UNIVERSIDADE DO MINHO

Licenciatura em Ciências da Computação

Análise Numérica

Duração: 2 horas e 30 minutos — 20 de novembro de 2021

TESTE 1 (COM CONSULTA)

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) Determina $b_i, i = 1, \cdots, 52$ e o expoente E para $x = 100$.

b) Mostra que o número

$$y = 100 + 2^{-46} + 2^{-47} + 2^{-48}$$

não pertence a $\mathcal{F}$.

c) Determina $fl(y)$, assumindo o modo do arredondamento para o mais próximo.

2. **a)** No Matlab, o comando

```
>> A=2^512; B=2^510; A^2-B^2
```

produz Inf. Porquê?

b) Calcula no Matlab o valor de $A^2 - B^2$ usando uma expressão alternativa. Na folha de respostas escreve o comando executado e o resultado obtido.

3. O desenvolvimento da função $log(1+x)$ em série de potências de x é

$$log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + ... + (-1)^{k+1}\frac{x^k}{k} + ...$$

a) Soma os primeiros 4 termos desta série para aproximar o valor de $log(1.001)$. Na folha de respostas escreve o(s) comando(s) executado(s) no Matlab e o resultado obtido em format long.

b) Determina um majorante para o erro de truncatura cometido na aproximação anterior. Justifica a tua resposta.

c) Achas que ocorreu cancelamento subtrativo na soma efetuada na alínea a)? Justifica a tua resposta.

4. No Matlab, o comando

```
>> z=1.001; ztil=z+4*1e-8; abs((z-ztil)/z), abs((log(z)-log(ztil))/log(z))
```

dá os resultados

```
ans =
3.996003993900677e-08
ans =
3.998001583172340e-05
```

Interpreta estes resultados e explica por que é que o segundo resultado é cerca de mil vezes maior do que o primeiro.

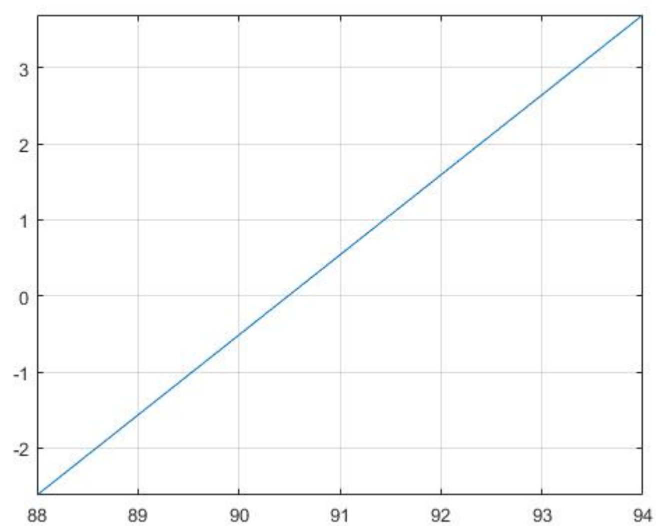
5. Na figura em baixo apresenta-se o gráfico da função definida por $f(x) = x + \sqrt{x} - 100$ num certo intervalo.

   a) Modifica ligeiramente o código da função bisec disponível na Blackboard para calculares um intervalo [a,b] de amplitude não superior a 10^{-7} que contem a raiz da equação $f(x) = 0$. Na folha de respostas explica em que consistiu a modificação, apresenta o comado executado no Matlab e os valores de a e b.

   b) A equação pode ser escrita na forma $x = 100 - \sqrt{x}$. A fórmula iterativa

$$x^{(k+1)} = 100 - \sqrt{x^{(k)}}$$

   pode ou não ser usada para calcular a raiz da equação? Justifica a tua resposta sem efetuares iterações.



6. **a)** Partindo da aproximação inicial $x^{(0)} = 0.9$, usa o método de Newton para calcular um zero do polinómio $p(x) = x^3 - 2.1x^2 + 1.2x - 0.1$. Termina as iterações quando duas aproximações sucessivas coincidirem no formato "short". Na folha de respostas apresenta as aproximações obtidas em todas as iterações efetuadas.

   b) Parece-te que o método exibiu convergência quadrática? Justifica a tua resposta.

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

RESOLUÇÃO

1. **a)** Uma vez que

$$100 = 64 + 32 + 4 = 2^6 + 2^5 + 2^2$$

tem-se

$$100 = (1.10010\cdots 0)_2 \times 2^6$$

isto é, para além do bit implícito, é

$$b_1 = b_4 = 1$$

e os restantes bits são iguais a 0. O expoente é $E = 6$. [nota: a função intdectobin desenvolvida nas aulas também podia ter sido usada para determinar os bits].

b) O sucessor de 100 em $\mathcal{F}$ é o número

$$100 + 2^{6-52} = 100 + 2^{-46}$$

e o sucessor deste é

$$100 + 2^{-46} + 2^{-46}.$$

Tem-se

$$100 + 2^{-46} < y = 100 + 2^{-46} + 2^{-47} + 2^{-48} < 100 + 2^{-46} + 2^{-46},$$

isto é, y está entre $100 + 2^{-46}$ e o respetivo sucessor. Assim se conclui que y não pertence a $\mathcal{F}$.

c) O ponto médio do intervalo

$$[100 + 2^{-46}, 100 + 2^{-46} + 2^{-46}]$$

é $100 + 2^{-46} + 2^{-47}$. Portanto, y é maior do que este ponto médio e

$$fl(y) = 100 + 2^{-46} + 2^{-46} = 100 + 2^{-45}.$$

2. **a)** Porque $A^2 = 2^{1024}$ e no formato duplo da norma IEEE o maior expoente que se pode representar é 1023. Assim, o cálculo de A^2 produz "overflow".

b) Uma vez que

$$A^2 - B^2 = (A + B) * (A - B),$$

no Matlab tem-se

```
>> A=2^512; B=2^510; (A+B)*(A-B)

ans =

    1.685337313933421e+308
```

3. **a)**

```
>> x=0.001; x-x^2/2+x^3/3-x^4/4

ans =

     9.995003330833332e-04
```

b) Uma vez que a série é alternada, o erro de truncatura é inferior ao valor absoluto do primeiro termo desprezado, $0.001^5/5 = 2 \times 10^{-16}$.

c) O cancelamento subtrativo ocorre quando a soma é de grandeza inferior à das parcelas somadas. Tal não acontece no caso presente porque a soma 9.99...e-04 é da ordem de grandeza da primeira parcela $x = 0.001$ e todas as outras parcelas têm grandeza inferior.

4. Um erro relativo aproximadamente igual a 4×10^{-8} no valor de $z = 1.001$ produz um erro relativo aproximadamente igual a 4×10^{-5} no valor de $log(z)$. Isto acontece porque o número de condição relativo da função $f(z) = log(z)$ é

$$\frac{z.f'(z)}{f(z)} = \frac{z \times \frac{1}{z}}{log(z)} = \frac{1}{log(z)}$$

e

```
>> z=1.001; 1/log(z)

ans =

     1.000499916708417e+03
```

O erro relativo em $log(z)$ é aproximadamente igual ao erro relativo em z multiplicado pelo número de condição calculado.

5. **a)** A função bisec, que implementa o método da bisseção, calcula um intervalo [a,b] que cumpre o critério de paragem e faz sair o valor médio raiz=(a+b)/2 mas não os valores de a e b. Podemos, na lista dos parâmetros de saída, incluir os valores de a e b isto é, o "header"da função

```
function [raiz, funevals] = bisec(f, a, b, tol)
```

é substituído por

```
function [a, b, raiz, funevals] = bisec(f, a, b, tol)
```

Com esta modificação e tendo em conta que a figura mostra que existe uma raiz da equação no intervalo [90,91]

```
>> [a,b] = bisec(@(x)x+sqrt(x)-100,90,91,1e-7)
```

dá

```
a =

   90.487507760524750


b =

   90.487507820129395
```

b) Uma vez que com $\phi(x) = 100 - \sqrt{x}$ é $\phi'(x) = \frac{1}{2\sqrt{x}}$, a condição $|\phi'(x)| < 1$ é satisfeita para $x > 1/4$ e a fórmula iterativa será convergente mesmo que a aproximação inicial não esteja muito próxima da raiz.

6. **a)** com

```
>> format short
>> p=[1, -2.1, 1.2, -0.1]
>> derp=[3, -4.2, 1.2]
>> x=0.9;
```

se repetirmos o comando

```
>> x=x-polyval(p,x)/polyval(derp,x)
```

obtemos as aproximações

```
0.9533
0.9773
0.9888
0.9944
0.9972
0.9986
0.9993
0.9997
0.9998
0.9999
1.0000
1.0000
```

b) A convergência não é quadrática. Isto acontece porque $r = 1$ é raiz dupla da equação $p(x) = 0$.