq \mathbf{b}-1$, $i = 1, \ldots,$

m é a mantissa e e o expoente da representação de a.

- Num computador um número real é habitualmente representado na base $\mathbf{b} = 2$.

O número de bits para representar um número é sempre finito, habitualmente dividido em 3 partes

$\pm$	expoente	mantissa

Por isso:

- apenas um número finito de números reais pode ser representado exatamente no computador. Todos os outros vão ser representados aproximadamente (com lei de corte ou arredondamento).
- o número zero tem uma representação especial.

Representação dos números reais em computador

- o conjunto dos números reais representáveis é limitado entre um valor mínimo absoluto a_{min} $(= \mathbf{b}^{L-1})$ e um valor máximo absoluto a_{max} $(= \mathbf{b}^U(1 - \mathbf{b}^{-n}))$, onde L e U são respetivamente o menor e maior expoentes representáveis e n é o número de bits para representar a mantissa.
- os números a tais que $|a| < a_{min}$ não são representáveis dando origem a um erro de *underflow*. Os números a tais que $|a| > a_{max}$ não são representáveis dando origem a um erro de *overflow*.

- VF($\mathbf{b}$, n, L, U) diz-se um sistema de vírgula flutuante na base $\mathbf{b}$ com n dígitos e designa o conjunto formado pelo zero e pelos números que podem ser escritos na forma normalizada como

$$\pm(0.a_1 a_2 a_3 \ldots a_n)_{\mathbf{b}} \times \mathbf{b}^e$$

onde $0 \leq a_i \leq \mathbf{b}-1$, $i = 1(1)n$, $a_1 \neq 0$, $L \leq e \leq U$, L, e e U inteiros.

Representação dos números reais em computador

A representação dos números num número de bits limitado dá origem a erros. Vejamos alguns exemplos em que se considera um sistema VF($\mathbf{1}0,\ 4,\ -9,\ 9$) com lei de arredondamento.

- Soma de números de ordem de grandeza diferentes:

  $S = 3.567 + 0.0004821 + 0.0004789 + 0.0004657$

  (1) $3.567 + 0.0004821 = 3.5674821$, mas o sistema VF($\mathbf{1}0,\ 4,\ -9,\ 9$) só tem 4 dígitos para representar a mantissa e portanto fica 3.567. Com as outras parcelas passa-se o mesmo.

  Assim $S = 3.567$.

  (2) $0.0004657 + 0.0004789 = 0.0009446$,
  $0.0009446 + 0.0004821 = 0.001427$, $0.001427 + 3.567 = 3.568$.

  O resultado é $S = 3.568$!

Representação dos números reais em computador

- Diferença de números muito próximos (cancelamento de dígitos significativos):

  $D = 3774 + 5.874 - 3779$

  (1) $3774 + 5.874 = 3780$
  $3780 - 3779 = 1$

  (2) $3774 - 3779 = -5$
  $5.874 - 5 = 0.874$!

- Exercício: Calcule, no computador, na sua linguagem preferida, $\sum_{i=1}^{100} 0.1$. Comente e justifique o resultado obtido.
- Exercício: Determine o valor de *eps*, epsilon máquina do seu computador (*eps* é a distância entre 1 e o primeiro número maior do que 1 representável no sistema de virgula flutuante). A distância entre 1000 e o primeiro número maior do que 1000 representável no sistema de virgula flutuante é *eps*?
- Exercício: Avaliar $f(x) = x(\sqrt{x+1} - \sqrt{x})$ para $x = 10^i,\ \ i = 1, 2, \ldots$. Justificar os resultados e propôr uma fórmula alternativa numericamente estável para o cálculo de $f(x)$.
- Veja: http: //catless.ncl.ac.uk/Risks/, usando a chave rounding, para uma lista de desastres provocados por erros de arredondamento.

Erro, erro absoluto

Seja X o valor exato de uma grandeza real e x um valor aproximado de X.

- $\Delta x = X - x$ é o **erro** de x em relação a X
- $|\Delta x|$ é o **erro absoluto** de x.
- Em geral dispomos de um majorante ε de $|\Delta x|$ tal que:

$$|X - x| \leq \varepsilon \iff x - \varepsilon \leq X \leq x + \varepsilon$$

- Ao mínimo do conjunto dos majorantes, é habitual chamar erro absoluto máximo de x.
- Um número com m casas decimais corretas (ie, corretamente arredondadas) é suposto afetado de um erro absoluto no máximo de

$$5 \times 10^{-(m+1)}$$

Erro relativo

O erro absoluto não dá ideia do rigor com que uma medição foi feita. Por exemplo considere-se um erro de um metro na medição da distância da Terra ao Sol, ou na medição da altura de um aluno. Define-se então

- $\left|\frac{\Delta x}{X}\right|$, **erro relativo** com que x representa X.

Exemplo: Sejam $X = 1/3$ e $Y = 1/3000$. Considerem-se as aproximações $x = 0.3333$ e $y = 0.0003$. Os erros absolutos são iguais $\Delta x = \Delta y \approx 3.3 \times 10^{-5}$ e no entanto o erro relativo de x é $\left|\frac{\Delta x}{X}\right| \approx 10^{-4}$ e o de y é $\left|\frac{\Delta y}{Y}\right| \approx 10^{-1}$.

- Conhecido um majorante, ε, do erro absoluto $|\Delta x|$ tem-se $|\Delta x| \leq \varepsilon$ e

$$|X| = |x + \Delta x| = |x - (-\Delta x)| \geq |x| - |\Delta x| \geq |x| - \varepsilon$$

e obtemos um majorante do erro relativo

$$\left|\frac{\Delta x}{X}\right| \leq \frac{\varepsilon}{|x| - \varepsilon}$$

Erro relativo

- Habitualmente $|\Delta x| \ll |x|$ e usa-se para estimativa do erro relativo

$$\left|\frac{\Delta x}{X}\right| \approx \left|\frac{\Delta x}{x}\right|$$

pois

$$\left|\frac{\Delta x}{X}\right| = \left|\frac{\Delta x}{x+\Delta x}\right| = \left|\frac{\Delta x}{x}\right| \left|\left(1+\frac{\Delta x}{x}\right)^{-1}\right| =$$

$$= \left|\frac{\Delta x}{x}\right| \left|1 - \frac{\Delta x}{x} + \left(\frac{\Delta x}{x}\right)^2 - \ldots\right| \approx \left|\frac{\Delta x}{x}\right|$$

porque se $|\Delta x| \ll |x|$ então $\left|\frac{\Delta x}{x}\right| \ll 1$

Nota: Dois números não nulos a e b dizem-se da mesma ordem de grandeza e escreve-se $|a| \approx |b|$ sse $1 \leq \frac{a}{b} < 10$ ou $0.1 \leq \frac{a}{b} < 1$. Diz-se que a é menosprezável com respeito a b e escreve-se $|a| \ll |b|$ se $\frac{a}{b} < 0.1$

Erro relativo

O erro relativo está associado ao número de algarismos significativos corretos. Seja $x \approx X$ um número com m algarismos significativos. Então

$$\left|\frac{\Delta x}{x}\right| \leq 5 \times 10^{-m}$$

Dem:
$x = a_0.a_1 \ldots a_{m-1} \times 10^p, \quad a_0 \neq 0, \quad p$ inteiro, $\Longrightarrow |\Delta x| \leq 5 \times 10^{-m+p}$

Ora

$$1 \times 10^p \leq |a_0| \times 10^p \leq |x| < 10 \times 10^p \iff 10^{-p-1} < \frac{1}{|x|} \leq \frac{1}{|a_0| \times 10^p} \leq 10^{-p}$$

$$\iff 10^{-p-1}|\Delta x| < \frac{|\Delta x|}{|x|} \leq \frac{|\Delta x|}{|a_0| \times 10^p} \leq 10^{-p}|\Delta x|$$

Como $|\Delta x| \leq 5 \times 10^{-m+p}$ fica

$$5 \times 10^{-(m+1)} < \frac{|\Delta x|}{|x|} \leq \frac{5 \times 10^{-m}}{|a_0|} \leq 5 \times 10^{-m}$$

Propagação do erro nos dados: erros numéricos na avaliação de uma função de uma variável independente

Problema

Dada uma função real f , x_0 um valor aproximado de x, com erro $\Delta x_0 = x - x_0$, qual é o erro com que $z_0 = f(x_0)$ aproxima $z = f(x)$?

- O erro $\Delta z_0 = z - z_0$ é tal que

$$\Delta z_0 = z - z_0 = f(x) - f(x_0) = f(x_0 + \Delta x_0) - f(x_0)$$

- Se f for contínua em $[x_0, x]$ e derivável em $]x_0, x[$, então

$$\exists\ c \in]x_0, x[: \quad z - z_0 = f(x) - f(x_0) = f'(c)(x - x_0)$$

isto é,

$$\exists\ c \in]x_0, x[: \quad \Delta z_0 = f'(c)\Delta x_0$$

e portanto

$$|\Delta z_0| = |f'(c)||\Delta x_0|$$

Propagação do erro

- Conhecido um majorante ε_0 de $|\Delta x_0|$, então o erro absoluto $|\Delta z_0|$ é tal que

$$|\Delta z_0| \leq |f'(c)|\varepsilon_0$$

e podemos calcular um majorante de $|\Delta z_0|$

$$|\Delta z_0| \leq \sup_{t \in I} |f'(t)|\varepsilon_0, \quad t \in I =]x_0 - \varepsilon_0, x_0 + \varepsilon_0[$$

- Na prática, se $|\Delta x_0|$ for pequeno, é habitual substituir c por x_0, e, menosprezando quantidades de 2^a ordem em Δx_0, calcular uma estimativa de $|\Delta z_0|$

$$|\Delta z_0| \approx |f'(x_0)|\varepsilon_0$$

Nota: Esta estimativa não tem significado se $f'(x_0) = 0$. Nesse caso devem usar-se, se possível, mais termos no desenvolvimento de Taylor:

$$f(x) = f(x_0) + f'(x_0)\Delta x_0 + \frac{1}{2} f''(x_0)(\Delta x_0)^2 + ...$$

Propagação do erro

- O erro relativo com que z_0 representa z pode ser estimado por

$$\left|\frac{\Delta z_0}{z}\right| \approx \left|\frac{\Delta z_0}{z_0}\right|$$

- Exercício: Dadas as expressões algébricamente equivalentes $\dfrac{1}{(2+\sqrt{3})^4}$ e $97-56\sqrt{3}$ calcule-as usando para $x=\sqrt{3}$ a aproximação $x_0 = 1.732$, obtida por arredondamento.

  Interprete os resultados e justifique-os.

Erros numéricos na avaliação de uma função de n variáveis independentes

Problema

Dada uma função real $w = f(x_1, x_2, \ldots, x_n)$ e sendo $x_i^0, \;\; i = 1(1)n$ valores aproximados de $x_i, \;\; i = 1(1)n$, qual é o erro com que $w_0 = f(x_1^0, x_2^0, \ldots, x_n^0)$ aproxima w?

- Queremos majorar ou estimar o erro $\Delta w_0 = w - w_0$ que resulta dos erros dos dados $\Delta x_i^0 = x_i - x_i^0, \;\; i = 1(1)n$
- $P_0 \equiv (x_1^0, x_2^0, \ldots, x_n^0)$
  $P \equiv (x_1, x_2, \ldots, x_n) = (x_1^0 + \Delta x_1^0, x_2^0 + \Delta x_2^0, \ldots, x_n^0 + \Delta x_n^0)$
- Se f for contínua num domínio que contenha o segmento $[P_0, P]$ e se existirem e forem contínuas em $]P_0, P[$ as derivadas parciais $\dfrac{\partial f}{\partial x_i}, \;\; i = 1(1)n$, então o teorema dos acréscimos finitos permite escrever

Erros numéricos na avaliação de uma função de n variáveis independentes

$$\Delta w_0 = w - w_0 = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)_C \Delta x_i^0$$

onde $C = (x_1^0 + \theta \Delta x_1^0, \ldots, x_n^0 + \theta \Delta x_n^0)$, $0 < \theta < 1$, isto é, C é um ponto interior do segmento $[P_0, P]$.

- E podemos, conhecidos majorantes ε_i de $|\Delta x_i^0|$, calcular um majorante do erro absoluto $|\Delta w_0|$:

$$|\Delta w_0| \leq \sum_{i=1}^{n} S_{x_i} \varepsilon_i$$

onde

$$S_{x_i} = \sup_T \left| \left(\frac{\partial f}{\partial x_i} \right)_T \right| \text{ e } T \in]x_1^0 - \varepsilon_1, \ldots, x_n^0 - \varepsilon_n), (x_1^0 + \varepsilon_1, \ldots, x_n^0 + \varepsilon_n)[$$

Erros numéricos na avaliação de uma função de n variáveis independentes

- ou calcular estimativas do erro absoluto $|\Delta w_0|$ e do erro relativo $\left| \frac{\Delta w_0}{w} \right|$:

$$|\Delta w_0| \approx \sum_{i=1}^{n} \left| \left(\frac{\partial f}{\partial x_i} \right)_{P_0} \right| \varepsilon_i$$

$$\left| \frac{\Delta w_0}{w} \right| \approx \frac{\sum_{i=1}^{n} \left| \left(\frac{\partial f}{\partial x_i} \right)_{P_0} \right|}{|f(P_0)|} \varepsilon_i$$

- Exercício: Calcular o valor de $0.12^{3.1}$ sabendo que os dados forma obtidos por arredondamento.

Erros numéricos na avaliação de uma função de n variáveis independentes

Exercício: Mostre que

- O erro absoluto de uma soma algébrica $X = \sum_{i=1}^{n} x_i$ é não superior à soma dos erros máximos absolutos das parcelas.
- O erro relativo de um produto $W = xy$ (ou de um quociente $W = x/y$) é a soma dos erros relativos dos fatores.

Então,

- o resultado de uma soma deve ser apresentado com o número de casas decimais da parcela que tiver menor número de casas decimais (cuidado quando n é muito grande!)
- num produto o resultado deve ser apresentado com o número de algarismos significativos do fator que tiver menor número de algarismos significativos.

Erros na resolução numérica de sistemas de séries numéricas

Cálculo de séries numéricas

- Dada uma série de termos reais

$$\sum_{k=1}^{\infty} a_k = \underbrace{a_1 + a_2 + \cdots + a_n}_{S_n} + \underbrace{a_{n+1} + a_{n+2} + \ldots}_{R_n}$$

$S_n \rightarrow$ soma parcial; $\quad R_n \rightarrow$ erro de truncatura

- Se a série for convergente podemos escrever

$$S = \sum_{k=1}^{\infty} a_k \iff S = S_n + R_n \iff S - S_n = R_n$$

e então

$$|S - S_n| = |R_n|$$

isto é, o erro absoluto com que S_n representa S é o valor absoluto do erro de truncatura.

Cálculo de séries numéricas

Problema

Calcular um valor aproximado de $S = \sum_{k=1}^{\infty} a_k$ com erro absoluto inferior a ε.

Temos de determinar n tal que $|S - S_n| < \varepsilon$, e em seguida somar os primeiros n termos da série.
O problema é então majorar o erro de truncatura, ie, determinar n tal que $|R_n| < \varepsilon$.
Os critérios de convergência das séries fornecem métodos para determinar o valor de n.

- Séries de termos alternadamente positivos e negativos

  Se $|a_{k+1}| < |a_k|$ e $\lim_{k\to\infty} a_k = 0$ então a série é convergente e

  $$|R_n| < |a_{n+1}|$$

  E, portanto, basta encontrar o 1^o termo de valor absoluto $< \varepsilon$ e somar os anteriores.

Cálculo de séries numéricas

- Séries de termos positivos
  - Critério de D'Alembert

    Se $\exists n_0 \in \mathbb{N}: \forall n > n_0 \quad \frac{a_{n+1}}{a_n} \leq L$ ou $\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = L$, então se

    $$\begin{cases} L < 1 \text{ a série converge e o erro é majorado por } R_n \leq a_{n+1}\dfrac{1}{1-L} \\ L > 1 \text{ a série diverge} \\ L = 1 \text{ nada se pode concluir} \end{cases}$$

    E, portanto, no caso $L < 1$ basta encontrar n tal que $a_{n+1}\dfrac{1}{1-L} < \varepsilon$ e somar os n primeiros termos.

Cálculo de séries numéricas

- Critério de Cauchy

  Se $\exists n_0 \in \mathbb{N}: \forall n > n_0 \quad \sqrt[n]{a_n} \leq L$ ou $\lim_{n\to\infty} \sqrt[n]{a_n} = L$, então se

  $$\begin{cases} L < 1 \text{ a série converge e o erro é majorado por } R_n \leq \dfrac{L^{n+1}}{1-L} \\ L > 1 \text{ a série diverge} \\ L = 1 \text{ nada se pode concluir} \end{cases}$$

- Critério de comparação com o integral

  Se $(a_n)_n$ é uma sucessão de termos positivos não crescente, ie, $a_1 \geq a_2 \geq \ldots$ e f uma função real contínua, não crescente e positiva definida em $[1, +\infty[$ tal que $f(n) = a_n$ então:

  $$\int_1^{+\infty} f(x)dx \text{ converge} \iff \sum_{k=1}^{\infty} a_k \text{ converge}$$

  e no caso de convergir $R_n \leq \int_n^{+\infty} f(x)dx$.

Cálculo de séries numéricas

Exercício: Calcular $\cos(\pi/4)$ com erro absoluto inferior a $\varepsilon = 10^{-5}$, usando o desenvolvimento em série de Taylor de $\cos(x)$

Nota:
Quando se calcula $S_n = a_1 + a_2 + \cdots + a_n$ temos de considerar os erros de representação dos números e os erros propagados na soma. Isto é, não se calcula S_n mas sim um valor aproximado $\widehat{S_n}$, devido aos erros de arredondamento. Sendo assim, o erro total é

$$|S - \widehat{S_n}| = |S - S_n + S_n - \widehat{S_n}| \leq \underbrace{|S - S_n|}_{\text{erro de truncatura}} + \underbrace{|S_n - \widehat{S_n}|}_{\text{erro de arredondamento}}$$

e devemos analisar se o erro de arredondamento é ou não menosprezável em relação ao erro de truncatura. Analisar com cuidado os casos em que n é muito grande ou ε é muito próximo da precisão máquina e eventualmente dividir o erro pelas duas parcelas.

Erros na resolução numérica de sistemas de equações lineares

Erros na resolução numérica de sistemas de equações lineares:

- Métodos diretos: método de eliminação gaussiana.
- Efeito dos erros de arredondamento.
- Pivotagem.

Erros na resolução numérica de sistemas de equações lineares

Dado um sistema de equações

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ \vdots & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n \end{cases} \qquad (1)$$

$$(1) \iff Ax = b, \quad A \in \mathbb{R}^{n\times n}, \quad \mathrm{x} \in \mathbb{R}^{n}, \quad \mathrm{b} \in \mathbb{R}^{n}$$

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

Sistemas triangulares

o método de eliminação gaussiana transforma o sistema num sistema equivalente em que a matriz é triangular superior

$$(1) \iff \begin{cases} a_{11}^{(1)} x_1 + a_{12}^{(1)} x_2 + \ldots + a_{1n}^{(1)} x_n &= b_1^{(1)} \\ a_{22}^{(2)} x_2 + \ldots + a_{2n}^{(2)} x_n &= b_2^{(2)} \\ \vdots & \\ a_{nn}^{(n)} x_n &= b_n^{(n)} \end{cases}$$

que se resolve por substituição para trás

$$\begin{cases} x_n &= \dfrac{b_n^{(n)}}{a_{nn}^{(n)}} \\ x_i &= \dfrac{1}{a_{ii}^{(i)}} \left(b_i^{(i)} - \displaystyle\sum_{j=i+1}^{n} a_{ij}^{(i)} x_j \right), \quad i = n-1(-1)1 \end{cases}$$

$\longrightarrow \quad a_{ii}^{(i)} \neq 0, \quad i = 1(1)n \quad \ldots$ É sempre possível?

Passo k do algoritmo de eliminação gaussiana

Passo k **:** elemento pivot - a_{kk}, elementos a eliminar - $a_{k+1k}, \ldots, a_{nk}$, elementos que são alterados - a verde

$$\left(\begin{array}{ccccccc|c} a_{11} & a_{12} & \ldots & a_{1k} & a_{1k+1} & \ldots & a_{1n} & b_1 \\ & a_{22} & \ldots & a_{2k} & a_{2k+1} & \ldots & a_{2n} & b_2 \\ & & \ddots & \vdots & \vdots & & \vdots & \vdots \\ & & & a_{kk} & a_{kk+1} & \ldots & a_{kn} & b_k \\ & & & a_{k+1k} & a_{k+1k+1} & \ldots & a_{k+1n} & b_{k+1} \\ & & & \vdots & \vdots & & \vdots & \vdots \\ & & & a_{nk} & a_{nk+1} & \ldots & a_{nn} & b_n \end{array} \right)$$

- para $i = k+1$ até n faça $\{$
  - se $a_{kk} \neq 0$ então $\{$ $\longleftarrow$ a_{kk} elemento pivot
    - $m_{ik} = \dfrac{a_{ik}}{a_{kk}}$ $\longleftarrow$ multiplicador
    - para $j = k+1$ até n faça $\{a_{ij} = a_{ij} - m_{ik} a_{kj}\ \}$
    - $b_i = b_i - m_{ik} b_k\ \}$

  $\}$

Propagação dos erros de arredondamento. Exemplo

- Exemplo: Resolver num sistema VF(10, 3, -9, 9) com arredondamento

$$\begin{cases} -1.41x_1 + 2x_2 = 1 \\ x_1 - 1.41x_2 + x_3 = 1 \\ 2x_2 - 1.41x_3 = 1 \end{cases}$$

- Passo 1

$$\left(\begin{array}{ccc|c} -1.41 & 2 & 0 & 1 \\ \boxed{1} & -1.41 & 1 & 1 \\ 0 & 2 & -1.41 & 1 \end{array}\right)$$

elemento pivot $a_{11} = -1.41 \neq 0$

$\longrightarrow$ eliminar a_{21}

$\Longrightarrow m_{21} = \dfrac{a_{21}}{a_{11}} = \dfrac{1}{-1.41} = -0.709$

multiplicar a linha 1 por $-m_{21}$ e somar na linha 2:

- $a_{21}^{(1)} = 0$
- $a_{22}^{(1)} = -1.41 - (-0.709) \times 2 = 0.01$
- $a_{23}^{(1)} = 1 - (-0.709) \times 0 = 1$
- $b_2^{(1)} = 1 - (-0.709) \times 1 = 1.71$

Propagação dos erros de arredondamento. Exemplo

e obtemos o sistema equivalente

$$\left(\begin{array}{ccc|c} -1.41 & 2 & 0 & 1 \\ 0 & 0.01 & 1 & 1.71 \\ 0 & \boxed{2} & -1.41 & 1 \end{array}\right)$$

- Passo 2

elemento pivot $a_{22}^{(1)} = 0.01 \neq 0$

$\longrightarrow$ eliminar $a_{32}^{(1)}$ $\Longrightarrow$ construir $m_{32} = \dfrac{2}{0.01} = 200$

- $a_{32}^{(2)} = 0$
- $a_{33}^{(2)} = -1.41 - 200 \times 1 = -201$
- $b_3^{(2)} = 1 - 200 \times 1.71 = -341$

e obtemos

$$\left(\begin{array}{ccc|c} -1.41 & 2 & 0 & 1 \\ 0 & 0.01 & 1 & 1.71 \\ 0 & 0 & -201 & -341 \end{array}\right)$$

Propagação dos erros de arredondamento. Exemplo

que se resolve por substituição para trás

- $x_3 = \dfrac{-341}{-201} = 1.70$
- $x_2 = \dfrac{1.71 - 1 \times 1.70}{0.01} = 1.00$
- $x_1 = \dfrac{1 - (0 \times 1.70 + 2 \times 1.00)}{-1.41} = 0.709$

- Calculemos a norma do vetor resíduo $||Ax - b||_2$:

$$Ax - b = A \begin{pmatrix} 0.709 \\ 1.00 \\ 1.70 \end{pmatrix} - b = \begin{pmatrix} 1.00 \\ 0.999 \\ -0.397 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ -0.001 \\ -1.40 \end{pmatrix}$$

$$\Longrightarrow \quad ||Ax - b||_2 = 1.4$$

Como o resíduo é grande a solução calculada não é uma boa aproximação da solução exata!

Pivotagem parcial sobre linhas. Exemplo

- No passo 2 trocar a segunda linha com a terceira na matriz aumentada (isto é trocar duas equações no sistema, obtendo um sistema equivalente)

$$\left(\begin{array}{ccc|c} -1.41 & 2 & 0 & 1 \\ 0 & 0.01 & 1 & 1.71 \\ 0 & \boxed{2} & -1.41 & 1 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|c} -1.41 & 2 & 0 & 1 \\ 0 & 2 & -1.41 & 1 \\ 0 & \boxed{0.01} & 1 & 1.71 \end{array}\right)$$

repetir o processo de eliminação: $m_{32} = \dfrac{0.01}{2} = 0.005$

- $a_{32}^{(2)} = 0$
- $a_{33}^{(2)} = 1 - m_{32} \times (-1.41) = 1.01$
- $b_3^{(2)} = 1.71 - m_{32} \times 1 = 1.71$

e obtemos

$$\left(\begin{array}{ccc|c} -1.41 & 2 & 0 & 1 \\ 0 & 2 & -1.41 & 1 \\ 0 & 0 & 1.01 & 1.71 \end{array}\right) \Longrightarrow x = \begin{pmatrix} 1.69 \\ 1.69 \\ 1.69 \end{pmatrix}, \quad Ax - b = \begin{pmatrix} -0.003 \\ -0.003 \\ -0.003 \end{pmatrix}$$

$$\Longrightarrow \quad ||Ax - b||_2 = 0.005$$

Pivotagem parcial sobre linhas

O resíduo é da ordem de grandeza de 10^{-3}, o que é aceitável atendendo a que estamos a resolver o sistema usando uma aritmética de 3 algarismos significativos.

O que aconteceu?

- Na primeira resolução o multiplicador $m_{32} = 200$ é muito grande porque o elemento pivot é $pivot = 0.01$ muito pequeno. Os erros de arredondamento cometidos no passo 1 vão propagar-se ao passo 2 multiplicados por 200 (vão ser muito ampliados).

  Por outro lado ao calcular a solução fazem-se subtracções de quantidades muito próximas dando origem a erros de cancelamento.
- Na segunda resolução usamos um elemento pivot maior, $pivot = 2$, o multiplicador é muito menor $m_{32} = 0.005$, e os erros do passo 1 vão ser multiplicados por 0.005.

Conclusão:

- A introdução de pivotagem conduz a resultados muito mais precisos.
- Para conferir estabilidade numérica ao processo de eliminação gaussiana devemos fazer escolha cuidadosa do elemento pivot.

Técnicas de escolha do pivot

As técnicas de escolha do pivot mais eficientes são as que garantem multiplicadores em módulo menores que 1.

- Pivotagem parcial sobre linhas:

  no passo de ordem k escolher para pivot o elemento

  $$a_{rk} : |a_{rk}| = \max_{k \leq i \leq n} |a_{ik}|$$

  e trocar a linha k com a linha r

  Note-se que com este processo se no final da pesquisa o elemento pivot é nulo, isto significa que o sistema é singular.
- Pivotagem total:

  no passo de ordem k escolher para pivot o elemento

  $$a_{rs} : |a_{rs}| = \max_{k \leq i,j \leq n} |a_{ij}|$$

  e trocar a linha k com a linha r e a coluna k com a coluna s.

  Esta escolha não é muito usada na prática porque a vantagem do ponto de vista de estabilidade é largamente ultrapassada pelos custos de implementação. Há que registar todas as trocas sobre colunas para reordenação da solução final.