UNIVERSIDADE DO MINHO

Licenciatura em Ciências da Computação

Análise Numérica

Duração: 2 horas — 30 de outubro de 2017

TESTE 1 (COM CONSULTA)

1. **a)** Dado um número na sua representação binária

$$b_k b_{k-1} \cdots b_0 \cdot b_{-1} b_{-2} \cdots b_{-m}$$

a função **horner**, desenvolvida nas aulas, pode ser usada no Matlab para obter a representação decimal do mesmo número. Explica como e justifica o procedimento.

b) Determina a representação decimal do número $(11.1011001)_2$. Na tua folha de respostas escreve o(s) comando(s) executados no Matlab e o resultado obtido.

2. 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) O número 0.1 (base 10) não pertence a $\mathcal{F}$ e é representado por outro número, que denotamos por $fl(0.1)$ e que é obtido por arredondamento. Determina um majorante para o erro $|0.1 - fl(0.1)|$ assumindo o modo de arredondamento "para o mais próximo".

b) Uma vez que, como se disse antes, 0.1 (base 10) não pertence a $\mathcal{F}$, podemos concluir que também 0.4 não pertence a $\mathcal{F}$. Porquê?

c) Os erros absolutos $|0.1 - fl(0.1)|$ e $|0.4 - fl(0.4)|$ são iguais? E os erros relativos $\frac{|0.1-fl(0.1)|}{0.1}$ e $\frac{|0.4-fl(0.4)|}{0.4}$? Justifica ambas as respostas.

3. Representemos por $atan(x)$ o valor do ângulo, em radianos, entre $-\frac{\pi}{2}$ e $\frac{\pi}{2}$, cuja tangente é x. Para valores $x \in]-1, 1[$, tem-se

$$atan(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots + (-1)^{k+1} \frac{x^{2k-1}}{2k-1} + \cdots$$

a) Se quisermos usar a série de potências anterior para aproximar o valor de $atan(0.5)$ com um erro de truncatura que é garantidamente menor do que 0.001, qual é o último termo que teremos que adicionar? Justifica a tua resposta.

b) No Matlab, soma os termos necessários para calcular $atan(0.5)$ com erro de truncatura inferior a 0.001. Na tua folha de respostas escreve o(s) comando(s) executado(s) e o resultado obtido.

4. Seja

$$X = \sqrt{1 + 10^{-14}}.$$

Para calcular $X - 1$, uma fórmula alternativa é $\frac{10^{-14}}{1+X}$ uma vez que $X - 1 = \frac{X^2-1}{X+1}$. No Matlab executa

```
>> X=sqrt(1+1e-14); X-1, 1e-14/(1+X)
```

escreve na tua folha de respostas os resultados obtidos e diz, justificando, qual dos dois resultados tem mais algarismos significativos corretos.

5. **a)** A partir de um intervalo inicial $[a_0, b_0]$, de amplitude igual a dois (isto é, $b_0 - a_0 = 2$), que contem uma raiz de uma dada equação, quantos passos k de bisseção são precisos para determinar um intervalo $[a_k, b_k]$ tal que

$$b_k - a_k < 10^{-14}?$$

Justifica a tua resposta.

b) Se for $[a_0, b_0] = [90, 92]$, será possível usar a função bisec, desenvolvida nas aulas, para produzir no Matlab um intervalo $[a_k, b_k]$ tal que $b_k - a_k < 10^{-14}$? Porquê?

6. Sendo a um número positivo, a raiz quadrada $\sqrt{a}$ é a raiz positiva da equação

$$x = (x + a/x)/2.$$

Daqui resulta a fórmula iterativa

$$x_{k+1} = (x_k + a/x_k)/2.$$

a) Tendo em conta que se $x_k > \sqrt{a}$ a fórmula anterior produz $x_{k+1} > \sqrt{a}$, mostra que a sucessão de aproximações gerada a partir de uma aproximação inicial $x_0 > \sqrt{a}$ converge para $\sqrt{a}$.

b) Usa a fórmula iterativa dada para, a partir de $x_0 = 1.5$, calculares $\sqrt{2}$ o mais exatamente possível. Deves escrever na tua folha de respostas todas as iterações produzidas no Matlab em *format long*.

questão	1a	1b	2a	2b	2c	3a	3b	4	5a	5b	6a	6b	Total
cotação	1.5	1.5	1.5	1.5	1.5	2	1.5	2	1.5	2	2	1.5	20

RESOLUÇÃO

1. **a)** Dado um polinómio de grau n na forma do vetor $[a(1), a(2), \cdots, a(n+1)]$ dos coeficientes (por ordem decrescente dos graus dos respetivos termos), usada na forma **px=horner(a,x)**, esta função calcula o valor do polinómio no ponto x, dado. Uma vez que o número representado em binário por

$$b_k b_{k-1} \cdots b_0 \cdot b_{-1} b_{-2} \cdots b_{-m}$$

é

$$b_k \times 2^k + b_{k-1} \times 2^{k-1} + \cdots + b_0 + b_{-1} \times 2^{-1} + b_{-2} \times 2^{-2} + \cdots + b_{-m} \times 2^{-m},$$

a parte inteira deste número é o valor do polinómio $[b(k), b(k-1), \cdots, b(0)]$ no ponto $x = 2$ e a parte decimal, que pode ser escrita na forma

$$b_{-m} \times (\frac{1}{2})^m + \cdots + b_{-2} \times (\frac{1}{2})^2 + b_{-1} \times (\frac{1}{2}) + 0$$

é o valor do polinómio $[b(-m), \cdots b(-2), b(-1), 0]$ no ponto $x = \frac{1}{2}$.

b) No Matlab,

```
horner([1,1],2)+horner([1,0,0,1,1,0,1,0],1/2)
```

dá o número 3.6953

2. **a)** Uma vez que

$$0.1 = 2^{-4} + 2^{-5} + \cdots,$$

o expoente (na representação normalizada) é $E = -4$. A distância entre dois números consecutivos de $\mathcal{F}$ com expoente $E = -4$ é 2^{-56} (tendo em conta que um bit igual a 1 na última posição da mantissa vale 2^{-52}). Com arredondamento "para o mais próximo" podemos concluir que

$$|0.1 - fl(0.1)| \leq 2^{-57}.$$

b) Uma vez que

$$0.1 = \left(1 + \sum_{j=1}^{\infty} b_j \times 2^{-j}\right) \times 2^{-4},$$

de $0.4 = 0.1 \times 2^2$ resulta

$$0.4 = \left(1 + \sum_{j=1}^{\infty} b_j \times 2^{-j}\right) \times 2^{-2},$$

para os mesmos bits $b_j,\ j = 0, \cdots, \infty$.

c) Tem-se

$$|0.1 - fl(0.1)| = \left| \left(1 + \sum_{j=1}^{\infty} b_j \times 2^{-j}\right) - fl\left(1 + \sum_{j=1}^{\infty} b_j \times 2^{-j}\right) \right| \times 2^{-4}$$

e

$$|0.4 - fl(0.4)| = \left| \left(1 + \sum_{j=1}^{\infty} b_j \times 2^{-j}\right) - fl\left(1 + \sum_{j=1}^{\infty} b_j \times 2^{-j}\right) \right| \times 2^{-2}$$

donde se conclui que

$$|0.4 - fl(0.4)| = 4 \times |0.1 - fl(0.1)|.$$

e os erros relativos são iguais.

3. **a)** Numa série alternada o erro de truncatura é inferior ao valor absoluto do primeiro termo que se despreza. Uma vez que

$$\frac{0.5^7}{7} = 0.0011$$

e

$$\frac{0.5^9}{9} = 2.1701e-04,$$

concluímos que teremos que adicionar ainda o termo $-\frac{0.5^7}{7}$ para garantir um erro de truncatura inferior a 0.001.

b) Executando no Matlab

```
>> x=0.5; x-x^3/3+x^5/5-x^7/7
```

obtemos o resultado 0.4635.

4. Os resultados obtidos são 4.8850e-15 e 5.0000e-15. O segundo resultado tem mais algarismos corretos do que o primeiro porque no cálculo de $X - 1$ ocorre cancelamento subtrativo uma vez que $X = 1.000000000000005$ coincide com 1 nos primeiros quinze algarismos significativos.

5. **a)** Trata-se do menor valor de k que satisfaz

$$\frac{2}{2^k} < 10^{-14}$$

que é $k = 48$.

b) Não é possível determinar a_k e b_k no intervalo $[90, 92]$ tais que $b_k - a_k < 10^{-14}$ porque neste intervalo a distância entre um número de $\mathcal{F}$ e o seu sucessor é $2^{-46} > 10^{-14}$.

6. **a)** A função iteradora é

$$\varphi(x) = \frac{1}{2}\left(x + \frac{a}{x}\right).$$

Tem-se

$$\varphi'(x) = \frac{1}{2}\left(1 - \frac{a}{x^2}\right)$$

e conclui-se que $|\varphi'(x)| < 1$ para todo $x > \sqrt{a}$ e o método converge para o ponto fixo de φ que é igual a $\sqrt{a}$.

b) No Matlab, começando com

```
>> a=2; x=1.5; format long
```

e executando sucessivamente

```
>> x=(x+a/x)/2
```

obtemos as aproximações

```
1.416666666666667
1.414215686274510
1.414213562374690
1.414213562373095
1.414213562373095
```