CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
DEPARTAMENTO DE MATEMÁTICA

Nome Sobrenome

Métodos Numéricos

SEROPÉDICA
2025

Nome Sobrenome

MÉTODOS NUMÉRICOS

Monografia apresentada à Banca Examinadora da Universidade Federal Rural do Rio de Janeiro, como parte dos requisitos para obtenção do título de Bacharel em Matemática sob orientação do Prof. Dr. Nome Sobrenome do Orientador

SEROPÉDICA

2025

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE MATEMÁTICA

COORDENAÇÃO DO CURSO DE GRADUAÇÃO EM MATEMÁTICA

A monografia "Métodos Numéricos", apresentada e defendida por NOME SOBRENOME, matrícula 2022019000-0, foi aprovada pela Banca Examinadora com conceito "X", recebendo o número 000.

Seropédica, 7 de agosto de 2025

BANCA EXAMINADORA:

Prof. Dr. Presidente da Banca
Orientador

Prof. Dr. Membro 1
Convidado 1

Prof. Dr. Membro 2
Convidado 1

Agradecimentos

Aqui está um agradecimento.

Resumo

Aqui está um resumo.

Abstract

Here is an (optional) abstract.

Sumário

Introdução **ii**

1 Pontos Flutuantes **1**

1.1 Aritmética de Ponto Flutuante 2

1.1.1 Precisão Simples e Precisão Dupla 3

1.1.2 Representação de Números em Sistemas de Ponto Flutuante 6

1.1.3 Representação Especial do Zero 9

1.2 Erros e Limitações 9

1.2.1 Erro Absoluto e Relativo 10

1.3 Perda de Significância em Operações com Pontos Flutuantes 11

1.4 Análise de Instabilidades e Casos Peculiares 13

1.4.1 Imprecisão de operações de Ponto flutuante 14

1.4.2 Discussão 17

2 Métodos Iterativos para Zeros de Função **18**

2.1 Localização de Raízes 18

2.2 Critério de Parada 20

2.3 Método do Ponto Fixo 21

2.3.1 Ordem de convergência 23

2.4 Método de Newton-Raphson 25

2.4.1 Demonstração Geométrica 26

2.4.2 Convergência 26

2.4.3 Ordem de Convergência 26

2.4.4 Ciladas 27

2.4.5 Fractais 27

Referências Bibliográficas **29**

Introdução

Capítulo 1

Pontos Flutuantes

Na matemática, um sistema numérico é um conjunto de regras e símbolos utilizados para representar quantidades através do que chamamos de números. Existem dois tipos de sistemas: os posicionais e os aditivos.

O sistema aditivo é aquele em que os números são representados pelas somas dos valores dos símbolos, geralmente agrupados lado a lado em ordem decrescente, como, por exemplo, os sistemas romano e egípcio. Já o sistema posicional leva em conta não só os dígitos mas também a posição que eles ocupam no número. A quantidade de símbolos diferentes que são utilizados para representar os dígitos está ligada à **base** desse sistema, e cada posição do dígito no número refere-se a uma potência dessa base. Por exemplo, no sistema decimal (base 10), usamos os dígitos de 0 a 9. No sistema binário (base 2), usamos os dígitos 0 e 1. E já no sistema hexadecimal (base 16), usamos de 0 a 9 e as letras A a F (que representam 10 a 15).

Com essas diferentes formas de representar um número, a escolha do sistema depende do contexto e da aplicação. No uso cotidiano, a base decimal é a mais utilizada. Já as bases binária e hexadecimal, são amplamente utilizadas na ciência da computação em operações aritméticas dos processadores e em algumas linguagens de programação para endereçamento de memória.

Um número N pode ser representado em uma base b no seguinte formato

$$N = \pm \sum_{i=-k}^{n} d_i b^i, \tag{1.1}$$

em que d_i são os dígitos na base b, k é o número de casas decimais à direita do ponto, e $n+1+k$ é o número de dígitos significativos. Vejamos alguns exemplos.

Exemplo 1.0.1. Vamos escrever o número 13 nas bases 10 e 2.

- Número na base decimal: $13 = 1 \times 10^1 + 3 \times 10^0 = 13_{10}$
- Número na base binária: $13 = 1 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 = 1101_2$

Exemplo 1.0.2. Agora vamos escrever o número 3,5625 nas bases 10 e 2.

- Número na base decimal:

$$3{,}5625 = 3 \times 10^0 + 5 \times 10^{-1} + 6 \times 10^{-2} + 2 \times 10^{-3} + 5 \times 10^{-4} = 3{,}5625_{10}$$

- Número na base binária:

$$3{,}5625 = 1 \times 2^1 + 1 \times 2^0 + 1 \times 2^{-1} + 0 \times 2^{-2} + 0 \times 2^{-3} + 1 \times 2^{-4} = 11{,}1001_2$$

1.1 Aritmética de Ponto Flutuante

A *aritmética de ponto flutuante* é o sistema adotado por computadores para que lidem com números reais utilizando uma notação compacta e eficaz. Essa técnica é utilizada para representar e manipular números reais de forma prática e eficiente. Ela permite representar números de grandezas diversas, que não podem ser armazenados com precisão, utilizando apenas números inteiros.

Um sistema de ponto flutuante F pode ser definido como

$$F(\beta, t, L, U)$$

cuja representação normalizada de um número real N nesse sistema é dada por

$$N = \pm(0.d_1 d_2 ... d_t)_\beta \times \beta^e \tag{1.2}$$

em que

- N é o número real;

- β é a base que a máquina opera;
- t é o número de dígitos na mantissa, tal que $0 \leq d_j \leq \beta - 1$, j = 1, ...,t, $d_1 \neq 0$;
- L é o menor expoente inteiro;
- U é o maior expoente inteiro;
- e é o expoente inteiro no intervalo [L,U].

No padrão IEEE 754 (usado na maioria dos sistemas eletrônicos), um número de ponto flutuante é dividido em três partes:

- **Sinal (S)**: 1 bit indicando se o número é positivo ($S = 0$) ou negativo ($S = 1$),
- **Expoente (E)**: campo que representa o expoente com viés (bias),
- **Mantissa (M)**: parte fracionária significativa do número.

A fórmula completa de reconstrução do número é:

$$\text{Valor} = (-1)^S \times (1.M) \times 2^{E-\text{bias}}$$

onde:

- S é o bit de sinal,
- $1.M$ indica que há um bit implícito "1"antes da mantissa nos números normalizados,
- bias é um valor constante que depende da precisão (por exemplo, 127 para 32 bits).

1.1.1 Precisão Simples e Precisão Dupla

Em sistemas computacionais, os números em ponto flutuante podem ser representados em diferentes níveis de precisão. Os dois mais comuns são:

- **Precisão Simples (32 bits)**
- **Precisão Dupla (64 bits)**

Esses formatos seguem o padrão `IEEE 754` de representação binária de números reais.

Comparação entre os formatos

Característica	**Precisão Simples (32 bits)**	**Precisão Dupla (64 bits)**
Bits totais	32	64
Bit de sinal	1	1
Bits de expoente	8	11
Bits de mantissa	23	52
Bias	127	1023
Intervalo do expoente real	-126 a $+127$	-1022 a $+1023$
Precisão (dígitos decimais)	Aproximadamente 7	Aproximadamente 16

Exemplo: Representação em Precisão Simples

Considere o número decimal $x = -12{,}25$. Sua representação em binário é:

$$x = -1100{,}01_2 = -1{,}10001 \times 2^3$$

Formato:

- Sinal: $s = 1$
- Mantissa (sem o bit oculto): 10001000000000000000000
- Expoente: $e = 3 + 127 = 130 = 10000010_2$

Portanto, o número seria representado, em binário de 32 bits, como:

1 10000010 10001000000000000000000

s_n	e	m
1	10000010	10001000000000000000000

Exemplo: Representação em Precisão Dupla

Vamos representar o número decimal $x = 12{,}375$ em ponto flutuante com **precisão dupla (64 bits)**.

1. **Conversão para binário:**

$$12{,}375_{10} = 1100{,}011_2 = 1{,}100011 \times 2^3$$

2. **Identificação dos componentes:**

- **Sinal (s)**: Como o número é positivo, $s = 0$
- **Expoente real (e)**: 3
- **Bias**: Para precisão dupla, bias = 1023
- **Expoente com bias**: $e + \text{bias} = 3 + 1023 = 1026$
- **Expoente em binário (11 bits)**: $1026_{10} = 10000000010_2$
- **Mantissa (m)**: Os bits após o ponto da parte fracionária normalizada: $100011000000\ldots$ (completando até 52 bits)

3. **Representação final (64 bits):**

$$\boxed{0\ 10000000010\ 1000110000000000000000000000000000000000000000000000}$$

Essa é a representação de 12,375 em ponto flutuante com precisão dupla.

Resumo:

- **Bits de sinal**: 0
- **Bits do expoente**: 10000000010
- **Bits da mantissa**: 100011 seguidos de zeros até completar 52 bits

Considerações

A escolha entre precisão simples e dupla depende da aplicação:

- **Precisão Simples**: adequada para aplicações com memória limitada e que não exigem alta precisão.

- **Precisão Dupla**: usada em aplicações científicas, cálculos de engenharia, simulações e algoritmos numéricos mais sensíveis. Apesar do ganho de precisão, o uso de precisão dupla demanda mais memória e tempo de processamento.

1.1.2 Representação de Números em Sistemas de Ponto Flutuante

Em máquinas que operam em sistemas de ponto flutuante, apenas um subconjunto finito de $\mathbb{R}$ pode ser representado de maneira exata. Por isso, frequentemente, é necessário limitar a quantidade de dígitos significativos na representação de números a fim de adequá-los ao sistema que a máquina opera. Dois dos principais processos empregados para este fim são o **truncamento** e o **arredondamento**.

O truncamento consiste na supressão de todos os dígitos após uma determinada posição, sem qualquer ajuste adicional no último dígito mantido. Formalmente, dado um número real x, sua aproximação truncada com n dígitos na base b é expressa por

$$T(x) = \sum_{i=-k}^{n-1} d_i b^i$$

onde os dígitos d_i com $i < -k$ são descartados.

O erro introduzido por este processo, dado por $E_T = x - T(x)$, é denominado *erro de truncamento*. Ele é limitado superiormente por

$$|x - T(x)| < b^{-k}. \tag{1.3}$$

Exemplo 1.1.1. Considere a aproximação com oito casas decimais de $\pi = 3,14159265$. Para truncar π com precisão de 4 casas decimais, descartamos todos os termos da sexta casa em diante. Assim, o valor truncado fica $T(\pi) = 3,1415$. O erro do truncamento é, nesse caso, $E_T = 3,14159265 - 3,1415 = 0,00009265$. Diante disso, podemos observar que $|E_T| < 10^{-4}$.

O arredondamento, por outro lado, ajusta o último dígito mantido com base no valor do primeiro dígito descartado, buscando minimizar o erro absoluto da aproximação. No arredondamento simétrico (ou clássico), se o primeiro dígito descartado for maior ou

igual a $\frac{b^{-n}}{2}$, incrementa-se o último dígito mantido em uma unidade; caso contrário, seu valor permanece inalterado.

Seja x um número real e $R(x)$ sua aproximação arredondada com n dígitos na base b. O erro de arredondamento satisfaz:

$$|x - R(x)| \leq \frac{1}{2}b^{-n}$$

Exemplo 1.1.2. Ainda considerando a mesma aproximação de $\pi = 3,14159265$. Para arredondar π com precisão de 4 casas decimais, vamos analisar o número da próxima casa em que queremos arredondar. Nesse caso, o número na quinta posição é 9, então vamos arredondar a quarta casa para cima

$$R(\pi) = 3,1416.$$

O erro do arredondamento é, nesse caso,

$$E_R = 3,14159265 - 3,1416 = -0,00000735.$$

Diante disso, podemos observar que $|E_R| \leq \frac{1}{2}10^{-4}$.

Em geral, o erro máximo introduzido pelo arredondamento é metade daquele introduzido pelo truncamento, razão pela qual o arredondamento tende a produzir aproximações mais precisas.

Para compreendermos melhor as limitações na representação de números em um sistema de ponto flutuante, vamos explorar o exemplo a seguir. Suponha que uma máquina opere no sistema $F(10, 5, -5, 5)$. Nesse sistema, os números serão representados da seguinte maneira

$$\pm(0.d_1d_2...d_t) \times 10^e, 0 \leq |d_j| \leq 9, d_1 \neq 0, e \in [-5, 5]. \tag{1.4}$$

O menor valor, em módulo, representado nesse sistema é

$$m = 0.10000 \times 10^{-5} = 10^{-6},$$

enquanto que o maior é

$$M = 0.99999 \times 10^5 = 99999.$$

$$G = \{x \in \mathbb{R} \mid x \text{ corresponde a um número representado no sistema } \mathcal{F}(\beta, t, L, U)\}$$

@LucasM: aqui vale a pena trocar por algo do tipo pontos flutuantes e inclusive inserir uma imagem mostrando que trata-se de um conjunto discreto

é o conjunto dos números que são representáveis por esse sistema de ponto flutuante. Nesse conjunto m é a mantissa Dado um número real x, as seguintes situações podem ocorrer:

Caso 1: $x \in G$ **(Número representável)**

Seja $x = 12237,76$.

Na forma normalizada temos $x = 0,1223776 \times 10^5$. Porém, esse número não pode ser representado precisamente no sistema F e, portanto, precisamos aplicar uma das técnicas de aproximação. Utilizando o truncamento o resultado é $\bar{x} = 0{,}12237 \times 10^5$. Já com o arredondamento $\bar{x} = 0{,}12238 \times 10^5$.

O número está dentro da faixa de expoente permitida e é representável com perda controlada de precisão.

Caso 2: $|x| < m$ **(Underflow)**

Seja $x = 0{,}582 \times 10^{-6}$.

O expoente é -6 , menor que o limite inferior $L = -5$. Portanto, o número não pode ser representado e ocorre **underflow**. Nesse caso, na maioria das vezes o valor é tratado como zero.

Caso 3: $|x| > M$ **(Overflow)**

Seja $x = 0{,}927 \times 10^6$

O expoente é $+6$, maior que $U = 5$. Portanto, ocorre **overflow** e o número não pode ser representado precisamente. Neste caso, o valor pode ser tratado como infinito ou como uma flag indicando o **overflow**.

Vejamos a seguir uma comparação da técnicas de arredondamento e de truncamento.

Considere uma máquina decimal com 3 dígitos na mantissa e expoentes variando de –4 a 4:

Número Real	**Arredondamento**	E_R	**Truncamento**	E_T
5,678	$0{,}568 \times 10^1$	$0{,}2 \times 10^{-3}$	$0{,}567 \times 10^1$	$0{,}8 \times 10^{-3}$
–192,73	$-0{,}193 \times 10^3$	$0{,}27 \times 10^1$	$-0{,}192 \times 10^3$	$0{,}73 \times 10^1$
3,14159	$0{,}314 \times 10^1$	$0{,}159 \times 10^{-2}$	$0{,}314 \times 10^1$	$0{,}159 \times 10^{-2}$
0,0000063	Underflow			
920000,0	Overflow			

1.1.3 Representação Especial do Zero

Na representação de ponto flutuante, um número real é geralmente expresso na forma normalizada:

$$N = \pm(d_1{,}d_2d_3 \ldots d_t)_\beta \times \beta^e,$$

com $d_1 \neq 0$, garantindo o aproveitamento máximo da precisão disponível e evitando representações redundantes. Contudo, o número zero não pode ser representado nesta forma, pois exigiria $d_1 = 0$, o que contraria a normalização.

Assim, o zero recebe uma *representação especial* denotada por

$$N = \pm(0{,}000 \ldots 0_t)_\beta \times \beta^{L-1},$$

em que L é o menor expoente permitido no sistema. Este tratamento especial assegura que o zero seja manipulado de forma única e consistente dentro do sistema de ponto flutuante, evitando, assim, a perda de informação ao realizar operações aritméticas que o envolvam. Essa perda de informação será discutida na seção 1.2.

1.2 Erros e Limitações

TODO: adicionar nessa seção uma subseção falando sobre epsilon de maquina, contextualizando para o experimento com as ULPS nos próximos capitulos

Erros em operações com pontos flutuantes podem se propagar e aumentar em cálculos mais complexos. Por exemplo, pequenos erros de arredondamento em etapas iniciais podem afetar significativamente o resultado final, especialmente em somas repetitivas ou subtrações de números muito próximos. Isso torna importante considerar a ordem das operações e o impacto da precisão em aplicações sensíveis.

1.2.1 Erro Absoluto e Relativo

Em cálculos numéricos, frequentemente lidamos com aproximações devido a arredondamentos e truncamentos. Para avaliar a precisão dessas aproximações, utilizamos as métricas de erro absoluto e erro relativo.

O erro absoluto mede a diferença entre o valor real $\boldsymbol{x_r}$ e o valor aproximado $\boldsymbol{x_a}$, ou seja, a quantidade exata de erro na aproximação. Ele é definido como:

$$\text{EA} = |x_a - x_r|. \tag{1.5}$$

Quanto menor for o erro absoluto, mais próximo o valor aproximado está do valor real. No entanto, essa métrica não fornece informações diretas sobre o impacto do erro em relação à magnitude do número em questão.

O erro relativo indica o quão preciso é o valor aproximado em relação ao valor real. Ele pode ser tratado como um decimal ou na forma de porcentagem. Ele é definido como

$$\text{ER} = \frac{|x_a - x_r|}{|x_r|} \quad \text{ou} \quad \frac{|x_a - x_r|}{|x_r|} \times 100\%. \tag{1.6}$$

Isso pode ser reescrito como

$$\text{ER} = \frac{EA}{|x_r|} \quad \text{ou} \quad \frac{EA}{|x_r|} \times 100\%. \tag{1.7}$$

Essa métrica é útil quando lidamos com valores de grandezas muito diferentes. Por exemplo, um erro absoluto de 0.1 pode ser insignificante se estivermos tratando de números na ordem de milhares, mas pode ser relevante se estivermos lidando com valores decimais.

Vamos analisar dois casos com mesmo erro absoluto mas diferentes erros relativos. Considere um valor real $x_{r1} = 10.5$ com uma aproximação de $x_{a1} = 10.3$. O seu erro absoluto é $|10.3 - 10.5| = 0.2$ e seu erro relativo $\dfrac{0.2}{10.5} \approx 0.019$ (ou 1.9%). Agora considere

um segundo valor real de $x_{r2} = 0.4$ com uma aproximação de $x_{a2} = 0.2$, seu erro absoluto é $|0.4 - 0.2| = 0.2$, o seu erro relativo será $\dfrac{0.2}{0.4} = 0.5$ ou seja, 50%.

Número Real	**Aproximação**	**Erro Absoluto**	**Erro Relativo**
10.5	10.3	0.2	0.019 ou 1.9%
0.4	0.2	0.2	0.5 ou 50.0%

Apesar dos dois casos terem o mesmo erro absoluto ($EA = 0.2$), o erro relativo relacionado à aproximação no primeiro caso representa menos de 2% do valor real, indicando que a aproximação é razoavelmente precisa. Em contrapartida, no segundo caso, o erro relativo é de 50%, um valor relativamente alto comparado ao primeiro, o que indica uma aproximação deficiente. Dessa forma, o erro relativo é uma ferramenta importante para interpretar melhor a qualidade de uma aproximação, independentemente da escala dos números envolvidos.

1.3 Perda de Significância em Operações com Pontos Flutuantes

A perda de significância (também conhecida como cancelamento catastrófico) ocorre de modo mais evidente quando há grande diferença de ordem de grandeza entre os números envolvidos na operação. Por exemplo, considere:

$$x = 1{,}000 \times 10^4 \quad \text{e} \quad y = 0{,}276 \times 10^{-2}.$$

Para realizar a soma, ambos os operandos precisam ser expressos com o mesmo expoente:

$$x = 1{,}000000000 \times 10^4, \quad y = 0{,}000000276 \times 10^4.$$

A soma exata seria:

$$x + y = (1{,}000000000 + 0{,}000000276) \times 10^4 = 1{,}000000276 \times 10^4.$$

Contudo, devido à precisão limitada do sistema (7 dígitos significativos), a aritmética de ponto flutuante armazena apenas:

$$x + y \approx 1{,}0000002 \times 10^4.$$

Uma parte do termo y é completamente desprezada, e a soma não resulta numericamente igual a $x + y$, evidenciando a perda catastrófica de significância.

Outro caso peculiar é a soma de um número muito grande com uma sequência de números pequenos. Dependendo da ordem em que as somas são realizadas, o número grande pode "mascarar"os pequenos, resultando em diferentes valores finais.

Por exemplo:

$$S = 10^8 + 10^{-1} + 10^{-2} + 10^{-3} + \ldots + 10^{-10}.$$

Se somarmos primeiro o número grande (10^8) e depois os números pequenos, muitos destes podem ser ignorados devido à falta de precisão da mantissa. Por outro lado, ao somar os números pequenos antes, o valor final será mais próximo do esperado.

Para ilustrar, suponha a seguinte ordem de cálculo:

- Caso 1: $S = 10^8 + (10^{-1} + 10^{-2} + \ldots + 10^{-10})$.
- Caso 2: $S = (10^{-1} + 10^{-2} + \ldots + 10^{-10}) + 10^8$.

No primeiro caso, muitos números pequenos são ignorados devido ao arredondamento. No segundo, o somatório dos números pequenos é calculado antes de adicionar o número grande, preservando mais informações significativas.

Prevenção da perda de significância em operações com Zero

O zero possui uma representação especial na aritmética de ponto flutuante devido à sua importância nas operações numéricas e à necessidade de evitar ambiguidades. Essa representação permite o tratamento adequado de operações que envolvem valores nulos, prevenindo erros numéricos e a perda de significância em cálculos sensíveis. Considere um sistema de ponto flutuante $F(10, 7, -5, 5)$. Sejam

$$x = 0{,}000 \times 10^{-6} \quad \text{(o número zero)}, \quad y = 0{,}276 \times 10^{-2}.$$

Na operação de soma:

$$x + y = 0{,}276 \times 10^{-2},$$

como x é exatamente zero, o resultado mantém integralmente os dígitos significativos de y.

@EnzoR: @LucasM por favor, verifique se o exemplo ficou bom

Se o zero não tivesse uma representação especial e fosse tratado como um número subnormal com expoente mínimo, o alinhamento das mantissas poderia comprometer a precisão de y, deslocando seus dígitos significativos e resultando em perda de informação. Nesse sentido, suponha que x não tivesse uma representação especial e fosse tratado como um número subnormal com o expoente mínimo -5, com mantissa ajustada para

$$x' = 0{,}000000 \times 10^{-5}.$$

Para somar x' e y, é necessário alinhar os expoentes, o que implica deslocar a mantissa de y para a direita:

$$y = 0{,}276 \times 10^{-2} = 0{,}00000276 \times 10^{-5}.$$

Ao realizar a soma,

$$x' + y = (0{,}000000 + 0{,}00000276) \times 10^{-5} = 0{,}00000276 \times 10^{-5}.$$

Devido à precisão limitada do sistema (7 dígitos significativos), essa operação pode fazer com que os dígitos significativos de y sejam deslocados para posições menos precisas, causando perda de informação relevante.

$$x' + y = 0{,}0000027 \times 10^{-5}.$$

Portanto, a representação especial do zero evita esse problema, preservando a precisão e assegurando a estabilidade numérica das operações.

1.4 Análise de Instabilidades e Casos Peculiares

@EnzoR: @LuisD preciso que verifique se a seção está coerente e se os exemplos estão corretos.

Na aritmética de ponto flutuante, certos casos resultam em erros devido à limitação da precisão e à maneira como os números são representados. A seguir, descrevemos dois exemplos clássicos que ilustram essas instabilidades.

1.4.1 Imprecisão de operações de Ponto flutuante

Considere o cálculo de $f(x) = x^{10} + 1 - x^{10}$ para $x \in [-60, 60]$. As partes envolvendo x são variáveis e a parte "1"é o literal, ou seja, um valor fixo. Analiticamente, o resultado deveria ser exatamente 1. No entanto, em implementações numéricas, pequenas imprecisões na representação de x^{10} podem levar a resultados instáveis, especialmente para valores de $|x|$ além de um limiar. Isso ocorre devido ao erro relacionado às operações de ponto flutuante, como mostrado na Figura 1.1.

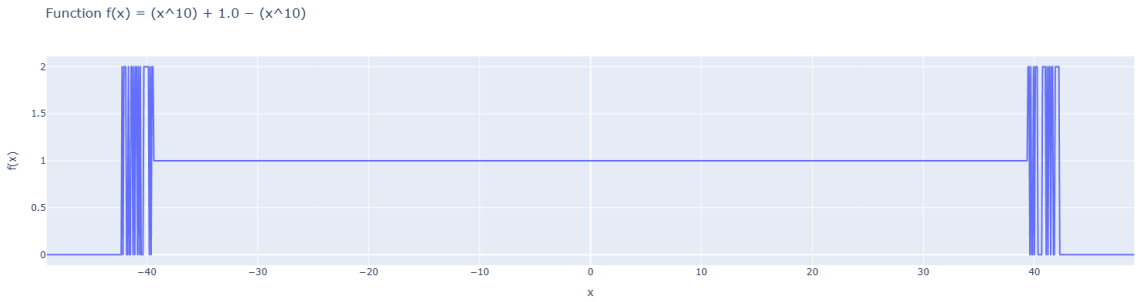

Figura 1.1: Comportamento da expressão $x^{10} + 1 - x^{10}$ no intervalo $[-40, 40]$.

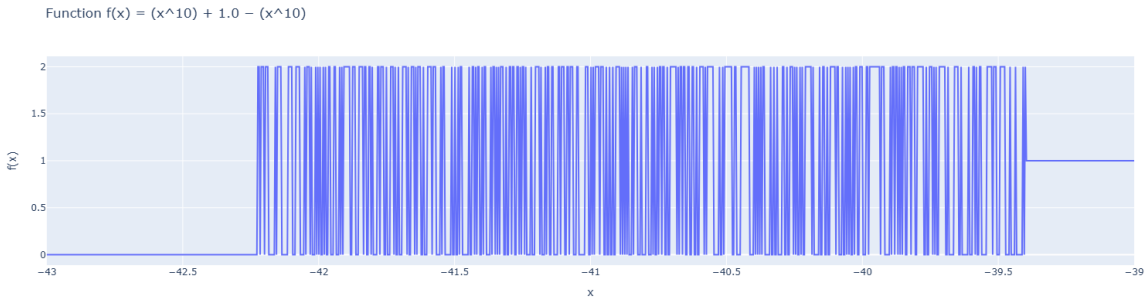

Figura 1.2: Comportamento da expressão $x^{10} + 1 - x^{10}$ no intervalo $[-43, -39]$.

Podemos observar que em um determinado limiar, a função para de se comportar como esperado $f(x) = 1$ e passa a assumir valores os de $f(x) = 2$ e $f(x) = 0$.

Vamos manipular essa expressão e ver como ela se comporta em diferentes situações.

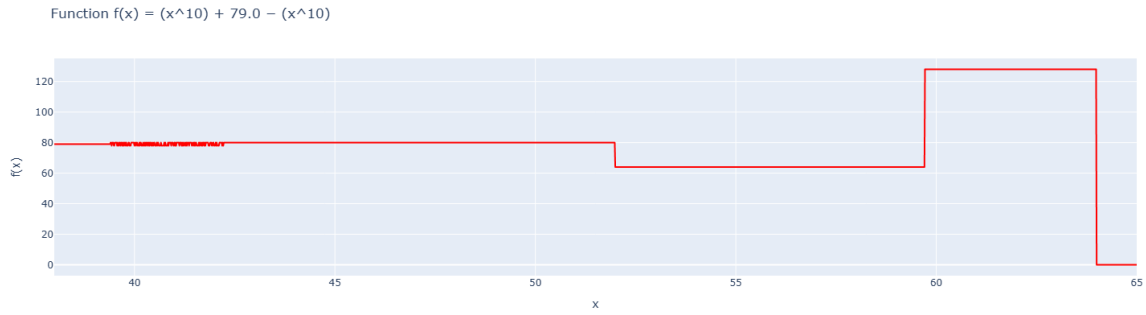

Figura 1.3: Comportamento da expressão $x^{10} + 79 - x^{10}$ no intervalo $[38, 65]$.

Na figura 1.3. é possível observar que a função assume mais de dois valores inesperados para $f(x)$, nesse caso, o conjunto imagem no intervalo $[38, 65]$ é $Im(f) = \{0{,}64{,}78{,}79{,}80{,}128\}$. O seguinte conjunto de valores para o literal foi testado $\{3, 12, 79, 98\}$, e acreditamos que para valores que não podem ser escritos como uma potência de base 2, esse padrão ocorra.

TODO: adicionar explicação

Explicação: blablabla

Uma dúvida comum é se a precisão desses sistemas afeta nessa expressão. Vamos comparar então um sistema de float32 (Cerca de 7-8 bits dígitos decimais de precisão) a um sistema float64 (Cerca de 15-16 dígitos decimais de precisão)

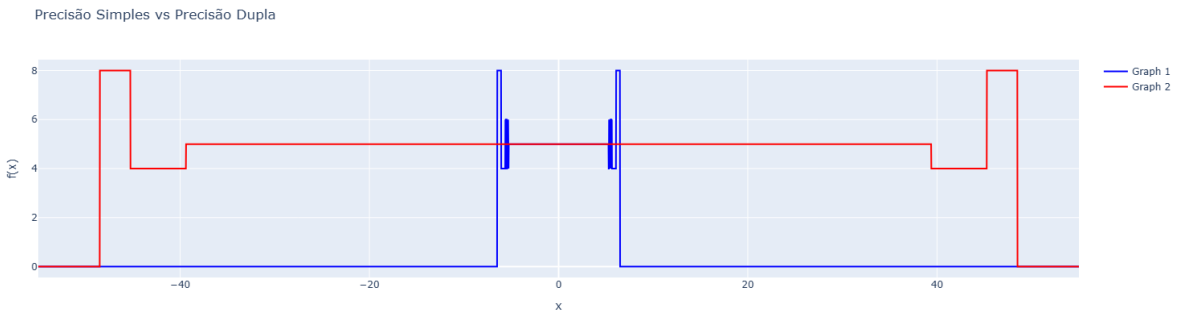

Figura 1.4: Comportamento da expressão $x^{10} + 2 - x^{10}$ no intervalo $[-40, 40]$.

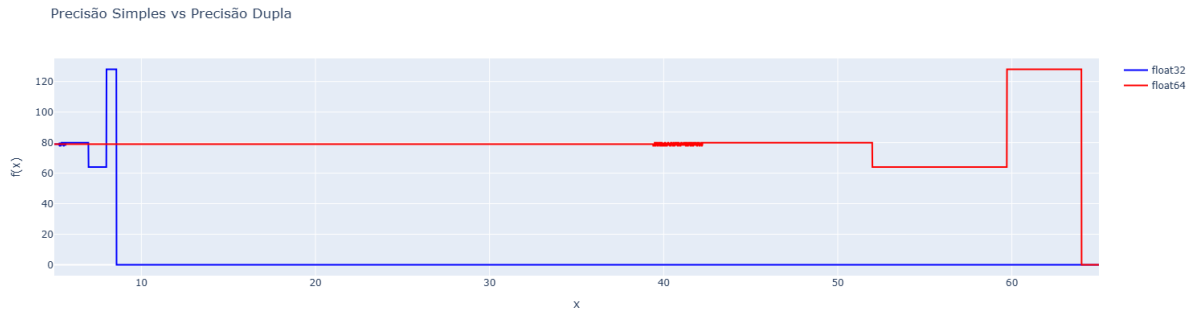

Figura 1.5: Comportamento da expressão $x^{10} + 79 - x^{10}$ no intervalo $[0, 60]$.

Dos testes experimentais, exemplificado nas figuras 1.4 e 1.5, foi concluído que o $|x|$ onde a instabilidade ocorre é menor é menor na precisão dupla.

Outro caso interessante é quando analisamos a função $f(x) = (x \times x \times x \times x \times x \times x \times x \times x \times x \times x) - x^{10}$. Vejamos o gráfico dessa função.

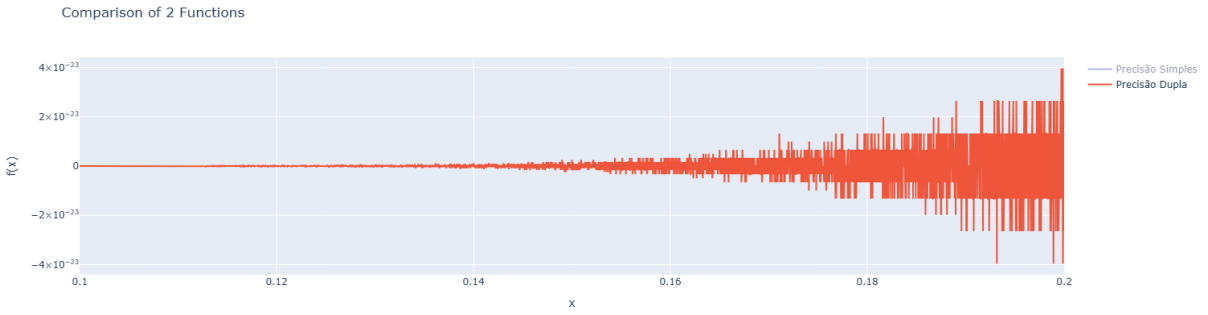

Figura 1.6: Comportamento da expressão $f(x) = (x \times x \times x \times x \times x \times x \times x \times x \times x \times x) - x^{10}$, com precisão dupla, no intervalo $[0{,}1, 0{,}2]$.

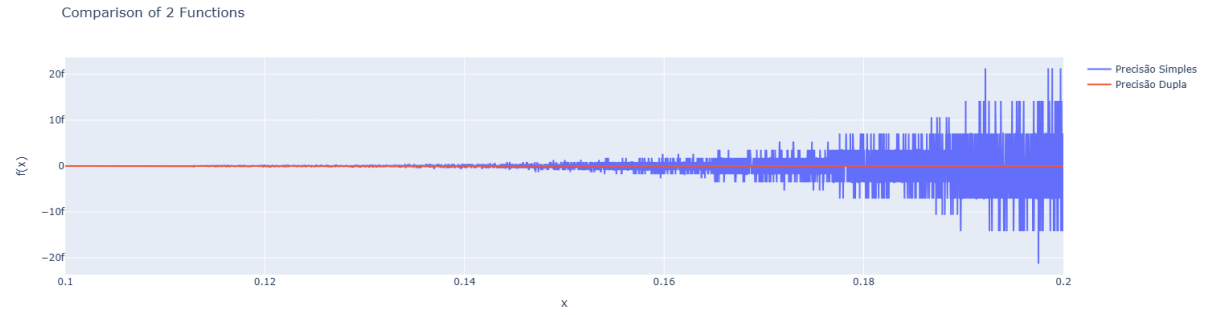

Figura 1.7: Comportamento da expressão $f(x) = (x \times x \times x \times x \times x \times x \times x \times x \times x \times x) - x^{10}$, com precisão simples, no intervalo $[0{,}1, 0{,}2]$.

A instabilidade na precisão dupla (Figura 1.7) produz imagem em torno de $[-4 \times 10^{-23}, 2 \times 10^{-23}]$, enquanto na precisão simples (Figura 1.6) está em torno

de $[-20^{-15}, 20^{-15}]$ ($[-20femto, 20femto]$). Observa-se, portanto, que a diferença de instabilidade é extremamente menor na precisão dupla em comparação com a simples.

Uma outra situação é a diferença de como o computador interpreta algumas funções se forem reescritas de maneiras diferentes. Seja $p(x) = (x-1)^6$ e $q(x) = x^6 - 6x^5 + 15x^4 - 20x^3 + 15x^2 - 6x + 1$, analiticamente essas funções são idênticas, porém existem problemas de cancelamento catastrófico na hora de analisarmos ambas em um sistema de ponto flutuante.

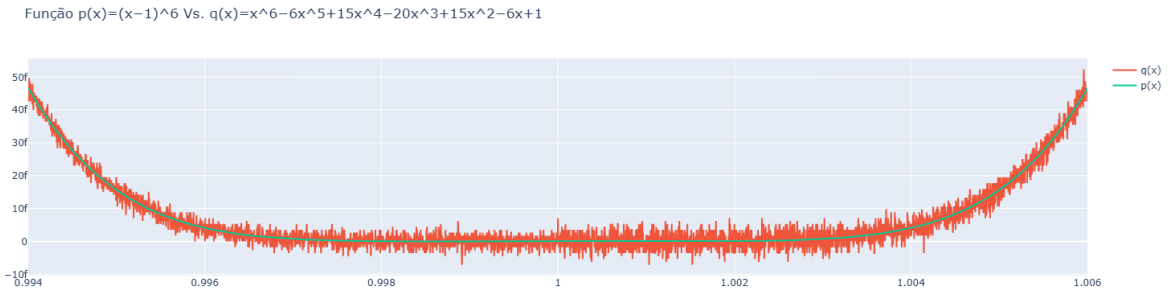

Figura 1.8: Comparação entre as funções $p(x) = (x-1)^6$ e $q(x) = x^6 - 6x^5 + 15x^4 - 20x^3 + 15x^2 - 6x + 1$.

TODO: adicionar explicação

Explicação: yada yada yada

1.4.2 Discussão

Esses exemplos destacam a importância da ordem das operações e da análise cuidadosa ao trabalhar com algoritmos numéricos. Técnicas como a reordenação de cálculos e o uso de formatos de precisão estendida podem ajudar a minimizar esses erros em contextos críticos.

Capítulo 2

Métodos Iterativos para Zeros de Função

Em muitas aplicações, as soluções buscadas se resumem a encontrar os zeros (ou raízes) de uma função. Entretanto, nem sempre é possível fazê-lo analiticamente, devido à natureza das componentes envolvidas na função como, por exemplo, funções polinomiais a partir do 3^{o} grau, somas de funções trigonométricas e logarítmicas, entre outras. Nesse ínterim, recorremos então a maneiras de obter valores aproximados para tais raízes.

@EnzoR: acho que devemos adiconar algo falando qual é o critério de parada.

Uma classe de métodos utilizados para aproximar raízes de funções são os **métodos iterativos**. A essência desses métodos está em, partindo de um chute inicial e de uma função apropriada φ, obter uma sequência x_k onde cada termo é obtido do anterior recursivamente como $x_{k+1} = \varphi(x_k)$. Essa sequência, sob certas hipóteses, converge para a raiz ξ da função.

Ao longo do capítulo, reservaremos o símbolo ξ para representar raízes de funções.

2.1 Localização de Raízes

Nos métodos que trataremos nesse capítulo, para garantir a convergência da sequência iterativa, é necessário que o primeiro termo esteja suficientemente próximo da raiz e, desse modo, faz-se necessário restringir as funções a intervalos que contenham raízes. Quando as funções envolvidas são contínuas, o resultado a seguir garante a existência de raízes em um intervalo [a,b] desde que as imagens dos extremos tenham sinais opostos.

Proposição 2.1.1. *Seja $f(x)$ uma função contínua no intervalo $[a, b]$. Se $f(a)f(b) < 0$, então há pelo menos uma raiz $\xi \in (a, b)$. Se, além disso, existir $f'(x)$ e $f'(x)$ preservar o sinal em (a, b), então a raiz é única.*

Por exemplo, considere a função $f(x) = x^3 - 9x + 3$. Utilizando a Proposição 2.1.1, observamos que

- $f(0)f(1) = -15 < 0$, portanto há raiz no intervalo (0, 1);
- $f(2)f(3) = -21 < 0$, portanto há raiz no intervalo (2, 3).

@EnzoR: acho que vale a pena citar quem é a, b, c e d antes, não?

Além disso, a derivada de $f(x)$ é $f'(x) = 3x^2 - 9$, cujas raízes são $\pm\sqrt{3}$, então nos intervalos (-∞, -$\sqrt{3}$), (-$\sqrt{3}$, $\sqrt{3}$) e ($\sqrt{3}$, ∞) o sinal da derivada é preservado. Portanto, nos intervalos (a, b) e (c, d) há apenas uma raiz.

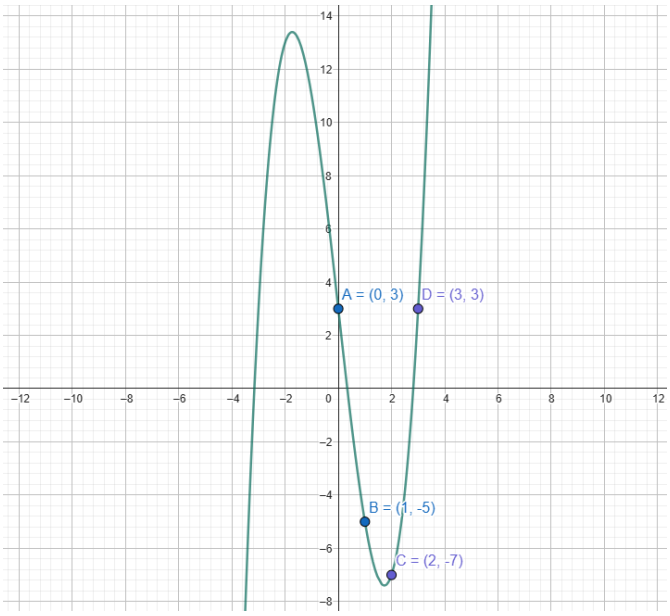

Figura 2.1: Raiz entre A e B, e entre C e D

Outra forma de localizar raízes de uma dada função $f(x)$ é escrevê-la como a diferença entre as funções $g(x) - h(x)$, pois se $f(\xi) = 0$ temos que $g(\xi) - h(\xi) = 0$ ou, equivalentemente, $g(\xi) = h(\xi)$. Graficamente, ξ é a abscissa do ponto de interseção entre as funções g(x) e h(x).

Da mesma função, podemos obter, por exemplo, as interseções entre x e $x^3 - 8x + 3$

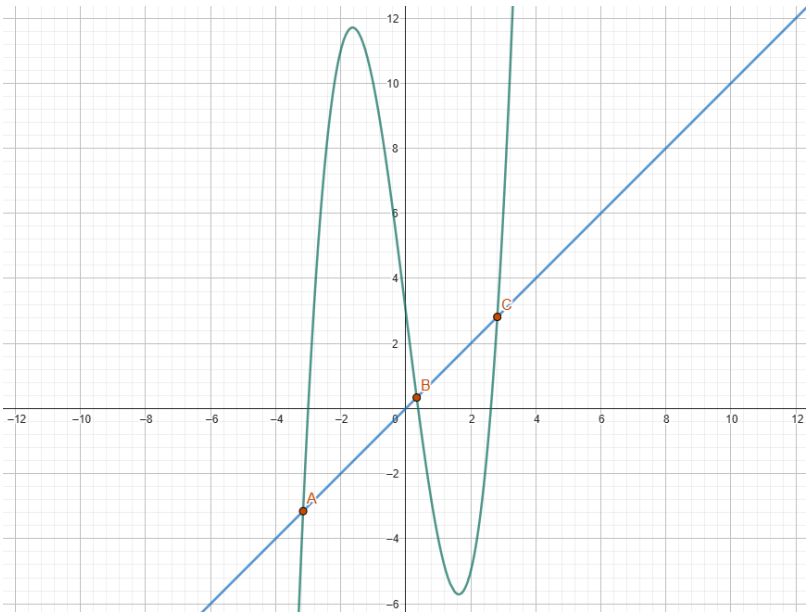

Figura 2.2: Interseções entre x e $x^3 - 8x + 3$

2.2 Critério de Parada

FIXME: Completar

O processo é repetido até que a diferença entre duas iterações consecutivas seja inferior a uma tolerância pré-estabelecida ou até que um número máximo de iterações seja atingido.

- $e_k = x_k - \xi$
- $e_k = f(x_k) - 0$

@LucasM: Definir raiz aproximada

2.3 Método do Ponto Fixo

O método do ponto fixo *é um método iterativo* que transforma o problema de buscar as raízes de uma função $f(x)$ no problema de encontrar os pontos fixos de uma outra função $\varphi(x)$, denominada de **função de iteração de ponto fixo**. A partir dessa função de iteração, uma sequência é construída recursivamente começando em um valor inicial x_0 que convergirá para a raiz ξ de $f(x)$, desde que sejam observadas certas condições sob a função $\varphi(x)$ e o dado inicial x_0.

O primeiro passo é gerar funções de iteração φ para $f(x)$, o que pode ser feito isolando x na equação $f(x) = 0$. Por exemplo, manipulando a função $x^3 - 9x + 3$ da seguinte forma

$$x^3 - 8x + 3 = x$$

obtemos a função de iteração $\varphi(x) = x^3 - 8x + 3$. Com a mesma lógica, outras possíveis funções de iteração para f são

a) $\varphi_1(x) = \dfrac{x^3}{9} + \dfrac{1}{3}$

b) $\varphi_2(x) = \sqrt[3]{9x - 3}$

c) $\varphi_3(x) = \dfrac{9}{x} - \dfrac{3}{x^2}$

d) $\varphi_4(x) = \sqrt{9 - \dfrac{3}{x}}$

e) $\varphi_5(x) = -\sqrt{9 - \dfrac{3}{x}}$

f) $\varphi_6(x) = x^3 - 8x + 3$

A forma geral da função de iteração é

$$\varphi(x) = x + A(x)f(x) \tag{2.1}$$

com $A(\xi) \neq 0$. Por exemplo, a $\varphi_1 = \dfrac{x^3}{9} + \dfrac{1}{3}$ na forma geral ficaria

$$\varphi_1(x) = x + \frac{1}{9}f(x)$$

em que $A(x) = \dfrac{1}{9}$. Nesse caso, pode-se observar que $\mathrm{A}(\xi) \neq 0$.

O resultado a seguir relaciona a raiz de uma função com pontos fixos de uma função de iteração associada a essa função.

Proposição 2.3.1. *Seja ξ uma raiz de uma função $f(x)$ e seja $\varphi(x)$ uma função de iteração associada a $f(x)$. Então, $f(\xi) = 0$ se, e somente se, $\varphi(\xi) = \xi$.*

Demonstração. ($\Rightarrow$) Pela forma geral da função de iteração temos que $\varphi(\xi) = \xi + A(\xi)f(\xi)$. Uma vez que $f(\xi) = 0$, então $\varphi(\xi) = \xi$.

($\Leftarrow$) Começando novamente pela forma geral da função de iteração, temos que $\varphi(\xi) = \xi + A(\xi)f(\xi)$. Como $\varphi(\xi) = \xi$, concluímos que $A(\xi)f(\xi) = 0$. Tendo como hipótese que $A(\xi) \neq 0$, então $f(\xi) = 0$. □

@LucasM: Aqui, antes de ir para o resultado principal, dar exemplos de funções de iteração que fazem a sequencia convergir e divergir. Inserir gráficos assim como no livro da Vera.

@DanielP: ok!

Sob condições a respeito da função de iteração, sua derivada e o dado inicial, a convergência da sequência iterativa é garantida, como pode-se observar a seguir.

@EnzoR: Vale a pena dar nome aos teoremas? Por exemplo: Ao invés de "Teorema 2.3.1"ser "Teorema 2.3.1 - Convergência de Funções de Iteração"ou algo assim. Acho que se fizermos isso, fica melhor de referenciar depois, pois assim o leitor não precisa ficar procurando o teorema pelo número.

Teorema 2.3.1. *Seja ξ uma raiz de f(x), isolada num intervalo I centrado nessa raiz. Considere uma função de iteração $\varphi(x)$ associada a f(x). Sob as seguintes hipóteses:*

i) $\varphi(x)$ *e* $\varphi'(x)$ *são contínuas em I,*

ii) $|\varphi'(x)| \leq M < 1$ *em I,*

iii) $x_0 \in I$*,*

a sequência $x_{k+1} = \varphi(x_k)$ *converge para a raiz* ξ*.*

Demonstração. Como $x_{k+1} = \varphi(x_k)$, subtraindo ξ de ambos os lados da igualdade e usando o fato de que $\varphi(\xi) = \xi$ temos

$$x_{k+1} - \xi = \varphi(x_k) - \varphi(\xi). \tag{2.2}$$

Pelo Teorema do Valor Médio (TVM) podemos escrever

$$\varphi(x_k) - \varphi(\xi) = \varphi'(c_k)(x_k - \xi) \tag{2.3}$$

com c_k entre x_k e ξ. Então, substituindo (2.3) em (2.2), temos

$$\begin{aligned} |x_{k+1} - \xi| &= |(x_k - \xi)\ \varphi'(c_k)| \\ &= |x_k - \xi|\ |\varphi'(c_k)| \\ &< |x_k - \xi| \end{aligned} \tag{2.4}$$

uma vez que $|\varphi'(x)| < 1$. Como $x_0 \in I$, podemos concluir que $x_k \in I$ para todo k já que, por (2.4), $|x_k - \xi| < |x_0 - \xi|$.

Na sequência, provaremos que x_k converge para a raiz ξ. Vamos começar mostrando que

$$|x_1 - \xi| \leq M\ |x_0 - \xi|. \tag{2.5}$$

Observe que, como $x_1 = \varphi(x_0)$, temos que $x_1 - \xi = \varphi(x_0) - \varphi(\xi)$. Pelo Teorema do Valor Médio temos que $\varphi(x_0) - \varphi(\xi) = (x_0 - \xi)\ \varphi'(c_0)$, para algum c_0 entre x_0 e ξ. Uma vez que $|\varphi'(x)| \leq M$ no intervalo I, a seguinte desigualdade é válida

$$\begin{aligned} |x_1 - \xi| &= |x_0 - \xi|\ |\varphi'(c_0)| \\ &\leq M\ |x_0 - \xi| \end{aligned}$$

e provamos a desigualdade (2.5). De modo similar prova-se que $|x_2 - \xi| \leq M\ |x_1 - \xi|$ que, combinado com (2.5), implica que $|x_2 - \xi| \leq M^2\ |x_0 - \xi|$. Repetindo o processo k vezes pode-se concluir que

$$|x_k - \xi| \leq M^k\ |x_0 - \xi|. \tag{2.6}$$

Como $0 < M < 1$, se k tende a infinito, M^k tende a 0 e, portanto, $M^k\ |x_0 - \xi|$ também tende a 0. Assim, provamos que

$$\lim_{k \to \infty} x_k = \xi, \tag{2.7}$$

ou seja, x_k converge para a raiz. □

2.3.1 Ordem de convergência

@LucasM: Dizer aqui o que significa na prática a ordem de convergência de um método. Perceba que pela definição ele está ligado ao erro cometido no processo iterativo.

A ordem de convergência de um método iterativo é uma medida de quão rapidamente a sequência de iterações converge para a solução desejada. Em termos práticos, isso significa que, se um método tem uma ordem de convergência alta, ele será capaz de reduzir o erro de aproximação de forma mais eficaz a cada iteração.

Definição 2.3.1. Seja $\{x_k\}$ uma sequência que converge para ξ e $e_k = x_k - \xi$ o erro na k-ésima iteração. Se existirem $p > 1$ e $C > 0$ tais que $\lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^p} = C$, então p é chamada de *ordem de convergência* da sequência e C é a *constante assintótica de erro.*

$$|e_{k+1}| \approx C|e_k|^p$$

Proposição 2.3.2. *Se*

$$\lim_{k\to\infty} \frac{e_{k+1}}{e_k} = C,\ 0 \leq |C| < 1, \tag{2.8}$$

então a convergência é pelo menos linear. O MPF tem convergência pelo menos linear.

Demonstração. Partindo de (2.2) e (2.3), e tomando o limite com k tendendo a infinito, podemos escrever (2.8) como

$$\lim_{k\to\infty} \frac{x_{k+1} - \xi}{x_k - \xi} = \lim_{k\to\infty} \varphi'(c_k)$$

com c_k entre x_k e ξ . Como por hipótese $\varphi'(x) < 1$ então $|C| < 1$, portanto, a convergência do MPF é pelo menos linear. □

2.4 Método de Newton-Raphson

TODO: talvez reescrever todas as condições como i) derivada limitada ii) f e g continuas , etc...

@DanielP: conv quadrática

@EnzoR: escolhendo a func de iter

O método de Newton é um caso particular do **Método do Ponto Fixo** amplamente utilizado para encontrar raízes

@EnzoR: seria raízes complexas, não?

reais de funções não lineares, cuja ordem de convergência é pelo menos quadrática. A função de iteração específica para este método produz uma sequência em que cada termo x_{k+1} corresponde, geometricamente, à **interseção da reta tangente** a $f(x)$ no ponto $(x_k, f(x_k))$ com o eixo x.

@DanielP: só tem que colocar no lugar certo

No método do ponto fixo, vimos que se a função satisfaz o Teorema 2.3.1 a sequência $x_{k+1} = \varphi(x_k)$ converge para ξ. Além disso, a desigualdade (2.6) na demonstração desse resultado nos diz que quanto menor for $|\varphi'(x)|$, mais rápido a sequência $\{x_k\}$ converge.

Dada uma função $f(x)$, o processo parte de uma estimativa inicial x_0 e aplica a seguinte fórmula iterativa, a partir da escolha da derivada em (2.1)

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \tag{2.9}$$

onde $f'(x_k)$ é a derivada da função f avaliada em x_k. Para que o método convirja para a raiz correta, é necessário que $f(x)$ seja continuamente diferenciável em uma vizinhança da raiz e que $f'(x) \neq 0$ nessa região. Além disso, a escolha adequada do ponto inicial x_0 é crucial para garantir a convergência do método.

Aplicando a derivada na forma geral de φ pela regra da cadeia obtemos $\varphi'(x) = 1+A'(x)f(x)+A(x)f'(x)$ e, calculando-a na raiz, resta $\varphi'(\xi) = 1+A(\xi)f'(\xi)$. Impomos que $\varphi'(\xi) = 0$, o que nos leva a $A(\xi) = \dfrac{-1}{f'(\xi)}$. Generalizando, temos $A(x) = \dfrac{-1}{f'(x)}$. Portanto, desde que $f'(x) \neq 0$, a forma da função de iteração do método de Newton-Raphson é

$$\varphi(x) = x - \frac{f(x)}{f'(x)}.$$

2.4.1 Demonstração Geométrica

2.4.2 Convergência

Teorema 2.4.1. *Sejam $f(x)$, $f'(x)$ e $f''(x)$ contínuas num intervalo I que contém a raiz ξ de f, supondo $f'(\xi) \neq 0$. Então, existe um intervalo $\overline{I} \subset I$, contendo a raiz ξ, tal que $x_0 \in \overline{I}$, a sequência x_k gerada pela função de iteração $\varphi(x) = x - \dfrac{f(x_k)}{f'(x_k)}$ convergirá para a raiz.*

Demonstração. Sendo o método de Newton-Raphson um caso particular do MPF, basta provar que para $\varphi(x) = x - \dfrac{f(x)}{f'(x)}$ as hipóteses do Teorema 2.3.1 são satisfeitas.

Primeiramente, observe que $\varphi'(x) = \dfrac{f(x)f''(x)}{[f'(x)]^2}$. Como $f'(\xi) \neq 0$ e $f'(x)$ é contínua em I, é possível obter $I_1 \subset I$ tal que $f'(x) \neq 0$ no intervalo I_1. Assim, a função f e suas derivadas primeira e segunda são contínuas em I_1 e, consequentemente, a função de iteração e sua derivada também.

Uma vez que a $\varphi'(x)$ é contínua em I_1 e $\varphi'(\xi) = 0$, é possível escolher $I_2 \subset I_1$ de modo que $|\varphi'(x)| < 1$ em I_2 tendo ξ como centro do novo intervalo.

Por fim, tomando, $\overline{I} = I_2$, satisfazem-se as hipóteses do Teorema 2.3.1. □

2.4.3 Ordem de Convergência

No MPF espera-se uma ordem de convergência ao menos linear, entretanto ao escolher uma função de iteração que satisfaça $\varphi'(\xi) = 0$, provaremos que sua ordem de convergência será ao menos quadrática.

Proposição 2.4.1.

@LucasM: Antes em Linguagem natural

A ordem de convergência do método de Newton é pelo menos quadrática.

Demonstração.

BUG: Ele está chamando de "Teorema 2.4.2", mas está se referenciando a uma demonstração.

Suporemos todas as hipóteses do Teorema 2.4.2. Partindo de $x_{k+1} = x_k - \dfrac{f(x_k)}{f'(x_k)}$, subtraindo ξ em ambos os lados da igualdade, obtemos

$$e_{k+1} = e_k - \frac{f(x_k)}{f'(x_k)} \tag{2.10}$$

O polinômio de Taylor de grau 2 para $f(x)$ centrado em x_k é

$$f(x) = f(x_k) + f'(x_k)(x - x_k) + \frac{f''(c_k)}{2}(x - x_k)^2$$

com c_k entre x e x_k. Assim, $f(\xi) = f(x_k) - f'(x_k)(x_k - \xi) + \frac{f''(c_k)}{2}(x_k - \xi)^2$, tomando $x = \xi$, dado que $f(\xi) = 0$, dividindo a equação pela derivada de f temos

$$\begin{aligned}
\frac{f''(c_k)}{2f'(x_k)}e_k^2 &= e_k - \frac{f(x_k)}{f'(x_k)} \\
\frac{f''(c_k)}{2f'(x_k)}e_k^2 &= e_{k+1} \\
\frac{e_{k+1}}{e_k^2} &= \frac{1}{2}\frac{f''(c_k)}{f'(x_k)}
\end{aligned}$$

$$\begin{aligned}
\lim_{k\to\infty} \frac{e_{k+1}}{e_k^2} &= \lim_{k\to\infty} \frac{1}{2}\frac{f''(c_k)}{f'(x_k)} \\
&= \frac{1}{2}\frac{f''[\lim_{k\to\infty}(c_k)]}{f'[\lim_{k\to\infty}(x_k)]} \\
&= \frac{1}{2}\frac{f''(\xi)}{f'(\xi)} \\
&= \frac{1}{2}\varphi''(\xi) \\
&= C
\end{aligned}$$

□

2.4.4 Ciladas

O Método de Newton é amplamente utilizado na prática devido à sua rapidez e precisão em condições ideais. Contudo, em algumas situações ele pode falhar ou convergir para raízes incorretas se essas condições não forem satisfeitas. Estes problemas geralmente estão associados a fatores como: pontos de máximo e mínimo, pontos de inflexão, multiplicidade da raiz e escolha inadequada de x_0.

2.4.5 Fractais

Referências Citadas

Como exemplo, citamos [1].

Referências Bibliográficas

[1] Márcia Aparecida Gomes Ruggiero and Vera Lúcia da Rocha Lopes. *Cálculo numérico: aspectos teóricos e computacionais*. Pearson/Makron, Sao Paulo, 1998.