UNIVERSIDADE DO MINHO

Licenciatura em Ciências da Computação

Análise Numérica

Duração: 2 horas e 30 minutos	10 de dezembro 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) Seja x um número tal que

$$x_- < x < x_+$$

onde $x_- = 1000$ e x_+ é o sucessor de 1000 em $\mathcal{F}$. Determina, justificando, um majorante para o erro

$$\frac{|x - fl(x)|}{|x|},$$

assumindo o modo de arredondamento para o mais próximo.

b) Todos os números inteiros positivos menores do que realmax pertencem a $\mathcal{F}$? Justifica a tua resposta.

2. Para a função exponencial tem-se

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ... + \frac{x^k}{k!} + \frac{x^{k+1}}{(k+1)!} + ...$$

a) Com $x = -0.1$, até que termo será necessário somar para garantir um erro de truncatura inferior a 10^{-16}? Justifica a tua resposta.

b) Usando a expressão do resto

$$R_k(x) = \frac{e^\theta}{(k+1)!} x^{k+1},$$

onde θ é um ponto que está entre 0 e x, mostra que para $x > 0$ tem-se

$$\frac{R_k(x)}{e^x} < \frac{x^{k+1}}{(k+1)!}$$

c) Para o valor *soma* calculado com função *expTaylor* (desenvolvida nas aulas PL) com $x = 0.1$ e $tol = 10^{-10}$, indica, justificando, um majorante para o erro relativo

$$\frac{e^x - soma}{e^x}$$

3. Sejam a e δ números tais que $a = 1 - \sqrt{1-\delta}$.

   a) Supõe que δ não é conhecido exatamente e calculamos $\widetilde{a} = 1 - \sqrt{1-\widetilde{\delta}}$, onde $\widetilde{\delta}$ aproxima δ com um pequeno erro relativo. Em que caso $\widetilde{a}$ terá um erro relativo bastante maior do que $\widetilde{\delta}$, quando $\widetilde{\delta}$ estiver muito próximo de 0 ou quando $\widetilde{\delta}$ estiver muito próximo de 1? Justifica o mais detalhadamente possível a tua resposta.

   b) Encontra uma expressão para calcular a equivalente à dada e que é numericamente mais estável.

4. Diz, justificando, se concordas ou não com a seguinte afirmação: se $\widetilde{x} \neq 0$ aproxima $x \neq 0$ com dez algarismos corretos , então $1/\widetilde{x}$ aproxima $1/x$ com o mesmo número de algarimos corretos.

5. **a)** Para aproximar $\sqrt{1000}$, no Matlab, executa

   ```
   >> f= ...; tol=2^-48; [raiz evals]=bisec(f,31,32,tol)
   ```

   completando a definição de f. Na tua folha de respostas deves escrever o comando completo executado no Matlab, o valor de raiz (em format long) e o número de iterações realizadas.

   b) Usando valores de $tol = 2^{-k}$ para valores de k maiores do que 48 não altera a aproximação obtida nem o número de iterações realizadas. Porquê?

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

RESOLUÇÃO

1. **a)** O erro absoluto $|x - fl(x)|$ depende do expoente do número mas não o erro relativo. Para qualquer número de $\mathcal{F}$ tem-se

$$\frac{|x - fl(x)|}{|x|} < 2^{-52}$$

seja qual for o modo de arredondamento utilizado. No caso do modo de arredondamento para o mais próximo, tem-se

$$\frac{|x - fl(x)|}{|x|} \leq 2^{-53}.$$

b) Há muitos números inteiros menores do que realmax que não pertencem a $\mathcal{F}$. Se x é um número de $\mathcal{F}$ com expoente E então sucessor de x é $x + 2^{E-52}$. Assim, com expoente $E = 53$ os números de $\mathcal{F}$ são

$$2^{53},\ 2^{53} + 2,\ 2^{53} + 4, ...$$

com expoente $E = 54$ são

$$2^{54},\ 2^{54} + 4,\ 2^{54} + 8, ...$$

com expoente $E = 55$ são

$$2^{55},\ 2^{55} + 8,\ 2^{55} + 16, ...$$

etc.

2. **a)** Uma vez que para $x = -0.1$ a série é alternada, o erro de truncatura (em valor absoluto) será inferior ao valor do primeiro termo desprezado. De

```
>> k=9; x=-0.1; x^k/factorial(k)

ans =

  -2.7557e-15

>> k=10; x=-0.1; x^k/factorial(k)

ans =

   2.7557e-17
```

concluímos que para garantir um erro de truncatura inferior a $1e - 16$ será necessário somar até $k = 9$ (inclusive).

b) Basta ter em conta que para θ entre 0 e x é

$$e^\theta < e^x.$$

Donde vem

$$R_k(x) < \frac{e^x}{(k+1)!}x^{k+1}$$

e

$$\frac{R_k(x)}{e^x} < \frac{x^{k+1}}{(k+1)!}.$$

c) A função expTaylor calcula a soma dos termos da série que são superiores a *tol* (tolerância dada). De

```
>> [soma n]=expTaylor(0.1,1e-10)

soma =

    1.1052


n =

     6


>> k=7; x=0.1; x^k/factorial(k)

ans =

   1.9841e-11
```

concluímos, do que se disse na alínea anterior, que

$$\frac{e^x - soma}{e^x} < 1.9841e - 11.$$

(nota: neste caso não usamos módulos porque o erro é positivo.)

3. **a)** Para valores de $\widetilde{\delta}$ próximos de 1, ocorrerá cancelamento subtrativo no cálculo de $1 - \widetilde{\delta}$ e será produzido um resultado próximo de zero com elevado erro relativo. Este erro relativo será no entanto pouco importante na subtração $1 - \sqrt{1 - \widetilde{\delta}}$. Já o mesmo não acontece para valores de $\widetilde{\delta}$ próximos de 0 porque ainda que o erro em $1 - \widetilde{\delta}$ seja pequeno, o cancelamento subtrativo na operação final causará um elevado erro relativo em $\widetilde{a}$ que será tanto maior quanto mais pequeno for $|\widetilde{a}|$.

   b) Para valores de $\widetilde{\delta}$ próximos de 0, a expressão

$$\frac{\widetilde{\delta}}{1 + \sqrt{1 - \widetilde{\delta}}}$$

é equivalente a $1 - \sqrt{1 - \widetilde{\delta}}$ e evita o cancelamento subtrativo referido.

4. A afirmação é verdadeira porque o número de condição relativo da função $f(x) = 1/x$ é

$$\left| x\frac{f'(x)}{f(x)} \right| = \left| x\frac{-1/x^2}{1/x} \right| = 1.$$

5. **a)** O valor procurado é a raiz da equação $x^2 = 1000$.

```
>> f=@(x) x^2-1000; tol=2^-48; [raiz evals]=bisec(f,31,32,tol)

raiz =

    31.622776601683793


evals =

    48
```

b) Tem-se

```
> f(raiz)

ans =

     0
```

e a função bisec termina a execução neste ponto com $a = b = raiz$ independentemente do valor de $tol > 0$. Observe-se, para x entre 32 e 32, 2^{-48} é distância entre x e o seu sucessor.