UNIVERSIDADE DO MINHO

Licenciatura em Ciências da Computação

Análise Numérica

Duração: 2 horas 30 minutos	22 de janeiro de 2019
EXAME FINAL (COM CONSULTA)	

Deves escrever na tua folha de respostas todos os comandos executados no Matlab.

1. No formato duplo da norma IEEE 754, um número x normalizado expressa-se na forma

$$x = \pm \left(1.b_1 b_2 \cdots b_{52}\right)_2 \times 2^E$$

onde $b_i = 0$ ou $b_i = 1$, para cada $i = 1, \cdots, 52$, e $-1022 \leq E \leq 1023$. Denotamos por $\mathcal{F}$ o conjunto dos números deste sistema.

a) É ou não verdade que há mais elementos de $\mathcal{F}$ no intervalo real [1,2] do que no intervalo real [2,3]? Porquê?

b) Sejam $fl(a)$ e $fl(b)$ as representações normalizadas obtidas por arredondamento dos números $a = 0.1$ e $b = 0.8$. Mostra que

$$\left|\frac{a - fl(a)}{a}\right| = \left|\frac{b - fl(b)}{b}\right|.$$

2. A função seno hiperbólico é definida por

$$senh(x) = \frac{exp(x) - exp(-x)}{2} \tag{1}$$

e também se pode expressar como soma de uma série de potências

$$senh(x) = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots + \frac{x^{2k+1}}{(2k+1)!} + \cdots \tag{2}$$

a) A partir da série anterior e sem conhecer o valor de $senh(x)$, podemos garantir que

$$\left|senh(x) - \left(x + \frac{x^3}{3!}\right)\right| < \frac{|x|^5}{5!}?$$

Porquê?

b) No código seguinte, que calcula a soma dos primeiros 10 termos da série (2) para um dado valor de x, completa as duas instruções que se encontram incompletas

```
termo=x; soma=termo;
for k=1:9
    termo=termo*.....
    soma=.....+termo;
 end
```

c) Para $x = 10^{-5}$, calcula $senh(x)$ usando a definição dada em (1) e a soma dos dois primeiros termos da série dada em (2). Apresenta os resultados obtidos em *format long*. Indica qual dos dois resultados te parece que tem menos algarismos corretos e explica qual é a causa do erro.

3. **a)** No Matlab executa

```
>> x=0.01:0.01:1; z=[0.015,0.995]; pz=interp1(x,log(x),z)
```

e interpreta os resultados obtidos.

b) No Matlab,

```
>> abs(pz(1)-log(0.015)),   abs(pz(2)-log(0.995))
```

produz os resultados 0.0589 e 1.2626e-05, respetivamente. Por que razão, o primeiro valor é muito maior do que o segundo?

4. Seja f definida por $f(x) = \frac{1}{(x+1)^2}$. Denotemos por p o polinómio de grau não superior a 2 que interpola f nos extremos e também no ponto médio do intervalo [0,0.2]. Da mesma maneira, denotemos por q, r, s e t os polinómios de grau não superior a dois que interpolam f nos extremos e também no ponto médio dos intervalos [0.2,0.4], [0.4,0.6], [0.6,0.8] e [0.8,1], respetivamente.

a) Para calcular a aproximação de $I = \int_0^1 f(x)dx$ dada por

$$\int_0^{0.2} p(x)dx + \int_{0.2}^{0.4} q(x)dx + \int_{0.4}^{0.6} r(x)dx + \int_{0.6}^{0.8} s(x)dx + \int_{0.8}^{1} t(x)dx$$

usa um dos códigos desenvolvidos nas aulas.

b) Determina um majorante do erro de truncatura cometido na aproximação.

5. **a)** No Matlab define a matriz

$$A = \begin{bmatrix} 1 & 2^{-53} & 3 \\ 4 & 2^{-40} & 1 \\ 2 & 3 & 4 \end{bmatrix}$$

e usa o código GaussElim para calcular a segunda coluna da matriz inversa de A.

b) Mostra que a solução obtida tem erro elevado e descreve detalhadamente a causa deste erro.

c) Qual a diferença entre os algoritmos implementados nos códigos GaussElim e GaussElimPP?

questão	1a	1b	2a	2b	2c	3a	3b	4a	4b	5a	5b	5c	Total
cotação	1,5	1,5	1,5	1,5	2	1,5	2	1,5	2	1,5	2	1,5	20

RESOLUÇÃO

1. a) Para cada valor do expoente E, existem 2^{52} mantissas diferentes. Para $E = 0$, os 2^{52} números diferentes estão todos no intervalo [1,2]. Para $E = 1$, os correspondentes 2^{52} números diferentes estão no intervalo [2,4] mas não estão todos no intervalo [2,3]. É pois verdade que há mais elementos de $\mathcal{F}$ no intervalo real [1,2] do que no intervalo real [2,3].

   b) Uma vez que $b = a \times 2^3$, as suas representações em $\mathcal{F}$ têm a mesma mantissa e só diferem no expoente, isto é, se

$$fl(a) = (1.b_1 b_2 \cdots b_{52})_2 \times 2^E$$

   (nota: não é necessário determinar os valores de $b_i \in \{0, 1\}$ e E.) então

$$fl(b) = (1.b_1 b_2 \cdots b_{52})_2 \times 2^{E+3}.$$

   Assim, obtemos para os erros absolutos

$$|b - fl(b)| = 2^3 \times |a - fl(a)|$$

   e

$$\frac{|b - fl(b)|}{|b|} = \frac{2^3 \times |a - fl(a)|}{2^3 \times |a|} = \frac{|a - fl(a)|}{|a|}$$

2. a) Nas séries alternadas, o erro de truncatura é, em valor absoluto, inferior ao valor do primeiro termo que se despreza. Mas a série dada não é alternada, para $x > 0$ todos os termos são positivos e para $x < 0$ todos os termos são não negativos. Em ambos os casos tem-se

$$\left| senh(x) - \left(x + \frac{x^3}{3!} \right) \right| = \frac{|x|^5}{5!} + \frac{|x|^7}{7!} + \cdots > \frac{|x|^5}{5!}.$$

   b) O código completo é

```
termo=x; soma=termo;
for k=1:9
    termo=termo*x^2/((2*k)*(2*k+1));
    soma=soma+termo;
 end
```

   c)
```
>>  x=1e-5; (exp(x)-exp(-x))/2

ans =

     1.000000000012102e-05

>> x+x^3/6

ans =
```

```
1.000000000016667e-05
```

Os valores calculados diferem nos últimos 5 algarismos. O valor calculado pela série é o mais exato que a série permite calcular uma vez que se adicionarmos mais termos o valor não se altera. A aproximação calculada a partir da expressão dada em (1) tem erro maior que é causado por cancelamento subtrativo em $exp(1e-5) - exp(-1e-5)$.

3. a)

```
>> x=0.01:0.01:1; z=[0.015,0.995]; pz=interp1(x,log(x),z)

pz =

   -4.2586   -0.0050
```

O resultado -4.2586 é o valor no ponto 0.015 do polinómio de grau 1 que interpola a função *log* nos nós 0.01 e 0.02; o resultado -0.0050 é o valor no ponto 0.995 do polinómio de grau 1 que interpola a função *log* nos nós 0.990 e 1.

b) O erro de truncatura na interpolação linear nos nós x_0 e x_1 é dado por

$$f(x) - p(x) = (x - x_0)(x - x_1)\frac{f''(\xi)}{2}.$$

Sendo $pz(1)$ o valor no ponto 0.015 do polinómio de grau 1 que interpola a função *log* nos nós 0.01 e 0.02 e $pz(2)$ o valor no ponto 0.995 do polinómio de grau 1 que interpola a função *log* nos nós 0.990 e 1, tem-se

$$log(0.015) - pz(1) = (0.015 - 0.01)(0.015 - 0.02) \times \frac{-1}{2\xi_1^2} = 2.5 \times 10^{-5} \times \frac{1}{2\xi_1^2}$$

e

$$log(0.995) - pz(2) = (0.995 - 0.99)(0.995 - 1) \times \frac{-1}{2\xi_2^2} = 2.5 \times 10^{-5} \times \frac{1}{2\xi_2^2}$$

onde ξ_1 está entre 0.01 e 0.02 e ξ_2 está entre 0.99 e 1. Assim, resulta que

$$|log(0.015) - pz(1)| \geq \frac{2.5 \times 10^{-5}}{2 \times 0.02^2} = 0.0313...$$

e

$$|log(0.995) - pz(2)| \leq \frac{2.5 \times 10^{-5}}{2 \times 0.99^2} = 1.27... \times 10^{-5}.$$

4. a) Trata-se da regra composta de Simpson com $h = 0.1$ e $n = 10$. Executando

```
S=simpson('1./(x+1).^2',0,1,10)
```

obtemos a aproximação

```
S=0.5000
```

b) Na regra composta de Simpson o erro de truncatura é majorado, em valor absoluto, por

$$\frac{h^4}{180}(b-a)f^{(iv)}(\eta)$$

onde $\eta \in [a,b]$. No caso presente, isto vale

$$\frac{10^{-4}}{180} \times \frac{120}{(\eta+1)^6}$$

onde η está entre 0 e 1, uma vez que se tem

$$f(x) = (x+1)^{-2},\ f'(x) = -2(x+1)^{-3}, \cdots, f^{(iv)}(x) = 120(x+1)^{-6}.$$

A derivada de quarta ordem atinge o valor máximo de 120 para $x = 0$ e um majorante para o erro de truncatura é dado por

```
>> (1e-4/180)*120

ans =

   6.6667e-0
```

5. a) A segunda coluna da inversa de A é a solução do sistema $Ax = b$ onde b é segunda coluna da matriz identidade.Com

```
>> A=[1, 2^-(53), 3; 4 , 2^-(40), 1; 2, 3, 4]; b=[0, 1, 0]';
>>  x=GaussElim(A,b)
```

obtem-se

```
x =

    0.2727
   -0.0606
   -0.0909
```

b) Se executarmos

```
>>  x1=GaussElimPP(A,b)
```

obtemos

```
x1 =

    0.2727
   -0.0606
   -0.0909
```

Embora os vetores x e $x1$ coincidam em "format short", a diferença é significativa como se pode ver de

```
>> norm(x-x1)

ans =

   2.9607e-05
```

No final do primeiro passo de redução tem-se a matriz ampliada

```
ans =

   1.0000e+00   1.1102e-16   3.0000e+00            0
            0   9.0905e-13  -1.1000e+01   1.0000e+00
            0   3.0000e+00  -2.0000e+00            0
```

e o elemento na posição (2,2) que vai ser usado como pivot é muito mais pequeno do que o elemento na posição (3,2) a eliminar. Isto produz um multiplicador muito grande que vai originar grandes erros de arredondamento.

c) O código GaussElimPP implementa o método de pivotação parcial que consiste em escolher para pivot, no início do k-ésimo passo de redução, para cada k=1,...,n-1, de entre os elementos A(k,k), A(k+1,k),...,A(n,k) aquele que tiver o maior valor absoluto. Se este elemento for A(p,k) onde $p > k$, então serão permutadas as linhas k e p da matriz ampliada. Com esta operação de equivalência garante-se que os multiplicadores terão valor absoluto não superior á unidade o que evita os erros de arredondamento mencionados antes.