UNIVERSIDADE DO MINHO

Licenciatura em Ciências da Computação

Análise Numérica

Duração: 2 horas e 30 minutos — 21 de novembro de 2020

TESTE 1 (COM CONSULTA)

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

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

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 $2^{-3} + 2^{-55}$ pertence a $\mathcal{F}$? Justifica a tua resposta.

b) Mostra que $x = 0.8$ não pertence a $\mathcal{F}$.

c) Denotando por $fl(0.8)$ o valor arredondado de 0.8, determina, justificando, o maior valor de k (inteiro) tal que

$$|0.8 - fl(0.8)| \leq 2^{-k}$$

(assume o modo do arredondamento para o mais próximo).

2. Para ilustrar que em aritmética de ponto flutuante a adição não é associativa, determina m inteiro tal que no Matlab a expressão

$$(4 + 2^{-m}) + 2^{-m} == 4 + (2^{-m} + 2^{-m})$$

produz o valor lógico 0.

3. O desenvolvimento da função *sin* em série de potências de x é

$$sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + ...$$

a) A função *senoTaylor* que se apresenta em baixo é para calcular, dado x, a soma dos termos que, em valor absoluto, são superiores a uma dada tolerância *tol*. Na antepenúltima linha do código Matlab, a instrução 'termo=' está incompleta. Na tua folha de respostas deves escrever a instrução completa.

```
function [soma, n]=senoTaylor(x, tol)
termo=x;
soma=0;
n=1;
while abs(termo)> tol
    soma = soma + termo;
    n=n+2;
    termo =
end
n=n-2;
```

b) Para $x = \pi - 1e-5$, podemos garantir que a soma dos termos até ao termo $\frac{x^{29}}{29!}$ (inclusive) aproxima o valor de $sin(x)$ com erro de truncatura inferior a 10^{-20}? Porquê?

c) Como sabes, é $sin(\pi - x) = sin(x)$ mas no Matlab, com $x = 10^{-7}$,

```
>> x=1e-7; [soma,n]=senoTaylor(pi-x,1e-22)
```

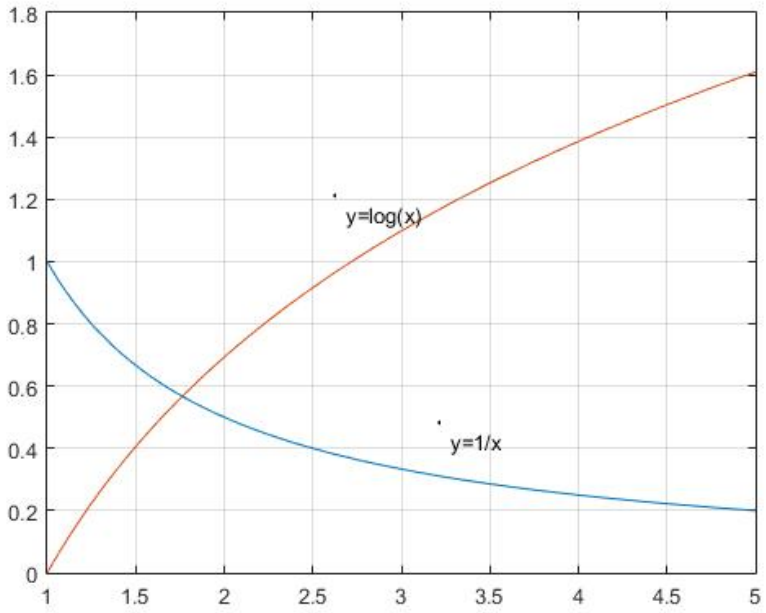
dá soma = 9.999999997787699e-08 e

```
>> x=1e-7; [soma,n]=senoTaylor(x,1e-22)
```

dá soma = 9.999999999999982e-08. Um destes resultados tem um erro relativo maior. Diz qual deles é e explica a causa do erro.

4. Diz, justificando detalhadamente, se concordas com a seguinte afirmação: *para valores de x muito próximos de 1, erros relativos em x produzem erros relativos bastante maiores em $f(x) = log(x)$.*

5. Para aproximar a abcissa do ponto em que se intersetam as curvas $y = log(x)$ e $y = 1/x$ (ver figura em baixo), escolhe uma função ϕ cujo ponto fixo seja o valor procurado e, a partir de uma aproximação inicial, no Matlab itera até que duas aproximações consecutivas coincidam nos primeiros cinco algarimos. Na tua folha de respostas deves escrever todas as iteradas obtidas no Matlab (em format short).



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

RESOLUÇÃO

1. **a)** Sim, o número tem a representação

$$(1.00\cdots01)_2 \times 2^{-3}$$

(todos os bits não explicitados são iguais a 0), logo pertence a $\mathcal{F}$.

b) Para determinar os bits $b_{-1}, b_{-2}, ...$ em

$$0.8 = b_{-1} \times 2^{-1} + b_{-2} \times 2^{-2} + b_{-3} \times 2^{-3} + b_{-4} \times 2^{-4} + ...$$

multiplicamos sucessivamente por 2 e extraímos do resultado a parte inteira (0 ou 1). De

$$1.6 = b_{-1} + b_{-2} \times 2^{-1} + b_{-3} \times 2^{-2} + b_{-4} \times 2^{-3} + ...$$

conclui-se que $b_{-1} = 1$. De novo multiplicando por 2 ambos os membros de

$$0.6 = b_{-2} \times 2^{-1} + b_{-3} \times 2^{-2} + b_{-4} \times 2^{-3} + ...$$

resulta

$$1.2 = b_{-2} + b_{-3} \times 2^{-1} + b_{-4} \times 2^{-2} + ...$$

e $b_{-2} = 1$. De

$$0.2 = b_{-3} \times 2^{-1} + b_{-4} \times 2^{-2} + ...$$

obtem-se

$$0.4 = b_{-3} + b_{-4} \times 2^{-1} + ...$$

e $b_{-3} = 0$. Agora, tem-se

$$0.8 = b_{-4} + ...,$$

donde se conclui que $b_{-4} = 0$ e também que 0.8 não tem representação finita em binário uma vez que é

$$(0.8)_{10} = (0.110011001100...)_2$$

c)

$$0.8 = (1.1001100...1100|1100...)_2 \times 2^{-1}$$

ou seja 0.8 está entre $x_- = (1.1001100...1100)_2 \times 2^{-1}$ e $x_+ = (1.1001100...1101)_2 \times 2^{-1}$ que distam entre si 2^{-53}. Se o arredondamento para o mais próximo for usado tem-se

$$|0.8 - fl(0.8)| \leq 2^{-54}.$$

2. Uma vez que se tem

```
>> 4+2^-50>4

ans =

     1

>> 4+2^-51>4

ans =

     0
```

conclui-se que é $m = 51$. Com efeito, no Matlab $(4 + 2^{-51}) + 2^{-51}$ produz o resultado 4 e $4 + (2^{-51} + 2^{-51})$ produz o resultado $4 + 2^{-50}$.

3. **a)** O código completo é o seguinte

```
function [soma, n]=senoTaylor(x, tol)

% calcula a soma dos termos da série de potências de x para a função
% seno até encontrar um termo cujo valor absoluto
% é inferior a uma tolerância tol.
% n é o grau do último termo somado.

termo=x;
soma=0;
n=1;
while abs(termo)> tol
    soma = soma + termo;
    n=n+2;
    termo = -termo*x^2/(n*(n-1));
end
n=n-2;
```

b) Numa série alternada, o erro de truncatura é inferior ao valor absoluto do primeiro termo que se despreza. Neste caso, será o termo

```
>> x=pi-1e-5; x^31/factorial(31)

ans =

   3.1375e-19
```

pelo que não se pode garantir um erro de truncatura inferior a 10^{-20}.

c) O primeiro resultado tem um erro relativo maior porque há perda de algarismos corretos devido ao cancelamento subtrativo. Isto acontece porque a soma é da ordem de grandeza de 10^{-7} e há termos cujo valor absoluto é da ordem de grandeza da unidade, por exemplo

```
 (pi-x)^3/factorial(3)

ans =

   5.167712286569766
```

O cancelamento subtrativo não está presente quando se usa $x = 1e{-}7$ como argumento da função senoTaylor porque agora os termos são de grandeza muito inferior. Com efeito, o maior termo é o primeiro $x = 1e - 7$, muito próximo do valor da soma calculada.

4. Tem-se (ver p. 35 das notas das aulas)

$$\left|\frac{f(x+\Delta x) - f(x)}{f(x)}\right| \approx \left|\frac{x.f'(x)}{f(x)}\right| \left|\frac{\Delta x}{x}\right|$$

o que mostra que o erro relativo no valor calculado da função é aproximadamente igual ao erro relativo no argumento da função multiplicado por $\left|\frac{x.f'(x)}{f(x)}\right|$ que é o número de condição relativo de f no ponto x. No caso presente, tem-se

$$\left|\frac{x.f'(x)}{f(x)}\right| = \left|\frac{x.\frac{1}{x}}{log(x)}\right| = \left|\frac{1}{log(x)}\right|$$

que assume valores tanto maiores quanto mais próximo de 1 estiver x. É portanto verdadeira a afirmação.

5. O que se quer é escrever na forma $x = \phi(x)$ a equação $log(x) = 1/x$ ou $f(x) = 0$ com $f(x) = log(x) - 1/x$.

   Não serve qualquer escolha de ϕ obter uma sucessão convergente. Uma boa escolha (assumindo que a raiz procurada é simples) é

$$\phi(x) = x - \frac{f(x)}{f'x)} = x - \frac{log(x) - 1/x}{1/x - 1/x^2}$$

```
>> fi=@(x) x-((log(x)-1/x)/(1/x+1/x^2))
fi =

    @(x)x-((log(x)-1/x)/(1/x+1/x^2))

>> x=1.5
x =

    1.5000

>> x=fi(x)

x =

    1.7351

>> x=fi(x)

x =

    1.7629

>> x=fi(x)
x =

    1.7632

>> x=fi(x)
x =

    1.7632
```