Introdução ao Cálculo Numérico

Calculando funções numericamente

Prof, Rudnei Dias da Cunha

- Seja numa calculadora científica ou numa linguagem de programação, sempre há disponível um conjunto de funções básicas, tais como as **trigonométricas**, **raiz quadrada**, **exponencial** e **logaritmo**.
- Essas funções podem ser calculadas usando diferentes técnicas, como:
  - Aproximação sucessiva;
  - Séries de potências;
  - Interpolação.
- Veremos, a seguir, exemplos dessas técnicas:

- Cálculo da raiz quadrada
  - Uma forma de calcular $x = \sqrt{a}$ é através de um método de aproximação sucessiva, denominado de *método babilônico*, o qual pode ser descrito pelos seguintes passos:

    1. Escolha $x_0 \cong \sqrt{a}$
    2. Calcule $x_{k+1} = \frac{1}{2}\left(x_k + \frac{a}{x_k}\right)$,

    para $k \geq 0$, até que $|x_{k+1} - x_k| < \epsilon$
  - O último valor calculado de x_{k+1} aproxima $\sqrt{a}$, sujeito ao valor de tolerância ϵ, o qual é escolhido de forma que o valor de x_{k+1} seja "exato" (i,e,, na precisão do computador).

- Cálculo da raiz quadrada
  - Vejamos um exemplo: suponha que se deseja calcular $x = \sqrt{12,34} = 3,5128336145$, usando como estimativa inicial o valor $x_0 = \frac{a}{2} = 6,17$.
  - As primeiras cinco iterações do método babilônico nos fornecem:

$$
\begin{aligned}
x_1 &= \frac{1}{2}(6,17 + 12,34/6,17) = 4,08500000000 \\
x_2 &= \frac{1}{2}(4,085 + 12,34/4,085) = 3,55290391677 \\
x_3 &= 3,51305957000 \\
x_4 &= 3,51283362135 \\
x_5 &= 3,51283361405
\end{aligned}
$$

Observe como o número de dígitos corretos aumenta rapidamente à medida que os valores x_k se aproximam do valor esperado!

- Observe que o método babilônico nada mais é do que o método de Newton-Raphson:
    - O problema que queremos resolver é calcular $x = \sqrt{a}$.
    - Um simples manipulação algébrica nos permite escrever esse problema na forma $f(x) = 0$:

$$x^2 - a = 0,$$

    de onde vemos que estamos buscando a raiz da função $f(x) = x^2 - a$.
    - Escrevendo a equação governante do método de Newton-Raphson para essa função, temos:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{{x_k}^2 - a}{2x_k} = \frac{1}{2}\left(x_k + \frac{a}{x_k}\right) \blacksquare$$

Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

- Outro método para calcular $x = \sqrt{a}$ é utilizar a *série de Taylor* para $\sqrt{s^2 + r}$:
  1. Escolha $s \cong \sqrt{a}$ e calcule $r = a - s^2$.
  2. Escolha um certo número de termos, N, da série

$$\sqrt{s^2 + r} = S_\infty = \sum_{n=0}^{\infty} \frac{(-1)^n (2n)!\, r^n}{(1-2n)(n!)^2 4^n s^{2n-1}}$$

  e calcule a *série truncada* S_N; tal número depende da *precisão* do computador e da *exatidão* desejada!

- Usando esse procedimento para calcular o valor de $x = \sqrt{12,34}$, temos $s = 3$ e $r = a - s^2 = 3,34$.
- Com $N = 8$ termos na série, obtemos:

$$S_2 = 3,5566666667$$
$$S_3 = 3,5050203703$$
$$S_4 = 3,51460362757$$
$$S_5 = 3,51460362757$$
$$S_6 = 3,51295827400$$
$$S_7 = 3,51300513731$$
$$S_8 = 3,51299100672$$

- Observe que, mesmo com oito termos, a aproximação não é tão boa quanto com o método babilônico, oferecendo apenas 3 dígitos decimais em concordância, e a um custo computacional bem maior!

- Vejamos agora o cálculo de e^x, baseado na série de Taylor, porém valendo-se de algumas propriedades matemáticas para acelerar a convergência da série.
- A série de Taylor para e^x é

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120} + O(x^6)$$

- Verifica-se que essa série converge rapidamente para e^x quando $x < 1$.

- Já que a série de Taylor é um polinômio, ela pode ser reescrita usando *multiplicações aninhadas*, com as quais eliminamos a necessidade de se calcular potências:

$$e^x = 1 + x\left(1 + \frac{x}{2}\left(1 + \frac{x}{3}\left(1 + \frac{x}{4}\left(1 + \frac{x}{5}\right)\right)\right)\right) + O(x^6)$$

a qual oferece, além disso, maior estabilidade numérica (dependendo do valor de x).

- Agora, para obtermos uma rápida convergência da série, devemos usá-la com um argumento que seja menor do que 1; para tal, escrevemos

$$z = \lfloor x \rfloor$$

$$f = x - z \therefore f < 1 \therefore$$

de onde obtemos

$$e^x = e^{z+f} = e^z e^f$$

$$e^f = 1 + f\left(1 + \frac{f}{2}\left(1 + \frac{f}{3}\left(1 + \frac{f}{4}\left(1 + \frac{f}{5}\right)\right)\right)\right).$$

- Como z é inteiro, podemos calcular e^z através de z multiplicações de

$$e^1 = 2{,}71828182846$$

(esse valor deve ser o mais preciso o possível).

- Agora, e^f irá convergir rapidamente, já que, por definição, $f < 1$; portanto, podemos calcular e^f com poucos termos, usando a expressão com multiplicações aninhadas.

- Por exemplo, para $x = 3{,}1723$, podemos calcular e^x usando o procedimento aqui ilustrado, com k termos na série para e^f, obtendo o resultado tabulado a seguir:

k	$e^x = e^z \text{taylor}(k,f)$	DIGSE
2	23,84441751477508800	3
3	23,86154083693713100	4
4	23,86227842403926000	5
5	23,86230384129080000	7
6	23,86230457118954000	9
7	23,86230458915547800	10
8	23,86230458954241800	12
9	23,86230458954982600	14
10	23,86230458954995400	15

- Podemos nos perguntar; mas por que não usar simplesmente a série de Taylor para e^x? Vejamos o que ocorre, para o mesmo valor $x = 3{,}1723$, nesse caso:

k	e^x=taylor(k,x)	DIGSE
2	9,20404364499999870	0
3	14,5247771000111640	0
4	18,7445177848441450	0
5	21,4217744597432790	0
6	22,8372846847070330	1
7	23,4787736970859630	1
8	23,7331481463321730	2
9	23,8228094869259190	2
10	23,8512527540024770	3

- Observe que:
  - Até $k = 5$, nenhum resultado é aceitável (DIGSE=0).
  - Para $k = 10$, DIGSE=3 nesse caso, ao passo que alcançamos anteriormente essa exatidão com $k = 2$!
  - Isso demonstra que o procedimento é eficiente (e também eficaz)!

- Considere agora que $y = e^x$ é um número em ponto-flutuante com mantissa m e expoente n tal que $y = m \times 2^n$, com n inteiro e m pequeno.
- Escrevendo a igualdade

$$m \times 2^n = e^x$$

e aplicando logaritmos dos dois lados da igualdade, obtemos

$$\ln y = \ln m + n \ln 2 = x.$$

- Como n deve ser um número inteiro, podemos calculá-lo como

$$n = \left\lfloor \frac{x}{\ln 2} \right\rfloor$$

- Podemos escrever, ainda,

$\ln m = x - n \ln 2 = u$, de onde $m = e^u$

e, pela escolha de n, temos $0 \leq u \leq \ln 2 < 1$.

- Dessa forma, podemos usar a série truncada de Taylor para calcular m, usando multiplicações aninhadas, sabendo que a convergência será rapidamente obtida, com poucos termos na série.

- Por exemplo, para $x = 3{,}1723$, podemos calcular $e^x = m \times 2^n$, com k termos na série para e^u, obtendo o resultado tabulado a seguir:

k	e^x=taylor(k,u)*2^n	DIGSE
2	23,0344568664383260	1
3	23,8012563153801050	2
4	23,8608481023319040	4
5	23,8625486184761610	4
6	23,8623444583434860	5
7	23,8623082003095420	6
8	23,8623048361557510	7
9	23,8623046034280860	9
10	23,8623045902212210	10

Observe que os resultados apresentam menor exatidão do que para *e^x=e^ztaylor(k,f)*, mas a vantagem aqui é o uso de aritmética binária!

- Vejamos agora como calcular a função

$$f(z) = \frac{e^z - 1}{z}, |z| > 0$$

quando:
  - $|z| \ll 1$;
  - ocorre *underflow* em e^z;
  - demais casos.
- Observe que

$$\lim_{z \to 0^+} f(z) = 1.$$

- Uma função $f(x)$: $\mathbb{R} \to \mathbb{R}$ pode ser analisada quanto à qualidade numérica de sua avaliação em x usando a definição de **número de condição da função**:

$$\boxed{k_f(x) = \begin{cases} \dfrac{|x||f'(x)|}{|f(x)|}, & |f(x)| > 0 \\ \infty, & c.c. \end{cases}}$$

- Quanto maior for o valor de $k_f(x)$, mais **mal condicionada** é a função.

- Quando $|z| \ll 1$, então e^z pode ou não sofrer *underflow*.
- Por exemplo, num computador com aritmética de ponto-flutuante IEEE-754 em precisão dupla, $e^z = 1$ para $|z| \leq 10^{-16}$:
  - Mas observe que, se $|z| \neq 0$, então $e^z \neq 1$!
- Por isso, ocorrerá cancelamento catastrófico no numerador da função quando $e^z \cong 1$ (ou seja, $z \cong 0$).

- Por exemplo, calculando $k_f(x)$ para valores de x próximos de $\varepsilon_M = 2{,}220446 \times 10^{-16}$ num computador com aritmética de ponto-flutuante IEEE-754 em precisão dupla, teremos:

x	$k_f(x)$
$4\varepsilon_M$	$8{,}8817841970012523 \times 10^{-16} = 4\varepsilon_M$
$2\varepsilon_M$	$4{,}4408920985006262 \times 10^{-16} = 2\varepsilon_M$
ε_M	$2{,}2204460492503131 \times 10^{-16} = \varepsilon_M$
$\frac{\varepsilon_M}{2}$	∞

- A tabela a seguir mostra alguns valores de $f(z)$, calculados num computador com aritmética de ponto-flutuante IEEE-754 em precisão dupla:

$\|z\|$	$f(z)$
$8{,}88178419700125232 \times 10^{-16}$	$1{,}00000000000000000$
$7{,}77156117237609574 \times 10^{-16}$	$1{,}14285714285714280$
$6{,}66133814775093924 \times 10^{-16}$	$1{,}00000000000000000$
$5{,}55111512312578270 \times 10^{-16}$	$0{,}80000000000000000$
$4{,}44089209850062616 \times 10^{-16}$	$1{,}00000000000000000$
$3{,}33066907387546962 \times 10^{-16}$	$1{,}33333333333333248$
$2{,}22044604925031308 \times 10^{-16}$	$1{,}00000000000000000$

- Analisemos os resultados na tabela:
  - Observe que os valores calculados para $f(z)$ oscilam, e para essa função, isso indica que há algum problema com o cálculo (já que $f(z)$ decresce monotonicamente com z).
  - Uma maneira de eliminar esse problema é utilizar novamente uma expansão em séries de Taylor da função e^z em torno de $z = 0$ e usá-la ao invés de e^z na função $f(z)$, obtendo uma função alternativa $F(z)$.

- Analisemos os resultados na tabela:
  - Por exemplo, utilizando uma expansão em séries de Taylor de ordem 10, truncada e substituindo-a na expressão para $f(z)$, obtemos

$$F(z) = 1 + \frac{z}{2!} + \frac{z^2}{3!} + \frac{z^3}{4!} + \cdots + \frac{z^8}{9!}$$

e, calculando o seu número de condição, verifica-se que ela é bem-condicionada:

x	$k_{\boldsymbol{F}}(x)$
$4\varepsilon_M$	$4{,}4408920985006327072 \times 10^{-16}$
$2\varepsilon_M$	$2{,}220446049250313088 \times 10^{-16} = \varepsilon_M$
ε_M	$1{,}110223024625156768 \times 10^{-16}$
$\frac{\varepsilon_M}{2}$	$5{,}551115123125782720 \times 10^{-17}$

Porém, quando devemos usar uma ou outra?

- A tabela a seguir mostra os valores de $f(z)$ e $F(z)$, calculados num computador com aritmética de ponto-flutuante IEEE-754:

$\|z\|$	$f(z)$	$F(z)$	Erro relativo
10^{-1}	1,05170918075647128	1,05170918075647601 6	$1{,}05563690508681528 8 \times 10^{-15}$
10^{-2}	1,00501670841679491 3	1,00501670841680557 1	$1{,}06049391489134370 4 \times 10^{-14}$
10^{-3}	1,00050016670838459 7	1,00050016670834174 3	$4{,}28331850173731078 4 \times 10^{-14}$
10^{-4}	1,00005000166714097 5	1,00005000166670843 2	$4{,}32521263609893670 4 \times 10^{-13}$
10^{-5}	1,00000500000696490 6	1,00000500001666670 1	$9{,}70174641409562534 4 \times 10^{-12}$
10^{-6}	1,00000049996218365 3	1,00000050000016682 5	$3{,}79831531714978560 0 \times 10^{-11}$
10^{-7}	1,00000004943368026 0	1,00000005000000169 5	$5{,}66321405876147212 8 \times 10^{-10}$
10^{-8}	0,99999999392252902 4	1,00000000499999997 0	$1{,}10774708851474401 6 \times 10^{-08}$
10^{-9}	1,00000008274037099 9	1,00000000050000004 1	$8{,}22403709165999964 8 \times 10^{-08}$
10^{-10}	1,00000008274037099 9	1,00000000005000000 4	$8{,}26903709908188310 4 \times 10^{-08}$

Calculando funções numericamente

- Observe que, agora, o comportamento da função $F(z)$ é o esperado, decrescendo monotonicamente com z.
- Note que, pela tabela, podemos determinar quando $F(z)$ deve ser utilizada; por exemplo, para $|z| \leq 10^{-5}$.