UNIVERSIDADE DO MINHO

Licenciatura em Ciências da Computação

Análise Numérica

Duração: 2 horas (+30 minutos de tolerância) 19 de novembro de 2022
teste 1 (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) Diz, justificando, se concordas com a afirmação seguinte "*se* $x \in \mathcal{F}$ *e* $y \in \mathcal{F}$, *então* $x + y \in \mathcal{F}$".

b) Mostra que $1/3 \notin \mathcal{F}$.

c) Assumindo o modo de arredondamento para o mais próximo, determina majorantes para os erros

$$|1/3 - fl(1/3)|$$

e

$$\frac{|1/3 - fl(1/3)|}{1/3},$$

onde $fl(1/3)$ representa o valor arredondado de 1/3.

2. Considera o seguinte desenvolvimento em série de potências de x

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

que é válido para $-1 < x \leq 1$.

a) Para calcular uma aproximação de $log(1.1)$ com erro absoluto garantidamente inferior a 10^{-10}, qual é o último termo que deve ser somado? Justifica a tua resposta.

b) Calcula a soma até ao termo que indicaste na alínea anterior. Na tua folha de respostas apresenta os comandos executados no Matlab e o resultado obtido em format long.

c) No Matlab executa

```
>> log(1.1)
```

e compara o valor obtido como que calculaste na alínea b). A diferença entre os dois valores é devida a cancelamento subtrativo no cálculo efetuado na alínea b) ou é outra a razão? Explica detalhadamente.

3. **a)** No Matlab,

```
>> format long, raiz= bisec (@(x) 1/(1-x),0,2,1e-10)
```

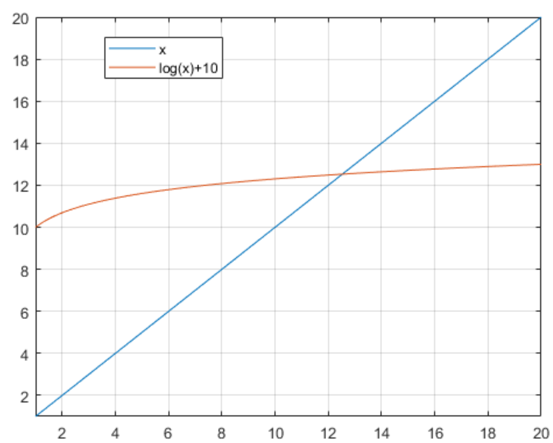
produz o resultado $raiz = 1.000000000029104$. Este valor aproxima um zero da função $f(x) = 1/(1-x)$? Explica detalhadamente.

b) Se tentares executar

```
>> raiz= bisec (@(x) 1/(1-x),0,2,1e-16)
```

o Matlab não termina a execução e terás de interromper o processo fazendo Ctrl-C. Analisa o critério de paragem implementado no código da função bisec e explica porque é que a execução daquele comando não termina.

4. **a)** Usa o método do ponto fixo para calculares o mais exatamente possível (no format long) a abcissa do ponto em que a curva de equação $y = log(x) + 10$ interseta a reta $y = x$ (ver figura em baixo). Na folha de respostas escreve os valores da aproximação inicial $x^{(0)}$ por ti escolhida e os valores produzidos na primeira e última iteração.



b) Mostra que a sequência produzida com $x^{(k+1)} = log(x^{(k)}) + 10$ converge para a raiz da equação $x = log(x) + 10$ qualquer que seja a aproximação inicial $x^{(0)} > 1$.

5. A sucessão de termo geral

$$\left(1 + \frac{1}{n}\right)^n$$

converge lentamente para o número de Neper $e = 2.71828...$ e é necessário usar valores de n muito grandes para obter boas aproximações daquele número. No entanto, há que ter cuidado na escolha do valor de n porque um pequeno erro relativo na representação da base $\left(1 + \frac{1}{n}\right)$ produzirá uma aproximação com erro muito grande. Por que é que tal acontece?

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

RESOLUÇÃO

1. **a)** A afirmação é falsa. Por exemplo, 1 e 2^{-53} pertencem ambos a $\mathcal{F}$ mas $1+2^{-53}$ não pertence porque o sucessor de 1 em $\mathcal{F}$ é $1+2^{-52}$.

   b) Vamos usar o algoritmo das multiplicações sucessivas para determinar a representação binária do número 1/3. De
   $$2\times 1/3 = 2/3 < 1$$
   e
   $$2\times 2/3 = 4/3 = 1+1/3$$
   concluímos que
   $$1/3 = 0\times 2^{-1} + 1\times 2^{-2} + 0\times 2^{-3} + 1\times 2^{-4} + 0\times 2^{-5} + \cdots$$
   Como não é finita a representação, então $1/3 \notin \mathcal{F}$.

   c) No intervalo [1/4, 1/2], a distância entre um número de $\mathcal{F}$ e o seu sucessor é 2^{-54}. No modo de arredondamento para o mais próximo, tem-se, para qualquer número de $\mathcal{F}$ entre 1/4 e 1/2
   $$|1/3 - fl(1/3)| \leq 2^{-55}.$$
   Para qualquer número x normalizado tem-se, no modo de arredondamento para o mais próximo,
   $$\frac{|x-fl(x)|}{|x|} \leq 2^{-53}.$$

2. **a)** Uma vez que a série é alternada, o erro de truncatura é, em valor absoluto, inferior ao valor absoluto do primeiro termo que se despreza. Uma vez que $0.1^9/9 > 10^{-10}$ e $0.1^{10}/10 < 10^{-10}$, para garantir uma aproximação com erro inferior a 10^{-10} deve adicionar-se ainda o termo $0.1^9/9$.

   b) No Matlab,
   ```
   >> format long;x=0.1;soma=0; for k=1:9, soma=soma+(-1)^(k+1)*x^k/k;end,soma
   ```
   produz o resultado soma=0.095310179813492.

   c) No Matlab,
   ```
   >> soma-log(1.1)

   ans =

        9.167139269905533e-12
   ```
   e a diferença é inferior à tolerância fixada 10^{-10}. Não há cancelamento subtrativo uma vez que o resultado $soma = 0.095...$ é da mesma ordem de grandeza do termo de maior ordem de grandeza, neste caso $x = 0.1$.

3. **a)** Obviamente, f definida por $f(x) = 1/(1-x)$ não tem zeros. O resultado produzido é o ponto médio de um intervalo [a,b] tal que $b-a < 10^{-10}$ e $f(a)*f(b) < 0$. A função não é contínua num intervalo que contenha o ponto $x = 1$ porque não está definida neste ponto. É verdade que $f(x) < 0$ para valores próximos de 1, à direita, e $f(x) > 0$ para valores próximos de 1, à esquerda mas, por não ser f contínua, tal não implica a existência de um zero entre a e b.

b) A função bisec não termina as iterações enquanto $b - a > tol$. O intervalo de menor amplitude tal que a e b pertencem a $\mathcal{F}$ e $f(a) \times f(b) < 0$ é $[a, b] = [1, 1 + eps]$. Como $eps = 2^{-52} \approx 2.2 \times 1e - 16$, é sempre $b - a > 1e - 16$.

4. **a)** Vamos tomar como aproximação inicial $x^{(0)} = 12$ e terminaremos quando duas aproximações sucessivas coincidirem em todos os algarismos no format long.

```
>> x=12; k=0;
>> x=log(x)+10, k=k+1
x =12.484906649788000
k = 1
....
x = 12.527963201982175
k =14
```

b) A função iteradora é $\phi(x) = log(x) + 10$. A condição $|\phi'(x)| < 1$, ou seja $|\frac{1}{x}| < 1$, é verdadeira para $x > 1$ e, portanto, qualquer que sejal $x^{(0)} > 1$, a sequência produzida com $x^{(k+1)} = log(x^{(k)}) + 10$ converge para o ponto fixo de $\phi(x) = log(x) + 10$ que é a solução de $x = log(x) + 10$.

5. O número de condição relativo de uma função f num ponto x é dado por $x.f'(x)/f(x)$. No caso de ser $f(x) = x^n$, o número de condição é n, isto é o erro relativo em f(x) pode ser n vezes maior do que o erro relativo em x. Portanto, para usar a expressão $\left(1 + \frac{1}{n}\right)^n$ deve garantir-se que a base $1 + \frac{1}{n}$ é representada exatamente (sem erro de arredondamento).