INSTITUTO FEDERAL
Paraná
Campus Pinhais

Análise numérica

Erros, bases e ponto flutuante

Gabriel V C Candido
gabriel.candido@ifpr.edu.br

Instituto Federal do Paraná - Pinhais

Sumário

Sumário

Cálculo numérico

- Nossa ideia é explorar métodos *construtivos* para a solução de modelos matemáticos.
- Geralmente, isso resulta em aproximações da solução exata. Portanto, também precisamos avaliar o quão longe estamos da resposta.

Cálculo numérico

- Cálculo numérico é a obtenção da solução pela aplicação de método numérico; a solução é, então, um conjunto de números, exatos ou aproximados.
- Método numérico é um algoritmo finito de operações envolvendo apenas números (operações aritméticas elementares, cálculo de funções, consulta a uma tabela de valores, ...).

Cálculo numérico

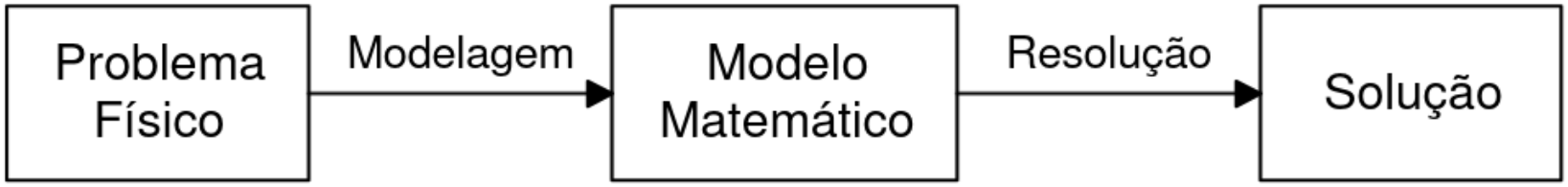

Figura: Fases de um problema. Fonte: SF 1

- Modelagem é a fase de obtenção do modelo matemático que descreve o comportamento do sistema físico/real;
- Resolução é a fase de obtenção da solução através da aplicação de métodos numéricos.

Sumário

Conceito

A noção de erro está presente em todos os campos do cálculo numérico.

Definição

Erro é a diferença entre o valor exato e o valor apresentado.

Erros na fase de modelagem

Raramente um modelo matemático consegue descrever corretamente um fenômeno do mundo físico. Normalmente existem várias simplificações envolvidas.

Exemplo

Medição de peso em uma balança. A gravidade na superfície da Terra não é constante, o peso pode mudar de acordo com altitude, pressão, ...

Erros na fase de modelagem

Outro exemplo

Equação do movimento com aceleração constante:

$$d = d_0 + v_0 * t + \frac{a*t^2}{2}$$

Erros na fase de modelagem

Outro exemplo

Equação do movimento com aceleração constante:
$d = d_0 + v_0 * t + \frac{a*t^2}{2}$

Determinar a altura de um edifício jogando uma bolinha e cronometrando: 3s

Erros na fase de modelagem

Outro exemplo

Equação do movimento com aceleração constante:
$d = d_0 + v_0 * t + \frac{a*t^2}{2}$

Determinar a altura de um edifício jogando uma bolinha e cronometrando: 3s

$d = 0 + 0 * 3 + (9.8 * 3^2)/2 = 44.1\text{m}$

Erros na fase de modelagem

Outro exemplo

E a resistência do ar? E a velocidade do vento? ...

Erros na fase de modelagem

Outro exemplo

E a resistência do ar? E a velocidade do vento? ...

Considere a precisão dos dados de entrada:

- Com 3.5s de cronômetro: $d = 60.025$m;
- Variação de 16% no cronômetro provocou uma variação de 36% na altura calculada.

Erros na fase de resolução

Para resolver os modelos matemáticos, muitas vezes precisamos de aproximações no cálculo, por conta do instrumento usado.

Conversão de bases, erros de arredondamento, erros de truncamento.

Como representar infinitos números reais entre A e B em um computador com memória finita? *Precisamos* de aproximações!

Toda medida tem erro!

Erro absoluto

Definição

Erro absoluto (EA) é a diferença entre o valor exato de um número x e o seu valor aproximado $\overline{x}$.

$$EA_x = x - \overline{x}$$

Erro absoluto

Geralmente não conhecemos o valor de x, impossibilitando o cálculo de EA_x. Mas podemos obter um limitante superior ou estimativa do módulo do EA.

Exemplo

$\pi \in (3.14, 3.15)$. Se escolhermos um valor dentro desse intervalo para π, então $|EA_\pi| = |\pi - \overline{\pi}| < 0.01$.

Erro relativo

Definição

Erro relativo (ER) é a divisão do erro absoluto pelo valor aproximado.

$$ER_x = EA_x/\overline{x}$$

Erro relativo

Exemplo

Suponha que temos $\alpha = 3876.373$ e aproximaremos apenas a parte inteira. Então
$EA_\alpha = |\alpha - \overline{\alpha}| = 0.373.$

Exemplo

Agora vamos fazer o mesmo com $\beta = 1.373$.
$EA_\beta = |\beta - \overline{\beta}| = 0.373.$

O EA é o mesmo, mas o efeito de aproximação em β é muito maior.

Erro relativo

O erro relativo "resolve" esse problema:

- $ER_\alpha = 0.373/3876 \approx 0.000096 < 10^{-4}$
- $ER_\beta = 0.373/1 = 0.373 < 10^{0}$

Erro relativo

Outro exemplo

- $\overline{x} = 2112.9$ e $|EA_x| < 0.1$, ou seja, $x \in (2112.8, 2113)$;
- $\overline{y} = 5.3$ e $|EA_y| < 0.1$, ou seja, $y \in (5.2, 5.4)$;

- $ER_x = EA_x/\overline{x} < 0.1/2112.9 \approx 4.7 \times 10^{-5}$;
- $ER_y = EA_y/\overline{y} < 0.1/5.3 \approx 0.02$;

Logo, x é representado com maior precisão que y.

Erro absoluto e relativo

Exercício

- O número de Euler e é aproximadamente $2.7182\ldots$. Dê o limitante superior do erro absoluto, similar ao que fizemos com π, considerando 4 casas decimais de e;
- Faça o mesmo com π, considerando 5 casas decimais;
- Utilize o erro relativo para mostrar qual dos dois números (e e π) está sendo representado com maior precisão.

Erros de truncamento

Truncar um número até a p-ésima casa decimal significa simplesmente desconsiderar todos os dígitos após essa casa.

Considere o número 2.345678. Podemos truncar considerando:

- Duas casas decimais: 2.34
- Três casas decimais: 2.345
- Quatro casas decimais: 2.3456

Erros de truncamento

Exemplo

Considere a seguinte fórmula para cálculo do seno:

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

Não podemos calcular infinitamente. Precisamos interromper o cálculo em algum momento: o objetivo é atingir uma precisão mínima.

Erros de arredondamento

Podemos definir diversas regras. Por exemplo: considerar o primeiro dígito a ser descartado. Se for ≥ 5, somamos um ao último dígito representado. Caso contrário, mantemos o último dígito.

Considere o número 2.3455. Podemos arredondar considerando:

- Uma casa decimal: 2.3
- Duas casas decimais: 2.35
- Três casas decimais: 2.346

Erros

Calcular a área de uma circunferência de raio 100m.

$$A = \pi * r^2$$

$$A = 3.14 * r^2 = 31400m^2$$
$$A = 3.1416 * r^2 = 31416m^2$$
$$A = 3.141592654 * r^2 = 31415.92654m^2$$

Sumário

Sistema decimal e binário

O sistema decimal, base 10, usa os dígitos de 0 a 9, de forma posicional. Portanto, os números podem ser representados como polinômios:
$347_{10} = 3 \times 10^2 + 4 \times 10^1 + 7 \times 10^0$.

A base binária também é posicional, usando os dígitos 0 e 1:
$10110_2 = 1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 0 \times 2^0$

Conversão para decimal

Para converter um número de uma base β para a base decimal, usamos o polinômio do número.

$0.001_2 = (?)_{10}$

Conversão de decimal

Para converter um número da base decimal para uma base β, usamos divisões sucessivas, e consideramos o resto (de trás para frente).

$23_{10} = (?)_2$

Conversão de decimal fracionário para binário

Algorithm 1 Conversão Fracionária $N_{10} \rightarrow N_2$

1: $k = 1, F = r_{10}$
2: **Faça:**
3: $\quad F = 2 \times F$
4: $\quad d_k = parteInteira(F)$
5: $\quad F = F - d_k$
6: $\quad k = k + 1$
7: **Enquanto** $(F > 0)$

Figura: Conversão de um decimal fracionário para binário.

INSTITUTO FEDERAL
Paraná
Campus Pinhais

Sumário

Ponto flutuante

Números na forma $\pm(d_1 d_2 \ldots d_t) \times \beta^e$.

- β é a base, geralmente binária
- e é o expoente

Ponto flutuante

Exemplo

Considere o exemplo na base decimal:
$\beta = 10; t = 3; e \in [-5, 5]$

Os números estão na forma $0.d_1d_2d_3 \times 10^e$.

Vamos tentar representar
$x = 235.89 = 0.23589 \times 10^3$. Como só temos 3 dígitos na mantissa, podemos:

- truncar: 0.235×10^3;
- arredondar: 0.236×10^3.

Ponto flutuante

Outro exemplo

Considere o exemplo na base decimal:
$\beta = 10; t = 3; e \in [-5, 5]$

Os números estão na forma $0.d_1 d_2 d_3 \times 10^e$.

Vamos tentar representar
$x = 875 \times 10^6 = 0.875 \times 10^9$.

Não podemos representar esse número por conta do expoente: *overflow*.

Ponto flutuante

Outro exemplo

Considere o exemplo na base decimal:
$\beta = 10; t = 3; e \in [-5, 5]$

Os números estão na forma $0.d_1 d_2 d_3 \times 10^e$.

Vamos tentar representar
$x = 0.00345 \times 10^{-5} = 0.345 \times 10^{-7}$.

Não podemos representar esse número por conta do expoente: *underflow*.

Ponto flutuante

Outro exemplo

Considere agora o exemplo na base binária: 3.5_{10}.

Convertendo para binário:
$11.1_2 = 11.1_2 \times 2^0 = 1.11_2 \times 2^1 = 0.111_2 \times 2^2$.

Considerando uma máquina
$\beta = 2; t = 4; e \in [-3, 3]$, o valor armazenado será
0.111×2^2.

Considerando uma máquina
$\beta = 2; t = 2; e \in [-3, 3]$, o valor armazenado será
0.11×2^2.

Ponto flutuante

Por conta dos erros na aritmética de ponto flutuante, a ordem das operações pode mudar seu resultado.

Supondo uma máquina $\beta = 10; t = 4; e \in [-4, 4]$, realize os seguintes cálculos considerando $x_1 = 0.3491 \times 10^4$ e $x_2 = 0.2345 \times 10^0$.

- $(x_2 + x_1) - x_1$
- $x_2 + (x_1 - x_1)$

Primeiro, precisamos igualar os expoentes e depois somar as mantissas.

Ponto flutuante

Supondo uma máquina $\beta = 10; t = 4; e \in [-4, 4]$, realize os seguintes cálculos considerando $x_1 = 0.3491 \times 10^4$ e $x_2 = 0.2345 \times 10^0$.

- $(x_2 + x_1) - x_1$
- $x_2 + (x_1 - x_1)$

$x_2 = 0.2345 \times 10^0 = 0.00002345 \times 10^4$. Mas a máquina posssui $t = 4$, portanto, esse número será representado como 0.0000×10^4.