CÁLCULO NUMÉRICO

LICENCIATURA EM MATEMÁTICA

CAPES

UAB
UNIVERSIDADE ABERTA DO BRASIL

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA CEARÁ

Ministério da Educação - MEC

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

Universidade Aberta do Brasil

Instituto Federal de Educação, Ciência e Tecnologia do Ceará

MINISTÉRIO DA EDUCAÇÃO
Universidade Aberta do Brasil
Instituto Federal de Educação, Ciência e Tecnologia do Ceará
Diretoria de Educação a Distância

Licenciatura em Matemática

Cálculo Numérico

Francisco Gêvane Muniz Cunha
Jânio Kléo de Sousa Castro

Fortaleza, CE
2010

CRÉDITOS

Presidente
Luiz Inácio Lula da Silva
Ministro da Educação
Fernando Haddad
Secretário da SEED
Carlos Eduardo Bielschowsky
Diretor de Educação a Distância
Celso Costa
Reitor do IFCE
Cláudio Ricardo Gomes de Lima
Pró-Reitor de Ensino
Gilmar Lopes Ribeiro
Diretora de EAD/IFCE e Coordenadora UAB/IFCE
Cassandra Ribeiro Joye
Vice-Coordenadora UAB
Régia Talina Silva Araújo
Coordenador do Curso de Tecnologia em Hotelaria
José Solon Sales e Silva
Coordenador do Curso de Licenciatura em Matemática
Zelalber Gondim Guimarães
Elaboração do conteúdo
Francisco Gêvane Muniz Cunha
Jânio Kléo de Sousa Castro
Colaborador
Marilia Maia Moreira
Equipe Pedagógica e Design Instrucional
Ana Claúdia Uchôa Araújo
Andréa Maria Rocha Rodrigues
Carla Anaíle Moreira de Oliveira
Cristiane Borges Braga
Eliana Moreira de Oliveira
Gina Maria Porto de Aguiar Vieira
Giselle Santiago Cabral Raulino
Glória Monteiro Macedo
Iraci Moraes Schmidlin
Jane Fontes Guedes
Karine Nascimento Portela
Lívia Maria de Lima Santiago
Lourdes Losane Rocha de Sousa
Luciana Andrade Rodrigues
Maria Irene Silva de Moura
Maria Vanda Silvino da Silva
Marília Maia Moreira
Saskia Natália Brígido Bastista

Equipe Arte, Criação e Produção Visual
Ábner Di Cavalcanti Medeiros
Benghson da Silveira Dantas
Davi Jucimon Monteiro
Diemano Bruno Lima Nóbrega
Germano José Barros Pinheiro
Gilvandenys Leite Sales Júnior
José Albério Beserra
José Stelio Sampaio Bastos Neto
Larissa Miranda Cunha
Marco Augusto M. Oliveira Júnior
Navar de Medeiros Mendonça e Nascimento
Roland Gabriel Nogueira Molina
Samuel da Silva Bezerra
Equipe Web
Aline Mariana Bispo de Lima
Benghson da Silveira Dantas
Fabrice Marc Joye
Igor Flávio Simões de Sousa
Luiz Bezerra de Andrade FIlho
Lucas do Amaral Saboya
Ricardo Werlang
Samantha Onofre Lóssio
Tibério Bezerra Soares
Thuan Saraiva Nabuco
Samuel Lima de Mesquita
Revisão Textual
Aurea Suely Zavam
Nukácia Meyre Araújo de Almeida
Revisão Web
Antônio Carlos Marques Júnior
Débora Liberato Arruda Hissa
Saulo Garcia
Logística
Francisco Roberto Dias de Aguiar
Virgínia Ferreira Moreira
Secretários
Breno Giovanni Silva Araújo
Francisca Venâncio da Silva
Auxiliar
Ana Paula Gomes Correia
Bernardo Matias de Carvalho
Isabella Britto
Maria Tatiana Gomes da Silva
Raíssa Miranda de Abreu Cunha
Wagner Souto Fernandes
Zuila Sâmea Vieira de Araújo

Catalogação na Fonte: Islânia Fernandes Araújo (CRB 3 – Nº917 615)

C972c Cunha, Francisco Gêvane Muniz

Cálculo numérico / Francisco Gêvane Muniz Cunha, Jânio Kléo Sousa de Castro; Coordenação Cassandra Ribeiro Joye. - Fortaleza: UAB/IFCE, 2010.

162p. : il. ; 27cm.

ISBN 978-85-475-0012-2

1. MATEMÁTICA - CÁLCULO 2. REPRESENTAÇÃO DOS NÚMEROS. 3. MÉTODOS NUMÉRICOS I. Castro, Jânio Kléo Sousa de. II. Joye, Cassandra Ribeiro. (Coord.) III. Instituto Federal de Educação, Ciência e Tecnologia do Ceará – IFCE IV. Universidade Aberta do Brasil V. Título

CDD – 519.40785

SUMÁRIO

APRESENTAÇÃO

Caro(a) aluno(a),

Seja bem-vindo a nossa disciplina de cálculo numérico, cujo objetivo central é estudar técnicas (ou métodos) numéricas para obter soluções de problemas que possam ser representados por modelos matemáticos. Assim, ganhamos uma importante ferramenta para a resolução de problemas oriundos da própria matemática, ou de outras áreas, estabelecendo um elo entre matemática e problemas práticos de áreas específicas.

Devemos destacar que a resolução de modelos matemáticos é muitas vezes complexa, envolvendo fenômenos não-lineares, podendo tornar impossível a descoberta analítica de soluções. Nestes casos, os métodos numéricos são ferramentas imprescindíveis a aproximação das soluções. Portanto, o cálculo numérico é fundamental na formação de profissionais das áreas de ciências exatas e engenharias.

Esperamos que você, caro(a) aluno(a), adquira habilidades para: compreender como os números são representados nas calculadoras e computadores e como são realizadas as operações nestes sistemas; conhecer e aplicar os principais métodos numéricos para a solução de certos problemas; estimar e analisar os erros obtidos; e propor soluções para minimizá-los ou mesmo, quando possível, eliminá-los.

A sua participação nas atividades e em cada aula será essencial para que você possa tirar o maior proveito da disciplina. Agradeceremos quaisquer contribuições no sentido de melhorar o nosso texto, estando à disposição para maiores esclarecimentos

Desejamos um bom curso a todos!

Gêvane Cunha e Jânio Kléo.

AULA 1

Representando números e calculando erros

Olá! Iniciaremos aqui os nossos estudos sobre o Cálculo Numérico. Nesta primeira aula, apresentamos uma breve visão sobre a disciplina, destacando, de modo geral, os conteúdos que serão abordados e procurando mostrar a importância dessa ferramenta para a resolução de diversos problemas que surgem, principalmente das ciências exatas e engenharias.

Nesta aula, trataremos ainda das formas de representação dos números em sistemas de numeração, enfatizando a representação em ponto flutuante, comumente adotada em sistemas digitais como calculadoras e computadores. Apresentaremos também noções de erro e de aproximação numérica, fundamentais para o trabalho com as técnicas do cálculo numérico.

Objetivos

- Formular uma visão geral do cálculo numérico
- Estabelecer, em linhas gerais, os conteúdos que serão abordados na disciplina
- Estudar noções de erro e de aproximação numérica
- Conhecer formas de representação numérica

TÓPICO 1

Cálculo numérico: Por que e para quê?

OBJETIVOS

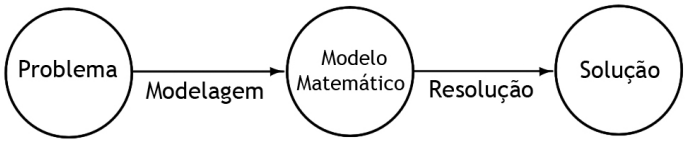
- Reconhecer a importância do cálculo numérico
- Conhecer princípios básicos usados em cálculo numérico
- Reconhecer problemas que podem ser resolvidos por cálculo numérico
- Estabelecer fases para a resolução de problemas reais

Neste tópico, estabelecemos as bases gerais para o nosso trabalho na disciplina, apontando os conteúdos que serão trabalhados. Com isso, estaremos realçando a importância do cálculo numérico e a sua utilidade como ferramenta para a resolução de problemas reais oriundos da própria Matemática, de outras ciências exatas e das engenharias.

Grande parte dos problemas matemáticos surge da necessidade de solucionar problemas da natureza, sendo que é possível descrever muitos fenômenos naturais por meio de modelos matemáticos (HUMES *et. al,* 1984). De acordo com Ohse (2005, p. 1):

> Desde que o homem começou a observar os fenômenos naturais e verificar que os mesmos seguiam princípios constantes, ele observou que estes fenômenos podiam ser colocados por meio de "fórmulas". Este princípio levou a utilização da matemática como uma ferramenta para auxiliar estas observações. Este é o princípio da matemática como um modelo, ou seja, modelar matematicamente o mundo em que vivemos e suas leis naturais.

A figura 1 apresenta, de forma sucinta, as etapas para solucionar um problema da natureza.



Fonte: Humes et. al (1984, p. 1)

Figura 1: Etapas para solucionar um problema da natureza.

O esquema da figura 1 mostra duas etapas fundamentais para a solução de um problema:

1. Modelagem do problema: etapa inicial que consiste na representação do problema por um modelo matemático conveniente. Em geral, o modelo é obtido a partir de teorias das área específicas que originaram o problema e, com vistas a tornar o modelo um problema matemático resolvível, podem conter simplificações do problema real. Dependendo da abordagem dada ao problema, é mesmo possível obtermos modelos matemáticos diferentes.
2. Resolução do modelo: etapa em que buscamos encontrar uma solução para o modelo matemático obtido na fase de modelagem. É nesta fase que necessitamos de métodos numéricos específicos para resolver o modelo correspondente.

A ideia de modelo matemático tem sido discutida por vários autores. Uma boa definição para a expressão modelo matemático é a de Biembengut e Hein (2000, p. 12), segundo a qual "um conjunto de símbolos e relações matemáticas que traduz, de alguma forma, um fenômeno em questão ou um problema de situação real, é denominado de modelo matemático".

Os métodos utilizados na resolução dos modelos matemáticos de problemas, nos vários ramos das engenharias ou ciências aplicadas, baseiam-se, atualmente, em uma de duas categorias: *métodos analíticos* e *métodos numéricos* .

ATENÇÃO!

Entendemos por método analítico aquele que, a menos de erros de arredondamentos, fornece as soluções exatas do problema real. Em geral, tais soluções são obtidas a partir de fórmulas explícitas. Por outro lado, um método numérico é constituído por uma sequência finita de operações aritméticas que, sob certas condições, levam a uma solução ou a uma aproximação de uma solução do problema.

Sempre que possível, e em especial quando desejamos exatidão na solução do problema, é preferível a utilização dos métodos analíticos na resolução dos modelos matemáticos. Tais métodos têm a vantagem de fornecer informações gerais em vez de particularizadas, além de uma maior informação quanto à natureza e à dependência das funções envolvidas no modelo.

No entanto, a resolução de modelos matemáticos obtidos na modelagem de problemas reais de diversas áreas é muitas vezes complexa e envolve fenômenos não-lineares, podendo tornar impossível

a descoberta de uma solução analítica para o problema dado. Nestes casos, e/ou quando for possível aceitar soluções aproximadas para os problemas reais, os métodos numéricos são ferramentas importantes para sua solução.

Para compreender melhor e diferenciar os métodos analíticos dos métodos numéricos, vejamos agora dois exemplos simples característicos.

GUARDE BEM ISSO!

Em um método numérico, uma solução aproximada é, em geral, obtida de forma construtiva: partindo de aproximações iniciais, vão sendo construídas novas aproximações até que uma aproximação considerada "boa" seja obtida. Desse modo, um método numérico pode ser escrito em forma de algoritmo com as operações (ou grupos de operações), podendo ser executadas repetidamente.

Exemplo 1:

Um método analítico para determinar (quando existem) os zeros reais de uma função quadrática

$f(x) = ax^2 + bx + c$, com $a \neq 0$

é dado pela fórmula de Bhaskara, a saber:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Desse modo, os zeros reais de $f(x) = x^2 - 5x + 6$ são

$$x_1 = \frac{-(-5) - \sqrt{(-5)^2 - 4 \times 1 \times 6}}{2 \times 1} = 2 \text{ e } x_2 = \frac{-(-5) + \sqrt{(-5)^2 - 4 \times 1 \times 6}}{2 \times 1} = 3$$

Exemplo 2:

Um método numérico para determinar uma aproximação para a raiz quadrada de um número real p, maior que 1, é o algoritmo de Eudoxo:

VOCÊ SABIA?

Eudoxo de Cnidos astrônomo, matemático e filósofo grego que viveu de 408 a.C a 355 a.C. Cnidos, onde nasceu, corresponde hoje à Turquia.

Do fato que $p > 1$, temos que $1 < \sqrt{p} < p$.

Escolhe-se, como uma primeira aproximação para $\sqrt{p}$, $x_0 = (1 + p)/2$, ou seja, a média aritmética entre 1 e p. Pode-se mostrar que $p/x_0 < \sqrt{p} < x_0$.

Escolhe-se como uma nova aproximação $x_1 = (p/x_0 + x_0)/2$, isto é, a média aritmética entre p/x_0 e x_0. Novamente, pode-se mostrar que $p/x_1 < \sqrt{p} < x_1$.

Continuando desse modo, podemos construir uma sequência de aproximações dada por:

$$x_n = \begin{cases} (1+p)/2 & \text{se} \quad n = 0 \\ (\frac{p}{x_{n-1}} + x_{n-1})/2 & \text{se} \quad n \geq 1 \end{cases}$$

A tabela 1 fornece os valores de algumas aproximações para $\sqrt{2}$ obtidas pelo algoritmo de Eudoxo. Para que se possa avaliar a precisão das aproximações, são fornecidos também os quadrados dessas aproximações. Trabalhando com 14 dígitos depois do ponto decimal, é possível observar que, na quinta aproximação, x_4 temos, x_4=2,00000000000000

Algoritmo de Eudoxo para $\sqrt{2}$		
n	x_n	x_n^2
0	1,50000000000000	2,25000000000000
1	1,41666666666667	2,00694444444444
2	1,41421568627451	2,00000600730488
3	1,41421356237469	2,00000000000451
4	1,41421356237310	2,00000000000000

Tabela 1: Algoritmo de Eudoxo para $\sqrt{2}$. Fonte: de Freitas (2000, p. 11).

SAIBA MAIS!

Para saber mais sobre o algoritmo de Eudoxo, consulte o artigo publicado na Revista do Professor de Matemática 45 intitulado *Raiz Quadrada Utilizando Médias* (CARNEIRO, 2001). Nele você encontrará as justificativas para o funcionamento deste formidável método, bem como conhecerá um procedimento generalizado para o cálculo aproximado de raízes quadradas de números reais maiores que 1 usando médias. Encontrará ainda uma discussão sobre a precisão do processo, calculando-se o erro cometido nas aproximações.

Grosso modo, o cálculo numérico tem por objetivo estudar *técnicas numéricas* ou *métodos numéricos* para obter soluções de problemas reais que possam ser representados por *modelos matemáticos*, ou seja, o cálculo numérico busca produzir respostas numéricas para problemas matemáticos.

Torna-se evidente que o cálculo numérico é uma disciplina fundamental para a formação de profissionais das áreas de ciências exatas e engenharias, pois possibilita que os alunos conheçam várias técnicas para a solução de determinadas classes de problemas, saibam escolher entre estes métodos os mais adequados

a um problema específico e aplicá-los de modo a obter soluções de seus problemas. Desse modo, o cálculo numérico estabelece uma ligação entre a Matemática e os problemas práticos de áreas específicas.

Antes de tudo, devemos deixar claro que este é apenas um curso introdutório de cálculo numérico. Nele, esperamos que você, caro (a) aluno (a), adquira habilidades para:

- Compreender como os números são representados nas calculadoras e computadores e como são realizadas as operações numéricas nestes sistemas digitais.
- Entender o que são *métodos numéricos* de aproximação, como e por que utilizá-los, e quando é esperado que eles funcionem.
- Identificar problemas que requerem o uso de técnicas numéricas para a obtenção de sua solução.
- Conhecer e aplicar os principais métodos numéricos para a solução de certos problemas clássicos, por exemplo, obter zeros reais de funções reais, resolver sistemas de equações lineares, fazer interpolação polinomial, ajustar curvas e fazer integração numérica.
- Estimar e analisar os erros obtidos devido à aplicação de métodos numéricos e propor soluções para minimizá-los ou mesmo, quando possível, eliminá-los.

VOCÊ SABIA?

Os métodos numéricos desenvolvidos e estudados no cálculo numérico servem, em geral, para a aproximação da solução de problemas complexos que normalmente não são resolúveis por técnicas analíticas.

A aplicação das técnicas desenvolvidas no cálculo numérico para a resolução de problemas envolve, normalmente, um grande volume de cálculos (ou seja, o esforço computacional é alto), tornando imprescindível o trabalho de forma integrada com calculadoras, preferencialmente, científicas, gráficas ou programáveis ou com ambientes computacionais programáveis, os quais normalmente dispõem de ferramentas algébricas, numéricas e gráficas, facilitando e possibilitando o trabalho.

Com o desenvolvimento de rápidos e eficientes computadores digitais e de avançados ambientes de programação, a importância dos métodos numéricos tem aumentado significativamente na resolução de problemas.

Neste tópico, esperamos ter deixado claro para você, caro aluno, o papel e a importância do cálculo numérico como ferramenta para a resolução de problemas reais em diversas áreas e, especialmente, nas ciências exatas e engenharias. No próximo tópico, faremos um breve estudo sobre erros. Uma vez que os métodos numéricos fornecem soluções aproximadas para os problemas, tal análise se torna essencial.

TÓPICO 2

Fontes de erros, erros absolutos e relativos

OBJETIVOS

- Conhecer as principais fontes de erros
- Determinar erros absolutos e relativos

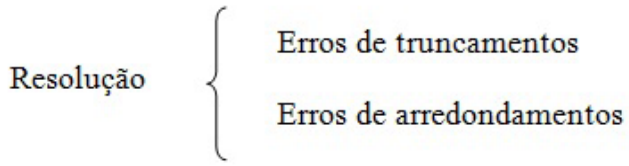
Você já deve ter percebido que, inerente ao processo de resolução de problemas reais via métodos numéricos, encontra-se o surgimento de erros. Neste tópico, iremos estudar várias fontes de erros que influenciam as soluções de problemas em cálculo numérico. Uma vez que os métodos numéricos fornecem soluções aproximadas para os problemas, tal análise se torna essencial. Veremos ainda as noções de erro absoluto e erro relativo, necessárias no decorrer de toda a disciplina.

Os erros cometidos para se obter a solução de um problema podem ocorrer em ambas as fases de modelagem e de resolução. Apresentaremos aqui as principais fontes de erros que levam a diferenças entre a solução exata e uma solução aproximada de um problema real, a saber:

- Erros nos dados.
- Simplificações na construção do modelo matemático.
- Erros de truncamentos.
- Erros de arredondamentos nos cálculos.

O esquema seguinte apresenta essas fontes de erros associadas à fase em que aparecem:

Modelagem { Erros nos dados
Simplificações na construção do modelo matemático



2.1 ERROS NOS DADOS

Os dados e parâmetros de um problema real são frequentemente resultados de medidas experimentais de quantidades físicas, de pesquisas ou de levantamentos e, portanto, são sujeitos a incertezas ou imprecisões próprias dos equipamentos de medições, dos instrumentos de pesquisas ou mesmo de ações humanas.

Tais erros surgem ainda da forma como os dados são armazenados no computador. Isso se deve ao fato de o computador usar apenas uma quantidade finita de dígitos para representar os números reais. Desse modo, torna-se impossível representar exatamente, por exemplo, números irracionais como as constantes matemáticas e e ϖ. Dependendo do sistema de numeração escolhido, até mesmo certos números racionais, inclusive inteiros, podem não ter uma representação exata em um determinado computador ou sistema eletrônico. A representação de números será objeto de estudo do próximo tópico dessa aula.

Há também a possibilidade de os dados serem originados pela solução numérica de outro problema que já carregam erros.

2.2 SIMPLIFICAÇÕES NA CONSTRUÇÃO DO MODELO MATEMÁTICO

Já vimos que, dependendo da abordagem dada ao problema, podemos ter modelos matemáticos diferentes. Muitas vezes, torna-se impossível obter um modelo matemático que traduza exatamente o problema real, enquanto, em outras, um tal modelo é demasiado complexo para ser tratado. Nesses casos, para obter um modelo tratável, necessitamos impor certas restrições idealistas de simplificações do modelo. O modelo matemático obtido então é um modelo aproximado que não traduz exatamente a realidade.

Devido às alterações e/ou simplificações, a solução de um modelo aproximado, ainda que exata, deve ser considerada suspeita de erros. É recomendável, então, que sejam feitos experimentos para verificar se as simplificações feitas são compatíveis com os dados experimentais, ou seja, é recomendável uma validação do modelo simplificado.

Desprezar a massa de um pêndulo ao se calcular o seu período, desprezar atritos ou resistências quando se trata de movimentos, dentre outras, são exemplos de simplificações de modelos.

2.3 ERROS DE TRUNCAMENTOS

Os erros de truncamento surgem quando processos infinitos ou muito grandes para a determinação de certo valor são interrompidos em um determinado ponto, ou seja, são substituídos por processos com uma limitação prefixada. Desse modo, podemos dizer que um erro de truncamento ocorre quando substituímos um processo matemático exato (finito ou infinito) por um processo aproximado correspondente a uma parte do processo exato. Ao consideramos um número finito de termos de uma série, estamos fazendo um truncamento da série.

Um exemplo claro desse tipo de erro pode ser visto quando calculamos e^x para algum número real x em um computador. O valor exato é dado pela série

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

Entretanto, por ser impossível somar os infinitos termos da série, fazemos apenas uma aproximação por um número finito de termos, ou seja, tomamos

$$e^x \cong \sum_{k=0}^{N} \frac{x^k}{k!}$$

em que N é um determinado número natural. Obviamente, à medida que N aumenta, mais precisa é a aproximação, ou seja, o erro de truncamento diminui.

2.4 ERROS DE ARREDONDAMENTOS

Os erros de arredondamento são aqueles que ocorrem no processo de cálculo de uma solução numérica, ou seja, surgem dos cálculos (operações aritméticas) existentes no método numérico. Tais erros estão associados ao fato de os computadores ou sistemas eletrônicos de cálculo utilizarem um número fixo de dígitos para representarem os números, isto é, são consequências de se trabalhar com o que chamamos aritmética de precisão finita.

Desse modo, sempre que o resultado de uma operação for um número que não pode ser representado exatamente no sistema de representação usado, necessitamos fazer arredondamentos, o que leva a desprezar dígitos e arredondar o número.

GUARDE BEM ISSO!

Em cálculo numérico, lidamos essencialmente com valores aproximados e a quase totalidade dos cálculos envolve erros. Assim não podemos usar métodos numéricos e ignorar a existência de erros.

Vale ressaltar que, mesmo quando as parcelas ou fatores de uma operação podem ser representados exatamente no sistema, não se pode esperar que o resultado da operação armazenado seja exato.

Uma vez que em nossa disciplina estaremos mais focados nos métodos numéricos, daremos maior ênfase aos erros de truncamento e de arredondamento.

Nosso principal interesse em conhecer as fontes de erros que ocorrem quando do uso de métodos numéricos reside na tentativa eliminá-los ou, pelo menos, de poder controlar o seu valor. Neste contexto, são de grande importância o conhecimento dos efeitos da propagação de erros e a determinação do erro final das operações numéricas.

Finalizamos este tópico apresentado as noções muito úteis de erro absoluto e erro relativo.

2.5 ERRO ABSOLUTO

Você já sabe que, ao resolvermos um problema real utilizando métodos numéricos, os resultados obtidos são geralmente aproximações do que seria o valor exato de uma solução do problema. Dessa forma, é inerente aos métodos se trabalhar com as aproximações e com os erros.

A informação sobre o erro que acompanha uma aproximação para a solução de um problema é fundamental para se conhecer a qualidade da aproximação e para termos uma noção mais clara sobre o valor exato da solução. Vejamos um exemplo:

Exemplo 3:

Considere a equação $2x^3 + 3x - 7 = 0$. Essa equação tem uma única raiz real. São aproximações para essa raiz os números 1,195000, 1,195175 e 1,195200. Agora, qual dessas aproximações é a mais exata, ou seja, qual delas mais se aproxima do valor exato da raiz? Para respondermos a esta pergunta, e para termos uma informação mais precisa sobre o valor exato da raiz, é necessário conhecer a qualidade da aproximação.

Apesar de, em geral, aumentando o esforço computacional, as aproximações poderem ser melhoradas, torna-se importante medir o quão próximo uma aproximação está do valor exato. Para quantificar essa informação, introduzimos a noção de erro absoluto.

Definição 1: Seja x um número e $\overline{x}$ uma sua aproximação, chama-se erro absoluto, e designa-se por EA_x, a diferença entre x e $\overline{x}$. Simbolicamente:

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

No caso de $x > \overline{x}$, ou seja, quando $EA_x > 0$, dizemos que $\overline{x}$ é uma aproximação por falta e, no caso de $x < \overline{x}$, ou seja, quando $EA_x < 0$, dizemos que $\overline{x}$ é uma aproximação por excesso.

Exemplo 4:

Como $3{,}14 < \pi < 3{,}15$, temos que 3,14 é uma aproximação de π por falta e 3,15 uma aproximação de π por excesso.

Entretanto, desde que, geralmente, não conhecemos o valor exato x (aliás, esta é a razão de procurarmos uma aproximação $\overline{x}$ para x), torna-se impossível determinar o valor exato do erro absoluto. Nesses casos, o que pode ser feito é a determinação de um limitante superior ou de uma estimativa para o módulo do erro absoluto.

No exemplo 2, uma vez que $\pi \in (3{,}14;\ 3{,}15)$, se tomarmos como aproximação para π, um valor $\overline{\pi}$ também pertence ao intervalo $(3{,}14;\ 3{,}15)$, teremos

$$|EA_\pi| = |\pi - \overline{\pi}| < 0{,}01\,,$$

que significa que o erro absoluto cometido é inferior a um centésimo.

Se $\varepsilon > 0$ é uma cota para $\mathrm{EA_x}$, ou seja, se $|\mathrm{EA_x}| < \varepsilon$, temos:

$$|EA_x| < \varepsilon \Leftrightarrow |x - \overline{x}| < \varepsilon \Leftrightarrow \overline{x} - \varepsilon < x < \overline{x} + \varepsilon\,.$$

Portanto, é possível precisar que o valor exato x (provavelmente não conhecido) está compreendido entre dois valores conhecidos: $\overline{\mathrm{x}} - \varepsilon$ e $\overline{\mathrm{x}} + \varepsilon$. Na prática, é desejável que uma cota para $\mathrm{EA_x}$ seja bem próxima de 0.

Contudo, o erro absoluto pode não ser suficiente para informar sobre a qualidade da aproximação. Para ilustrar isso, consideremos duas situações: a primeira foi adaptada de Ruggiero e Lopes (1996, p. 13), e a segunda de Freitas (2000, p. 18):

SAIBA MAIS!

Um número $\varepsilon > 0$ tal que $|\mathrm{EA_x}| < \varepsilon$ é chamado **cota** para o erro $\mathrm{EA_x}$.

ATENÇÃO!

Para descrever o intervalo $(3{,}14;\ 3{,}15)$, usamos o separador ponto-e-vírgula (;) em vez de vírgula (,) como fazemos normalmente. Para evitar confusão, faremos isso sempre que algum dos extremos tiver parte fracionária (que precisa ser separada da parte inteira por vírgula).

Situação 1

Seja um número x com uma aproximação $\bar{x} = 2112,9$ tal que $|EA_x| < 0,1$, o que implica $x \in (2112,8; 2113)$ e seja um número y com uma aproximação $\bar{y} = 5,3$ tal que $|EA_y| < 0,1$, o que implica $y \in (5,2; 5,4)$. Note que os limites superiores para os módulos dos erros absolutos são os mesmos. Podemos dizer que os números estão representados por suas aproximações com a mesma precisão?

Situação 2

Considere $x = 100$; $\bar{x} = 100,1$ e $y = 0,0006$; $\bar{y} = 0,0004$. Assim, $EA_x = 0,1$ e $EA_y = 0,0002$. Como $|EA_y|$ é muito menor que $|EA_x|$, é possível afirmar que a aproximação $\bar{y}$ de y é melhor que a aproximação $\bar{x}$ de x?

Para responder os questionamentos acima, é preciso comparar, em ambas as situações, a ordem de grandeza de x e de y. Uma primeira análise nos permite afirmar que as grandezas dos números envolvidos são bastante diferentes. Para a situação 1, é possível concluir ainda que a aproximação para x é mais precisa que a aproximação para y, pois as cotas para os erros absolutos são as mesmas (0,1), e a ordem de grandeza de x é maior que a ordem de grandeza de y. Já para a situação 2, a ordem de grandeza de x é também maior que a ordem de grandeza de y, mas, como a cota para o erro em x é maior que aquela para o erro em y, precisamos fazer uma análise mais cuidadosa. Para tanto, introduzimos a noção de erro relativo.

Definição 2: Seja x um número e $\bar{x} \neq 0$ uma sua aproximação, chama-se erro relativo, e designa-se por ER_x, a razão entre EA_x e $\bar{x}$. Simbolicamente:

$$ER_x = \frac{EA_x}{\bar{x}} = \frac{x - \bar{x}}{\bar{x}}.$$

Ao produto $100 \times ER_x$, chamamos erro percentual ou percentagem de erro.

Exemplo 5:

Vamos calcular cotas para os erros relativos cometidos nas aproximações na Situação 1. Temos

$$|ER_x| = \frac{|EA_x|}{|\bar{x}|} < \frac{0,1}{2112,9} \cong 4,73 \times 10^{-5}$$

e,

$$|ER_y| = \frac{|EA_y|}{|\bar{y}|} < \frac{0,1}{5,3} \cong 1,89 \times 10^{-2}.$$

Isso confirma que a aproximação para x é mais precisa que a aproximação para *y*. De fato, um erro da ordem de 0,1 é bem menos significativo para *x* que é da ordem de milhares do que para *y* que é da ordem de unidades.

Exemplo 6:

Vamos calcular os erros relativos e os erros percentuais cometidos nas aproximações na Situação 2. Temos

$$ER_x = \frac{EA_x}{\bar{x}} = \frac{0,1}{100,1} \cong 9,99 \times 10^{-4}$$

$$100 \times ER_x \cong 100 \times 9,99 \times 10^{-4}\% \cong 0,1\%$$

e

$$ER_y = \frac{EA_y}{\bar{y}} = \frac{0,0002}{0,0006} \cong 3,33 \times 10^{-1}$$

$$100 \times ER_x \cong 100 \times 3,33 \times 10^{-1}\% = 33,3\%$$

ATENÇÃO!

Do mesmo modo que para o erro absoluto, na maior parte dos casos, não é possível a determinação exata do erro relativo. Isso porque, em geral, não se conhece o valor exato de *x*, mas apenas uma aproximação $\bar{x}$. A partir de uma cota para o erro absoluto, podemos calcular uma cota para o erro relativo.

Portanto, ao contrário do que poderia parecer, a aproximação para *x* é mais precisa que a aproximação para *y*. Assim, um erro da ordem de 0,1 para *x*, que é da ordem de centenas, é menos significativo que um erro de 0,0002 para *y*, que é da ordem de décimos de milésimos.

Conhecemos, neste tópico, as principais fontes geradoras de erros quando do uso de métodos numéricos para a resolução de problemas reais. Vimos ainda formas de medir os erros cometidos ao se tomar uma aproximação para um determinado valor.

No próximo tópico faremos uma breve apresentação sobre representação de números.

TÓPICO 3

Representação de números e aritmética de ponto flutuante

OBJETIVOS

- Apresentar formas de representação numérica
- Conhecer sistemas de numeração
- Aprender a representar números em ponto flutuante

Reservamos este último tópico para tratar das formas de representação dos números em sistemas de numeração. Daremos ênfase à representação dos números em *ponto flutuante,* comumente adotada em sistemas digitais como calculadoras e computadores.

A necessidade de contar e de registrar o total de objetos contados é muita antiga e o homem utilizou vários processos de fazê-los. Desde a contagem via correspondência um a um, com o registro por meio de marcas (uma para cada objeto), passando pelas contagens por agrupamentos que facilitavam as contagens de grandes quantidades de objetos, foram muitos os avanços alcançados. Outra necessidade marcante era a de fazer medições e registrar os resultados dessas medições.

À medida que se civilizava, a humanidade foi apoderando-se de modelos abstratos para os registros das contagens e das medições, os números. Dessa forma os números surgiram, principalmente, da necessidade de o homem contar e medir. De acordo com Lima (2003, p. 25), os "números são entes abstratos, desenvolvidos pelo homem como modelos que permitem contar e medir, portanto avaliar as diferentes quantidades de uma grandeza".

Associados ao conceito de *número* estão os conceitos de *numeral* e de *sistema de numeração*, fundamentais para que se possam representar os números. Em linhas breves, podemos dizer que

1. Um **número** é uma noção matemática que serve para descrever uma quantidade ou medida.

2. Um **numeral** é um símbolo ou conjunto de símbolos que representam um número.
3. Um **sistema de numeração** é um conjunto de numerais que representam os números. Para tal, é fixado um número natural b, $b>1$, denominado base do sistema de numeração e são utilizados elementos do conjunto $\{0,1,2,\ldots,b-1\}$, denominados **algarismos** ou **dígitos** do sistema de numeração.

No nosso dia a dia, estamos acostumados a lidar com o sistema de numeração de base 10 ou sistema de numeração decimal. Esse sistema que utiliza 10 dígitos – 0, 1, 2, 3, 4, 5, 6, 7, 8 e 9 – para a representação dos números é o mais utilizado para a comunicação entre as pessoas. No caso de representações no sistema de numeração decimal, a indicação da base torna-se desnecessária, por isso costumamos omiti-la. Assim, a menos que seja especificada outra base, sempre que falamos em um número ou escrevemos o seu numeral, referimo-nos a eles no sistema de numeração decimal.

ATENÇÃO!

A rigor, sempre que escrevemos o numeral que representa um número, deveríamos indicar a base do sistema de numeração adotado.

Uma importante característica do sistema de numeração decimal é o fato de ele ser posicional, ou seja, nele o valor de cada símbolo é relativo, dependendo da sua posição no número.

Exemplo 7:

No número 46045 temos

1. o primeiro algarismo 4 ocupa a posição das dezenas de milhares, valendo 4 dezenas de milhares ou $4\times10000=40000$ unidades ou ainda 4×10^4 unidades.
2. o algarismo 6 ocupa a posição das unidades de milhar, valendo 6 unidades de milhar ou $6\times1000=6000$ unidades ou ainda 6×10^3 unidades.
3. o algarismo 0, ocupando a posição das centenas, indica ausência de centenas ou $0\times100=0$ unidades ou ainda 0×10^2 unidades.
4. o segundo algarismo 4 ocupa a posição das dezenas, valendo 4 dezenas ou $4\times10=40$ unidades ou ainda 4×10^1 unidades.

5. o algarismo 5 ocupa a posição das unidades, valendo $5\times1=5$ unidades ou ainda 5×10^0 unidades.

Logo, 46045 significa $4\times10^4+6\times10^3+0\times10^2+4\times10^1+5\times10^0$.

O próximo teorema é bem conhecido e estabelece que qualquer número natural pode ser representado de modo único em uma base qualquer.

> **Teorema 1:** *Seja B um inteiro maior que 1, então cada* $N\in\mathbb{N}$ *admite uma representação única da forma*
>
> $$N=a_m\times B^m+a_{m-1}\times B^{m-1}+\ldots+a_2\times B^2+a_1\times B^1+a_0,$$
>
> *em que* $a_m\neq0$ *e* $0\leq a_i<B$, *para toda i com* $0\leq i\leq m$.

A demonstração desse teorema pode ser vista nos livros de Teoria dos Números. Para exemplificar, vamos representar um determinado número em algumas bases bem conhecidas.

Exemplo 8:

Representar o número 69 nas bases 2 (binária), 8 (octal), 10 (decimal) e 16 (hexadecimal). Temos

$$69=1\times2^6+0\times2^5+0\times2^4+0\times2^3+1\times2^2+0\times2^1+1\times2^0$$
$$69=1\times8^2+0\times8^1+5\times8^0$$
$$69=6\times10^1+9\times10^0$$
$$69=4\times16^1+5\times16^0$$

Portanto, 69 é escrito como 1000101 na base 2, 105 na base 8, 69 na base 10 e 45 na base 16. Usando uma notação com o numeral entre parênteses e base como índice, temos que 69 é escrito como $(1000101)_2$, $(105)_8$, $(69)_{10}$ e $(45)_{16}$. Assim,

$(1000101)_2=(105)_8=(69)_{10}=(45)_{16}$.

A figura 2 apresenta a representação nas bases binária, octal, decimal e hexadecimal dos números de 1 a 20.

BINÁRIA	OCTAL	DECIMAL	HEXADECIMAL
00001	01	01	01
00010	02	02	02
00011	03	03	03
00100	04	04	04
00101	05	05	05
00110	06	06	06

00111	07	07	07
01000	10	08	08
01001	11	09	09
01010	12	10	0A
01011	13	11	0B
01100	14	12	0C
01101	15	13	0D
01110	16	14	0E
01111	17	15	0F
10000	20	16	10
10001	21	17	11
10010	22	18	12
10011	23	19	13
10100	24	20	14

Figura 2: Representação dos números de 1 a 20 em diferentes bases.

O teorema 1 apresenta a representação de números inteiros positivos em uma base qualquer. Entretanto, ele pode ser generalizado para a representação de números reais positivos de modo natural. Assim, se *B* é um inteiro maior que 1, então o número

$$a_m a_{m-1} \ldots a_2 a_1 a_0 , a_{-1} a_{-2} \ldots$$

representa, na base 10, o número

ATENÇÃO!

Na representação amam-1...a2a1a0,a-1 a-2... , a vírgula (,) separa a parte inteira da parte fracionária. Essa é a notação mais comum no Brasil. Alguns autores, entretanto, talvez influenciados pela notação usada pelos ingleses e americanos, usam o ponto (.) como separador.

$$\overbrace{a_m \times B^m + a_{m-1} \times B^{m-1} + \ldots + a_2 \times B^2 + a_1 \times B^1 + a_0 \times B^0}^{\text{Parte Inteira}} + \overbrace{a_{-1} \times B^{-1} + a_{-2} \times B^{-2} + \ldots}^{\text{Parte Fracionária}},$$

em que $a_m \neq 0$ e $0 \leq a_i < B$, para toda *i* com $0 \leq i \leq m$.

Exemplo 9:

$$(1101,101)_2 = 1\times 2^3 + 1\times 2^2 + 0\times 2^1 + 1\times 2^0 + 1\times 2^{-1} + 0\times 2^{-2} + 1\times 2^{-3} = 13,625$$
$$(470,75)_8 = 4\times 8^2 + 7\times 8^1 + 0\times 8^0 + 7\times 8^{-1} + 5\times 8^{-2} = 312,953125$$
$$(142,857)_{10} = 1\times 10^2 + 4\times 10^1 + 2\times 10^0 + 8\times 10^{-1} + 5\times 10^{-2} + 7\times 10^{-3} = 142,857$$
$$(D3,A2)_{10} = 13\times 16^1 + 3\times 16^0 + 10\times 16^{-1} + 2\times 16^{-2} = 107,6328125$$

Para facilitar a representação física, a definição das operações aritméticas

SAIBA MAIS!

A representação de números reais em certa base no formato parte inteira, vírgula (ou ponto), parte fracionária, como mostrado na figura 3, é também chamada representação em *ponto fixo*.

Parte Inteira	.	Parte Fracionária

Figura 3: Representação de números reais em ponto fixo.

e a comunicação entre as máquinas digitais, é necessário fazer uso de outros sistemas de representação. Os computadores comumente operam no *sistema binário* (base 2), o qual usa apenas dois algarismos (0 e 1), correspondentes aos estados ausência ou presença de sinal elétrico, respectivamente. Outras bases também são ou foram utilizados.

Assim, é importante conhecer a *representação de números* em bases diferentes da base decimal e a conversão de números de uma para outra base é uma tarefa muitas vezes necessária. Vale destacar que um mesmo número pode ter representação finita (exata) em uma base, mas sua representação em outra base pode ser infinita. Por conseguinte, a própria representação de um número em uma determinada base pode ser uma fonte de erros. De acordo com Ruggiero e Lopes (1996, p. 3-4), na interação entre o usuário e o computador:

> ... os dados de entrada são enviados ao computador pelo usuário no sistema decimal; toda esta informação é convertida para o sistema binário, e as operações todas serão efetuadas neste sistema. Os resultados finais serão convertidos para o sistema decimal e, finalmente, serão transmitidos ao usuário. Todo este processo de conversão é uma fonte de erros que afetam o resultado final dos cálculos.

VOCÊ SABIA?

De modo geral, qualquer número (inteiro ou fracionário) pode ser expresso no formato número x base$^{\text{expoente}}$, em que variam a posição da vírgula e o expoente ao qual elevamos a base. Essa representação é denominada *representação em ponto flutuante*, pois o ponto varia sua posição de acordo com o expoente escolhido. Na forma normalizada, o número é representado movendo-se a vírgula de forma que o número seja menor que 1, o mais próximo possível de 1. Isso significa que o primeiro dígito significativo virá imediatamente após a vírgula.

Por outro lado, a representação em ponto fixo, ainda que cômoda para cálculos no papel, não é adequada para processamento nos computadores ou calculadoras. Nestes sistemas, costuma-se usar uma representação denominada representação em *ponto flutuante normalizada*. Nela, um número é representado na forma

$$\pm 0, d_1 d_2 \ldots d_t \times B^e ,$$

em que, para cada i = 1, 2, ..., t, d_i é um

inteiro com $0 \leq d_i < B$ e $d_1 \neq 0$, e é um inteiro no intervalo tal que $l \leq e \leq u$. O número $0, d_1 d_2 \ldots d_t$ é chamado de *mantissa*, *B* é a *base* do sistema, *t* é o *número de algarismos na mantissa* (algarismos significativos) e *l* e *u* são, respectivamente, os limites inferior e superior para o expoente *e*.

Observe que a representação em ponto flutuante normalizada corresponde a um deslocamento da vírgula na representação em ponto fixo que se dá pela multiplicação do número por uma correspondente potência da base do sistema.

Para fixar melhor a representação em ponto flutuante normalizada, vejamos alguns exemplos:

Exemplo 10:

Considere uma máquina *S* com representação em ponto flutuante normalizada na base binária, com t = 8 e $e \in [-5, 5]$. Temos, então:

o número $n_1 = 0{,}10100110 \times 2^3$ representado em *S* corresponde, na base 10, a 5,1875 e o número $n_2 = 0{,}10100111 \times 2^3$ representado em *S* corresponde, na base 10, a 5,21875. Como exercício, verifique essas correspondências.

Perceba que nesse sistema, n_1 e n_2 são dois números consecutivos. Portanto, não é possível representar em *S* qualquer número compreendido entre 5,1875 e 5,21875. Assim, o 5,2, por exemplo, não tem representação exata em *S*. Esta perda de precisão se dá porque o número de dígitos na mantissa não é suficiente.

Exemplo 11:

Considerando a mesma máquina S do exemplo 7, temos

maior número real representado: $M = +0{,}11111111 \times 2^5$ que corresponde a $+31{,}875$.

menor número real representado: $-M = -0{,}11111111 \times 2^5$ que corresponde a $-31{,}875$.

menor número real positivo representado: $m = +0{,}10000000 \times 2^{-5}$ que corresponde a $+0{,}015625$.

maior número real negativo representado: $-m = -0{,}10000000 \times 2^{-5}$ que corresponde a $-0{,}015625$.

Como exercício, verifique essas correspondências.

Portanto, por falta de expoentes maiores que $u = 5$, não é possível representar em *S* números que sejam menores que $-M$ ou maiores que *M*, isto é, não é possível representar números *x* tais $|x| > M$. Nestes casos, a máquina costuma

retornar um erro de *overflow*. Por outro lado, por falta de expoentes menores que $l = -5$, também não é possível representar em S números que são menores que estão entre $-m$ e m, ou seja, não é possível representar números x tais $|x| < m$. Nestes casos, a máquina costuma retornar um erro de *underflow*.

Dos exemplos acima, podemos concluir que, quanto maior o intervalo para o expoente, maior será a faixa de números que um sistema pode representar; e, quanto maior o número de algarismos para a mantissa, maior será a precisão da representação. Vejamos mais um exemplo, este extraído de Ruggiero e Lopes (1996, p. 12):

Exemplo 12:

Veja a representação de alguns números em um sistema de aritmética de ponto flutuante de três dígitos para $B = 10$, $l = -4$ e u = 5:

x	Representação por arredondamento
3,42	$0{,}342 \times 10^{1}$
200,65	$0{,}201 \times 10^{3}$
85,7142	$0{,}857 \times 10^{2}$
0,0041887...	$0{,}419 \times 10^{-2}$
9999,99	$0{,}100 \times 10^{5}$
0,0000078	*Underflow*
123456,789	*Overflow*

Tabela 2: Representação em ponto flutuante com arredondamento.

ATENÇÃO!

Vale ressaltar que as operações de adição e multiplicação em aritmética de ponto flutuante não gozam das propriedades associativas e distributivas.

Finalizamos este tópico, fazendo três observações importantes sobre a representação e a aritmética de ponto flutuante normalizada:

1. A *adição* de dois números em aritmética de ponto flutuante é feita com o alinhamento dos pontos decimais, do seguinte modo: a mantissa do número de menor expoente é deslocada para a direita até que os expoentes se igualem, ou seja, o deslocamento é de um número de casas igual à diferença dos expoentes. Somam-se as mantissas e repete-se o expoente e, se necessário, faz-se a normalização.

Exemplo:

Em um sistema de base 10 com $t = 4$, temos

$$\begin{aligned} 0,4370 \times 10^5 + 0,1565 \times 10^3 &= 0,4370 \times 10^5 + 0,0016 \times 10^5 \\ &= (0,4370 + 0,0016) \times 10^5 \\ &= 0,4386 \times 10^5 \end{aligned}$$

O zero em ponto flutuante é representado por mantissa nula (0,00...0) e com o menor expoente disponível. Caso o expoente não fosse o menor possível, mesmo a mantissa sendo nula, poderia ocasionar a perda de dígitos significativos na adição deste zero a um outro número. Isso se dá pela forma como a adição é realizada em aritmética de ponto flutuante.

Exemplo:

Em um sistema de base 10 com $t = 4$, temos

$$\begin{aligned} 0,0000 \times 10^0 + 0,1428 \times 10^{-2} &= 0,0000 \times 10^0 + 0,0014 \times 10^0 \\ &= 0,0014 \times 10^0 \\ &= 0,1400 \times 10^{-2} \end{aligned}$$

A *multiplicação* de dois números em aritmética de ponto flutuante é feita multiplicando-se as mantissas dos números e somando-se os expoentes; em seguida, se necessário, faz-se a normalização.

Exemplo:

Em um sistema de base 10 com $t = 4$, temos

$$\begin{aligned} 0,4370 \times 10^5 \times 0,1565 \times 10^3 &= (0,4370 \times 0,1565) \times 10^{5+3} \\ &= 0,6839 \times 10^{-1} \times 10^5 \\ &= 0,6839 \times 10^4 \end{aligned}$$

Nesta aula, fizemos uma breve introdução ao estudo do Cálculo Numérico, apresentando a sua importância para a resolução de diversos problemas reais nas mais diversas áreas, especialmente ciências exatas e engenharias. Uma vez que o Cálculo Numérico trabalha com aproximações, demos algumas noções de erros, apontando como surgem e de que modo podemos medi-los. Finalmente, apresentamos formas de representação dos números, enfatizando a r*epresentação em ponto flutuante*.

SAIBA MAIS!

Você pode aprofundar seus conhecimentos consultando as referências que citamos e/ou visitando páginas da internet. Abaixo, listamos uma página interessante que pode ajudá-lo nessa pesquisa. Bons estudos!

http://www.profwillian.com/_diversos/download/livro_metodos.pdf

AULA 2

Zeros reais de funções reais

Caro (a) aluno (a),

Nesta segunda aula, abordaremos um importante problema que aparece com muita frequência em diversas áreas: encontrar zeros reais de funções reais. Iniciaremos fazendo uma breve introdução de apresentação do problema. Daremos também o significado geométrico para os zeros reais de funções reais e veremos como fazer a localização ou isolamento de tais zeros utilizando como recursos o tabelamento e a análise gráfica da função. Então, vamos ao problema!

Objetivos

- Contextualizar o problema de determinar zeros de funções
- Apresentar técnicas para resolver o problema
- Rever conceitos e resultados necessários do cálculo
- Localizar zeros reias de funções reais

TÓPICO 1

Conhecendo o problema e sua importância

OBJETIVOS

- Conhecer o problema e constatar sua importância
- Dar o significado geométrico de zeros reais de funções reais
- Conhecer a ideia geral dos métodos iterativos para resolver o problema

Neste tópico, introduziremos o problema geral de determinar a existência de e de calcular zeros reais de funções reais e conheceremos a sua importância para as mais diversas áreas do conhecimento humano, justificando assim a sua inclusão entre os problemas que são objetos de estudo do cálculo numérico. Faremos ainda a interpretação geométrica e estabeleceremos a ideia central dos métodos numéricos iterativos para a obtenção de zeros reais de funções reais. Iniciaremos com uma definição.

Definição 1: *Dada uma função* $f:\mathbb{R}\to\mathbb{R}$ *(função real de uma variável real), chama-se zero de f a todo* $a\in\mathbb{R}$ *tal que* $f(a)=0$.

GUARDE BEM ISSO!

O problema de determinar zeros de uma função aparecerá sempre que tivermos de resolver uma equação.

Portanto, o problema de *determinar os zeros reais de uma função f* (que é o problema no qual estamos interessados) equivale ao problema de *determinar as raízes reais da equação* $f(x)=0$, ou seja, determinar os valores $a\in\mathbb{R}$ que satisfazem $f(a)=0$.

Vejamos algumas situações em que este problema aparece.

EXEMPLO 1:

Considere um circuito elétrico composto apenas de uma fonte de tensão V e

de uma resistência R, como ilustrado na figura 1a. O modelo matemático para calcular a corrente que circula no circuito é conhecido como *Lei de Kirchoff*, sendo dado pela equação

$$V - Ri = 0 .$$

Este é um modelo bem simples: uma equação linear a uma incógnita cuja única raiz é dada por $i = V / R$. Agora, como indicado na figura 1b, se introduzirmos neste circuito elétrico um diodo *D* (dispositivo ou componente eletrônico semicondutor usado como retificador de corrente elétrica), o modelo matemático para determinar a corrente que circula no circuito será dado pela equação:

$$V - Ri - \frac{kT}{q}\ln\left(\frac{i}{I_s} + 1\right) = 0$$

em que k e I_s são constantes, q é a carga do elétron e T é a temperatura do dispositivo (BUFFONI, 2002).

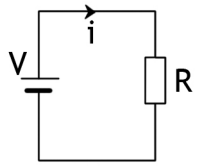


Figura 1a: Circuito elétrico

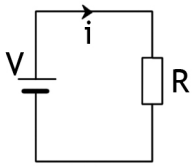


Figura 1b: Circuito elétrico

Exemplo 2:

Para encontrar a quantidade de ácido que se ioniza em uma solução em equilíbrio, o modelo matemático (obtido de teorias da química) é dado pela equação

$$x^2 + k_a x - k_a C_0 = 0 ,$$

em que k_a indica a constante de ionização do ácido e C_0 representa a concentração inicial do ácido (BERLEZE E BISOGNIN, 2006). Este modelo é de uma equação quadrática e suas raízes (reais ou não) são dadas pela conhecida *fórmula de Bhaskara*.

Exemplo 3:

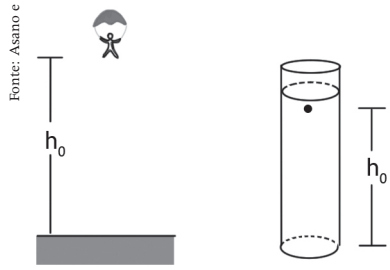
O tempo de queda de um paraquedista ou de uma bolinha dentro d'água

SAIBA MAIS!

As Leis de Kirchhoff são bastante utilizadas em circuitos elétricos mais complexos. Acesse o site http://www.infoescola.com/eletricidade/leis-de-kirchhoff/ e conheça mais sobre as leis desse brilhante físico.

(ASANO e COLLI, 2007, p. 90-93):

> "Imagine um paraquedista que abre seu paraquedas no instante $t = 0$, da altura h_0, ou, alternativamente, uma bolinha que parte do repouso à altura h_0 dentro de um tubo cheio d'água, e cai sob a força da gravidade. Levando em conta que a queda não é completamente livre, isto é, o meio oferece resistência ao movimento, quanto tempo levará a queda do paraquedista e da bolinha?"



Fonte: Asano e Colli (2007, p. 90)

Figura 2: Tempo de queda.

Resolver este problema corresponde a obter as raízes da equação $h(t) = h_0$, em que

$$h(t) = A + Bt - Ce^{-Dt},$$

com A, B, C e D sendo constantes que dependem da constante de aceleração da gravidade à superfície terrestre g, da altura inicial h_0, da massa do corpo m, da velocidade inicial do corpo v_0 e da velocidade para a qual a força de resistência do meio é exatamente igual à força da gravidade mg. Equivalentemente, o problema consiste em obter os zeros da função *f*, dada por

$$f(t) = h(t) - h_0.$$

Para maiores detalhes, incluindo a dedução da equação acima, veja a referência Asano e Colli (2007, p. 90).

VOCÊ SABIA?

Dada uma função $f : \mathbb{R} \to \mathbb{R}$, os zeros de *f* correspondem às abscissas dos pontos em que o gráfico de *f* intercepta o eixo das abscissas. De fato, $f(a) = 0 \Leftrightarrow (a, 0) \in \text{Graf}(f)$.

Os exemplos acima são de situações concretas e mostram a importância do problema de obter zeros reais de funções reais ou, equivalentemente, de determinar as raízes reais de equações. No primeiro caso do exemplo 1 e no exemplo 2, pela simplicidade dos modelos, as raízes são obtidas de modo exato através de fórmulas, dispensando o uso de métodos numéricos específicos. Já no segundo

caso do exemplo 1 e no exemplo 3, os modelos não são tão simples, não havendo fórmulas explícitas para o cálculo das raízes. Nesses casos, os métodos numéricos tornam-se indispensáveis.

Apesar de certas equações (como as polinomiais) poderem apresentar raízes complexas, o nosso interesse será somente nas raízes reais das equações, ou seja, nos zeros reais das funções correspondentes. Há uma interpretação gráfica para os zeros reais de funções reais:

Para a função f cujo gráfico está esboçado abaixo (figura 3), temos que os números x_1, x_2 e x_3 são zeros reais de f.

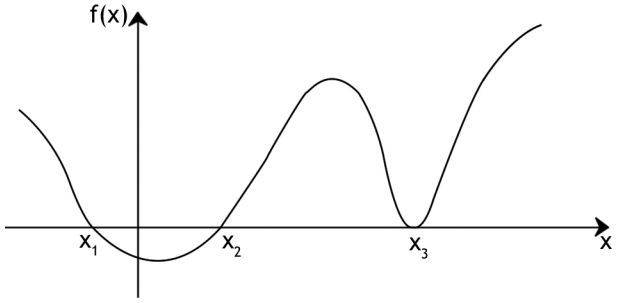


Figura 3: Zeros reais de uma função real

Até agora, já sabemos a importância de calcular zeros reais de funções reais e o significado geométrico de tais zeros. Você deve está se perguntando:

Como calcular os zeros reais de uma dada função?

É o que pretendemos responder a partir de agora. Sabemos que, para certas funções, como as polinomiais afins ou quadráticas, tais zeros podem ser obtidos diretamente através de fórmulas. Entretanto, existem funções (e, na maioria dos problemas reais, é isto que ocorre) para as quais não existem ou são muito complexas as fórmulas para o cálculo exato de seus zeros. Nesses casos, precisamos recorrer a *métodos numéricos*. Tais métodos podem ser utilizados no cálculo de um zero real (caso exista) de qualquer função contínua dada.

VOCÊ SABIA?

1. Em geral, um método (processo ou procedimento) numérico iterativo calcula uma sequência de aproximações de um zero de f, cada uma mais precisa que a anterior. Assim, a repetição do processo fornece, em um número finito de vezes, uma aproximação a qual difere do valor exato do zero por alguma precisão (tolerância) prefixada.
2. O cálculo de cada nova aproximação é feito utilizando aproximações anteriores, porém as aproximações iniciais que o processo exigir devem ser fornecidas.

Em geral, salvo raras exceções, os métodos numéricos iterativos não fornecem os zeros exatos de uma função f. Eles podem, entretanto, ser usados para o cálculo de aproximações para estes zeros.

A princípio, obter apenas uma aproximação para o zero (e não seu valor exato) da função f pode parecer uma limitação, mas ela não é uma limitação tão séria, pois, com os métodos numéricos que trabalharemos, será possível obter aproximações "boas" ou "satisfatórias". Para sermos mais precisos, a menos de limitações de máquinas, é possível encontrar um zero de uma função com qualquer *precisão prefixada*. Isso significa que a aproximação pode ser tomada tão próxima do valor exato do zero quanto se deseje.

Relembre que a diferença entre o valor exato de um zero $\overline{x}$ de f e de um seu valor aproximado $\tilde{x}$ é chamada erro absoluto (ou, simplesmente, erro). Como vimos na aula 1, por não conhecer o valor exato $\overline{x}$, não podemos determinar o valor exato do erro. Nestes casos, o que se costuma fazer é delimitar o erro, ou seja, exigir que $|\overline{x}-\tilde{x}|<\delta$ para algum $\delta>0$ previamente escolhido. Desse modo, temos $\tilde{x}-\delta<\overline{x}<\tilde{x}+\delta$ e diremos que $\tilde{x}$ é uma aproximação de $\overline{x}$ com precisão δ.

Obviamente, será interessante que a sequência $x_1, x_2, x_3, \ldots$ gerada por um processo iterativo convirja para algum $\overline{x}\in\mathbb{R}$. Neste caso, dizemos também que o processo iterativo converge para $\overline{x}$. Você já deve ter visto o conceito de convergência de uma sequência em disciplinas anteriores, entretanto vamos relembrá-lo:

> **Definição 2:** *Uma sequência* $x_1, x_2, x_3, \ldots$, *denotada por* $(x_n)_{n\in\mathbb{N}}$, *converge para* $\overline{x}$, *se* $\lim_{n\to\infty} x_n = \overline{x}$. *Ou seja, se dado* $\varepsilon>0$, $\exists N\in\mathbb{N}$ *tal que qualquer que seja* $n>N$, $|x_n-\overline{x}|<\varepsilon$. *Isto será indicado por* $x_n\to\overline{x}$.

Os métodos numéricos iterativos para o cálculo de um zero real de uma função real *f* que apresentaremos envolvem duas fases:

→ **Fase 1 - Isolamento ou localização dos zeros:** consiste em achar intervalos fechados disjuntos $[a,b]$, cada um dos quais contendo exatamente um zero de *f*.

→ **Fase 2 - Refinamento:** consiste em, partindo de aproximações iniciais escolhidas em um determinado intervalo obtido na fase 1, melhorar (refinar) sucessivamente as aproximações até obter uma aproximação para o zero de *f* que satisfaça uma precisão prefixada.

Neste tópico, apresentamos o problema de calcular zeros reais de funções reais e percebemos sua importância. Demos também o significado geométrico de tais zeros e vimos a necessidade do uso de métodos numéricos iterativos para resolver este problema. No próximo tópico, trataremos da fase inicial de isolamento dos zeros de uma função.

TÓPICO 2

Isolamento ou localização de zeros reais

OBJETIVOS

- Construir tabelas e esboçar gráficos de funções
- Isolar ou localizar zeros reais de funções reais
- Classificar métodos iterativos para a fase de refinamento

O conhecimento de um intervalo $[a, b]$ que contém um único zero $\overline{x}$ de uma função real f é uma exigência de alguns métodos numéricos iterativos para a determinação de uma aproximação $\tilde{x}$ para $\overline{x}$. Para outros, a exigência é de uma aproximação inicial x_0 de $\overline{x}$. De todo modo, conforme vimos, para o cálculo dos zeros reais de f, os métodos iterativos pressupõem uma fase inicial de isolamento ou localização desses zeros. Reservamos este tópico para abordarmos especificamente esta primeira fase. Vale ressaltar que o sucesso nessa fase é fundamental para que possamos obter êxito também na segunda fase.

Nosso objetivo será, portanto, obter intervalos fechados disjuntos $[a, b]$ que contenham zeros isolados de f. Para tanto, necessitaremos estudar o comportamento de f, sendo úteis as seguintes ferramentas ou estratégias:

→ Tabelamento da função.

→ Análise gráfica da função.

Na aula 1, já deixamos claro que, para o trabalho nessa disciplina, será fundamental o uso de uma calculadora (científica, gráfica ou programável) e/ou de um *software* com ferramentas algébricas, numéricas e gráficas. Sugerimos uma calculadora científica para a computação numérica. Você pode obter uma na tela de seu computador. É uma ferramenta do sistema operacional *Windows* que é encontrada pelo caminho:

Iniciar - Todos os programas – Acessórios – Calculadora.

Se for possível, recomendamos ainda que vocês utilizem algum dos softwares que foram trabalhados na disciplina Informática Aplicada ao Ensino do segundo semestre. Finalmente, devemos dizer que os gráficos apresentados nesta e nas demais aulas serão gerados com o auxílio do *software Mathematica 6.0*.

Para o isolamento de zeros via tabelamento da função, serão úteis dois resultados do cálculo. Suas demonstrações podem ser encontradas na maioria dos livros de Cálculo. Veja, por exemplo, Lima (2004).

Teorema 1 (Teorema de Bolzano): *Seja* $f:\mathbb{R}\to\mathbb{R}$ *uma função contínua num intervalo fechado* $[a,b]$. *Se* $f(a)\cdot f(b)<0$, *então f tem pelo menos um zero no intervalo aberto* (a,b).

Este teorema diz que se uma função contínua em um intervalo fechado troca de sinal nos extremos desse intervalo, ela possui zeros reais nele. Graficamente, pela continuidade de *f*, este resultado parece ser bastante natural. Vejamos um exemplo:

Exemplo 4:

Seja $f:\mathbb{R}\to\mathbb{R}$, dada por $f(x)=\text{sen}(x)+\cos(x)$. Desde que *f* é contínua em $\mathbb{R}$, ela é contínua em qualquer intervalo $[a,b]$. Temos também que $f(-\pi)=\text{sen}(-\pi)+\cos(-\pi)=0-1=-1$ e $f(2\pi)=\text{sen}(2\pi)+\cos(2\pi)=0+1=1$.

Portanto, $f(-\pi)\cdot f(2\pi)=-1<0$. Logo, pelo teorema 1, f tem zeros no intervalo $(-\pi,2\pi)$. A figura 4, abaixo, mostra que *f* tem três zeros em $(-\pi,2\pi)$.

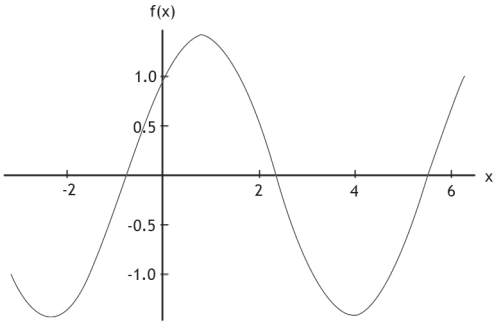


Figura 4: Gráfico de $f(x)=\text{sen}(x)+\cos(x)$ em $[-\pi,2\pi]$

VOCÊ SABIA?

Aqui, sen(x) e cos(x) são calculadas para x em radianos (rad) e não em graus (º). Nestes casos, ao usar a calculadora, você deve habilitar para o modo Radianos.

O Teorema de Bolzano, satisfeitas suas condições, garante a existência de zeros em um intervalo, mas não diz nada a respeito da quantidade deles. Pode haver apenas um (caso em que o zero estaria isolado), dois, três (como no Exemplo

VOCÊ SABIA?

A constante matemática e é conhecida como número de Euler (em homenagem ao matemático suíço Leonhard Euler) ou constante de Napier (em homenagem ao matemático escocês John Napier). Este número irracional é a base da função logaritmo natural e seu valor aproximado com 4 (valor usado acima) e com 30 casas decimais (dígitos após a vírgula) é, respectivamente:

$$e \cong 2{,}7183$$

e

$$e \cong 2{,}718281828459045235360287$$

4) ou até uma infinidade deles. Para garantir a unicidade do zero, é suficiente o seguinte teorema:

Teorema 2: *Sob as hipóteses do teorema 1, se a derivada* f' *de* f *existir e preservar o sinal no intervalo aberto* (a, b)*, então* f *tem um único zero em* (a, b).

Dizer que f' preserva o sinal em (a, b) é o mesmo que afirmar que

$f'(x) > 0, \forall x \in (a, b)$ ou $f'(x) < 0, \forall x \in (a, b)$.

Isso significa que a função f é, respectivamente, estritamente crescente ou estritamente decrescente no intervalo (a, b). Vejamos mais um exemplo:

Exemplo 5:

Seja $f : \mathbb{R} \to \mathbb{R}$, dada por $f(x) = -x + 2e^{-x}$. Desde que f é contínua em $\mathbb{R}$, ela é contínua em qualquer intervalo $[a, b]$. Temos também que

$$f(0) = -0 + 2e^{-0} = -0 + 2 \cdot 1 = 2$$

$$\text{e } f(3) = -3 + 2e^{-3} = -3 + \frac{2}{e^3} < -3 + \frac{2}{2{,}7182^3} < -2.9 .$$

Portanto, f muda de sinal nos extremos do intervalo $[0, 2]$. Logo, pelo Teorema 1, f tem zeros no intervalo $(0, 2)$. Por outro lado, temos

$f'(x) = -1 - 2e^{-x} = -1 - \frac{2}{e^x} < 0$, para todo $x \in \mathbb{R}$.

Assim, f' preserva o sinal em $(0, 2)$. Mais precisamente, $f'(x) < 0, \forall x \in (0, 2)$, o que implica que f é estritamente decrescente em $(0, 2)$. Logo, pelo Teorema 2, f tem um único zero no intervalo $(0, 2)$. A Figura 5, abaixo, comprova este fato.

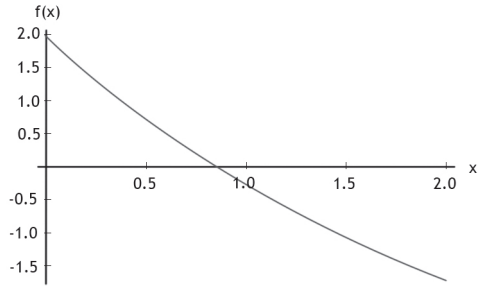


Figura 5: Gráfico de $f(x) = -x + 2e^{-x}$ em $[0, 2]$

Os Teoremas 1 e 2 são grandes aliados para o isolamento dos zeros reais de uma função real f via tabelamento da função. Esta estratégia consiste em construir uma tabela com valores de f para diversos valores de x e observar as mudanças de sinal de f e o sinal da derivada f' nos intervalos em que f mudou de sinal nos extremos. Algumas vezes, certas características próprias das funções ajudarão. Vamos isolar os zeros de algumas funções usando a estratégia de tabelamento?

Exemplo 6:

Seja $f:\mathbb{R}\to\mathbb{R}$, dada por $f(x)=x^4-9x^3-2x^2+120x-130$. Desde que f é contínua em $\mathbb{R}$, ela é contínua em qualquer intervalo $[a,b]$. Vamos construir uma tabela com valores de f para alguns valores de x e observar as mudanças de sinal de ocorridas. Temos

x	-10	-5	-4	-3	0	1	2	3	4	5	7	10
$f(x)$	17470	970	190	-184	-130	-20	46	50	-2	-80	-74	1870
SINAL	+	+	+	-	-	-	+	+	-	-	-	+

Pelas variações de sinal, podemos dizer que f tem zeros nos intervalos $[-4,-3]$, $[1,2]$, $[3,4]$ e $[7,10]$. Desde que f é um polinômio de grau 4, f tem no máximo 4 zeros reais distintos (este é um resultado que você deve ter visto na disciplina *Matemática Básica II*. Reveja-o). Portanto, podemos afirmar que f tem exatamente 4 zeros reais distintos e eles estão isolados nos intervalos listados acima.

Exemplo 7:

Seja $f:(0,+\infty)\to\mathbb{R}$, dada por $f(x)=\ln x+x\sqrt{x}$. Temos que f é contínua em $(0,+\infty)$, como produto e soma de funções contínuas. Logo, f é contínua em qualquer intervalo $[a,b]$ contido em $(0,+\infty)$. Vamos construir uma tabela com valores (ou valores aproximados) de f para alguns valores de x e observar as mudanças de sinal que ocorrem. Temos

x	0,01	0,1	0,5	1	2	3	5	10
$f(x)$	-9,60	-7,27	-5,34	-4,00	-1,48	1,29	7,79	28,93
SINAL	-	-	-	-	-	+	+	+

Pelas variações de sinal, podemos dizer que f tem zeros no intervalo $[2,3]$. A derivada de f está definida em $(0,+\infty)$ e é dada por

ATENÇÃO!

Você já deve ter esboçado gráficos de algumas funções na disciplina de Cálculo I. Sabe, portanto, que esta tarefa requer um estudo detalhado do comportamento da função, destacando-se a determinação de intervalos de crescimento e decrescimento, pontos de máximo e de mínimo, concavidade, pontos de inflexão, assíntotas horizontais e verticais, dentre outros. Isso envolve o estudo da função e de suas derivadas. O tabelamento de valores da função para alguns valores de x é também útil.

$$f'(x)=\frac{1}{x}+\frac{3\sqrt{x}}{2}.$$

Perceba que $f'(x)>0$ para todo $x>0$, ou seja, f é estritamente crescente em seu domínio de definição. Assim, f' preserva o sinal em $(2,3)$. Logo, podemos afirmar que f possui um único zero no intervalo $(2,3)$.

Além do tabelamento com a análise de mudanças de sinal da função, o isolamento dos zeros reais de uma função real f pode ser feito também por meio da análise gráfica da função. Para tanto, torna-se necessário esboçar o gráfico de f e obter intervalos que contenham as abscissas dos pontos em que o gráfico de f intercepta o eixo dos x.

Vejamos um primeiro exemplo. Neste apresentamos as ferramentas do cálculo para esboçar o gráfico. Entretanto, como dissemos, usaremos o *software Mathematica 6.0* para gerar os nossos gráficos.

Exemplo 8:

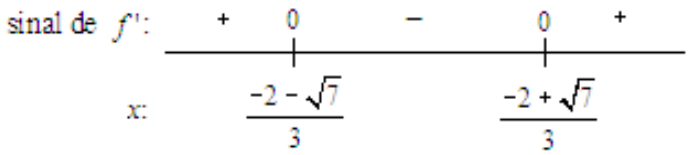
Seja $f:\mathbb{R}\to\mathbb{R}$, dada por $f(x)=x^3+2x^2-x-1$.

Temos

$$f'(x)=3x^2+4x-1$$

$$\Rightarrow \quad f'(x)=0\Leftrightarrow 3x^2+4x-1=0\Leftrightarrow x=\frac{-2-\sqrt{7}}{3} \text{ ou } x=\frac{-2+\sqrt{7}}{3}.$$

Logo, o sinal de f' é:

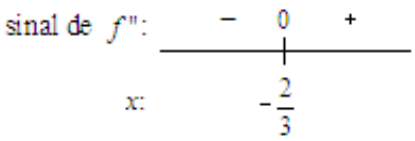


Portanto, f é crescente nos intervalos $\left(-\infty,\frac{-2-\sqrt{7}}{3}\right]$ e $\left[\frac{-2+\sqrt{7}}{3},+\infty\right)$ e é decrescente no intervalo $\left[\frac{-2-\sqrt{7}}{3},\frac{-2+\sqrt{7}}{3}\right]$. Os valores $x=\frac{-2-\sqrt{7}}{3}$ e $x=\frac{-2+\sqrt{7}}{3}$ são abscissas de pontos de máximo e de mínimo local de f, respectivamente.

Temos ainda

$$f''(x) = 6x + 4 \quad \Rightarrow \quad f''(x) = 0 \Leftrightarrow 6x + 4 = 0 \Leftrightarrow x = -\frac{2}{3}.$$

Logo, o sinal de f'' é:



Desse modo, a concavidade de *f* é voltada para baixo no intervalo $\left(-\infty, -\frac{2}{3}\right]$ e é voltada para cima no intervalo $\left[-\frac{2}{3}, +\infty\right)$. O valor $x = -\frac{2}{3}$ é abscissa de ponto de inflexão de *f*.

Temos também que f está definida e é contínua em $\mathbb{R}$ e que $\lim_{x\to-\infty} f(x) = -\infty$ e $\lim_{x\to+\infty} f(x) = +\infty$. Logo, *f* não possui assíntotas verticais nem horizontais.

Com essas informações, e com o auxílio da tabela seguinte com valores exatos (ou aproximados) de *f* para alguns valores de *x*, fica mais simples esboçar o gráfico de *f*:

x	$f(x)$
$-2,5$	$-1,625$
-2	1
$\frac{-2-\sqrt{7}}{3} \cong -1,5586$	1,6311
-1	1
$-\frac{2}{3} \cong -0,6667$	0,2593
$-0,5$	$-0,125$
0	-1
$\frac{-2+\sqrt{7}}{3} \cong 0,2153$	$-1,1126$
0,5	-0,875
1	1

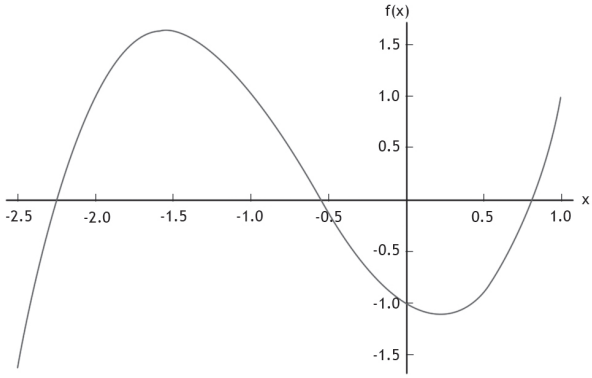


Figura 6: Gráfico de $f(x) = x^3 + 2x^2 - x - 1$ em $[-2,5; 1]$

ATENÇÃO!

Para descrever o intervalo $[-2,5; 1]$, usamos o separador ponto-e-vírgula (;) em vez de vírgula (,) como fazemos normalmente. Para evitar confusão, faremos isso sempre que algum dos extremos tiver parte fracionária (que precisa ser separada da parte inteira por vírgula).

GUARDE BEM ISSO!

O uso de um software matemático adequado torna a tarefa de esboçar os gráficos bem mais simples. Alguns desses softwares são *Mathematica, Maple, Graphmatica, Winplot,* dentre outros. Você deve ter trabalhado com o *Winplot* na disciplina de Informática Aplicada ao Ensino. Ele é um *software* livre e pode ser baixado do link http://www.baixaki.com.br/download/winplot.htm.

Podemos concluir que f tem um zero em cada um dos intervalos $[-2,5; 2]$, $[-0,6667; -0,5]$ e $[0,5; 1]$.

A menos que se use um software matemático, para certas funções, a tarefa de esboçar o gráfico não é nada fácil. Isso porque o estudo detalhado do comportamento de uma função f cuja expressão analítica seja mais complexa pode ser bastante laborioso. Em alguns desses casos, é mais conveniente, partindo da equação $f(x) = 0$, obter uma equação equivalente $f_1(x) = f_2(x)$, em que f_1 e f_2 sejam funções mais simples e de análise gráfica mais fácil. Os intervalos de isolamento dos zeros de f procurados podem ser obtidos considerando as abscissas dos pontos de intersecção dos gráficos de f_1 e f_2. De fato, se a é um zero de f, então:

$$f(a) = 0 \quad \Leftrightarrow \quad f_1(a) = f_2(a).$$

Logo, a é abscissa de um ponto comum dos gráficos de f_1 e f_2. Vejamos um exemplo:

Exemplo 9:

Seja $f : \mathbb{R} \to \mathbb{R}$, dada por $f(x) = -1 + x + x\cos(x)$. Temos que

$$-1+x+x\cos(x)=0 \Leftrightarrow x(1+\cos(x))=1 \Leftrightarrow 1+\cos(x)=\frac{1}{x}\cdot$$

Portanto, isolar os zeros de *f* é equivalente a obter intervalos cada um dos quais contendo a abscissa de um dos pontos de intersecção dos gráficos de f_1 e f_2 (figura 8), no qual $f_1(x)=1+\cos(x)$ e $f_2(x)=\frac{1}{x}$, que são mais simples de ser esboçados do que o gráfico de *f*.

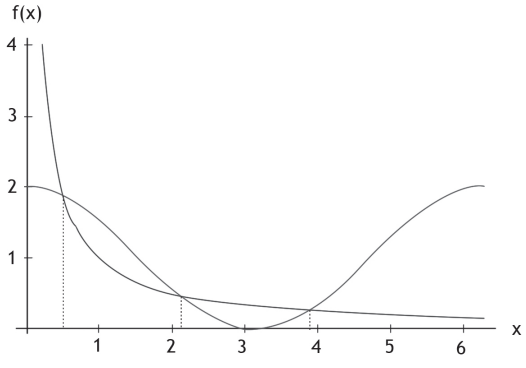


Figura 7: Gráficos de $f_1(x)=1+\cos(x)$ e $f_2(x)=\frac{1}{x}$ em $[0,2\pi]$.

GUARDE BEM ISSO!

Você deve esboçar os gráficos de f_1 e f_2 em um mesmo sistema de coordenadas cartesianas no plano para visualizar melhor os pontos de intersecção.

Dos gráficos de f_1 e f_2, podemos concluir que *f* tem um zero em cada um dos intervalos $[0,1]$, $[2;2,5]$ e $[3,5;4]$. Entretanto, não podemos afirmar que isolamos todos os zeros de *f*. Na verdade, *f* possui uma infinidade de zeros em $\mathbb{R}$.

O tabelamento e a análise gráfica da função são recursos complementares para o isolamento dos zeros. O trabalho com essas duas ferramentas simultaneamente pode tornar a fase de isolamento mais eficiente, permitindo obter intervalos de *amplitudes* bem pequenas.

GUARDE BEM ISSO!

Quanto menor for a amplitude do intervalo que contém o zero, mais eficiente será a fase de refinamento.

Agora você já sabe como fazer o isolamento dos zeros de uma função *f*. Na próxima aula, veremos métodos iterativos específicos para a fase refinamento. De acordo com Camponogara e Castelan Neto (2008, 33-34), tais métodos são de três tipos:

1. **Métodos de quebra:** requerem um intervalo fechado $[a,b]$ que contenha um único zero de *f* e tal que $f(a)\cdot f(b)<0$, ou seja, tal que a função troque de sinal nos extremos do intervalo. Então, partindo o intervalo em dois outros intervalos, verifica-se qual deles contém a raiz desejada.

Prossegue-se repetindo o procedimento com o subintervalo obtido.

2. **Métodos de ponto fixo:** Partindo de uma aproximação inicial x_0, constrói-se uma sequência $(x_j)_{j=1}^{n}$ na qual cada termo é obtido a partir do anterior por $x_{j+1} = g(x_j)$, em que g é uma função de iteração. Dependendo das propriedades de g, surgem diferentes tipos de métodos de ponto fixo, dentre eles o conhecido *Método de Newton*.
3. **Métodos de múltiplos passos:** Generalizam os métodos de ponto fixo. Constrói-se uma sequência $(x_j)_{j=1}^{n}$, utilizando vários pontos anteriores: x_j, x_{j-1}, ..., x_{j-p} para determinar o ponto x_{j+1}.

Sob certas condições, teremos que a raiz $\overline{x}$ será dada por $\overline{x} = \lim_{j\to\infty} x_j$, em que $(x_j)_{j\in\mathbb{N}}$ é a sequência gerada pelo método.

Nesta aula, conhecemos o problema de obter zeros de funções e vimos várias situações em que este problema aparece de forma contextualizada, caracterizando a importância deste problema nas mais diversas áreas. Abordamos também formas de localizar ou isolar os zeros reais de funções reais, um requisito necessário pelos métodos numéricos iterativos para a determinação de aproximações para os zeros de funções. Na próxima aula, apresentaremos métodos iterativos específicos para a fase de refinamento.

SAIBA MAIS!

Amplie seus conhecimentos consultando as referências e os sites citados. Para um maior aprofundamento, você deverá pesquisar também outras referências ou visitar outras páginas da internet. Abaixo, listamos algumas páginas interessantes que podem ajudá-lo nessa pesquisa. Bons estudos!

1. www.ime.usp.br/~asano/LivroNumerico/LivroNumerico.pdf
2. www.professores.uff.br/salete/imn/calnumI.pdf
3. http://www.das.ufsc.br/~camponog/Disciplinas/DAS-5103/LN.pdf

AULA 3

Método iterativos para celular zeros e funções

Olá aluno (a),

Esta é nossa terceira aula. Nela, continuaremos abordando o problema de encontrar zeros reais de funções reais. Veremos alguns dos principais métodos numéricos iterativos para obter tais zeros, destacando-se *método da bissecção, método da posição falsa, métodos do ponto fixo e método de Newton-Raphson.*

Objetivos

- Saber utilizar métodos numéricos iterativos
- Calcular aproximações para zeros reais de funções reais
- Estudar a convergência de alguns métodos
- Conhecer critérios de parada de algoritmos

TÓPICO 1

Métodos iterativos para refinamento de zeros: funcionamento e critérios de parada

OBJETIVOS

- Conhecer a ideia geral dos métodos iterativos para refinamento de zeros
- Apresentar fluxograma de funcionamento dos métodos iterativos
- Estabelecer critérios de proximidade

Neste primeiro tópico, conheceremos o *modus operandi* dos métodos iterativos para calcular zeros de funções. Mais precisamente, veremos como estes métodos fazem o refinamento da aproximação inicial obtida na fase de isolamento dos zeros, ou seja, como eles calculam aproximações para os zeros reais de uma função f que estejam suficientemente próximas dos zeros.

Na aula anterior, vimos que, utilizando aproximações anteriores para calcular as novas aproximações, um método numérico iterativo constrói uma sequência de aproximações $x_1, x_2, x_3, \ldots$ de um zero de f. Veremos que, sob certas condições, a sequência construída converge para o valor exato do zero de modo que, em um número finito de repetições do procedimento, é possível obter uma aproximação que satisfaça uma precisão prefixada.

Os métodos iterativos são compostos, basicamente, de pelo menos três módulos:

Inicialização: onde são fornecidos os dados iniciais (como aproximações iniciais ou intervalos iniciais) e/ou feitos alguns cálculos iniciais.

Atualização: aqui se calcula (geralmente, por meio de alguma fórmula) uma nova aproximação.

Parada: módulo que estabelece quando parar o processo iterativo.

O fluxograma seguinte mostra como os métodos iterativos fazem o refinamento dos zeros.

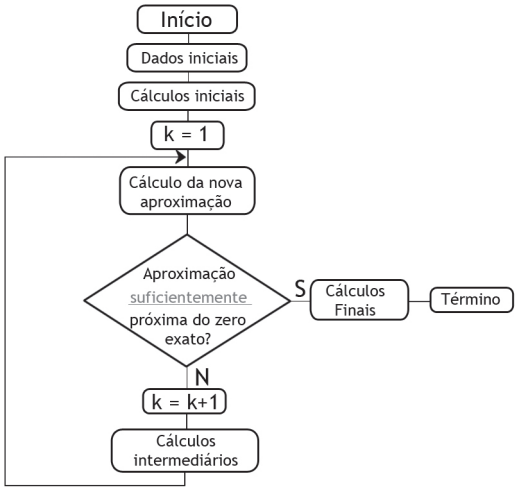


Figura 1: Fluxograma da fase de refinamento

A *inicialização* corresponde à fase de localização ou isolamento dos zeros e isto é o que vimos na aula 2. A *atualização* é o módulo que caracteriza cada método iterativo e corresponde à forma particular que cada um tem de calcular uma nova iteração. Este módulo é o nosso foco de estudo nesta aula. Antes, porém, falaremos um pouco mais sobre o módulo de *parada*.

VOCÊ SABIA?

Não podemos repetir um processo numérico iterativo infinitamente, ou seja, em algum momento, precisamos pará-lo. Para **parar** as iterações de um processo numérico iterativo, devemos adotar os chamados critérios de parada. Obviamente, esses critérios dependerão do problema a ser resolvido e da precisão que necessitamos obter na solução.

O diagrama de fluxo anterior sugere que os métodos iterativos, para obter um zero real de uma função f , fazem um teste de parada, dado pela pergunta:

A *aproximação atual está suficientemente próxima do zero exato de f*?

Mas, o que significa estar suficientemente próxima? Qual o significado de aproximação ou de zero aproximado? Especificamente, há várias formas de fazer o teste de parada do processo iterativo. Concentraremos-nos em quatro delas. Suporemos que $\overline{x}$ é um zero (exato) de f, e que x_k é a aproximação (zero aproximado) calculada na k-ésima iteração. Sejam ainda ε_1 e ε_2 precisões (tolerâncias) prefixadas.

1. $|\overline{x} - x_k| < \varepsilon_1$: a distância entre $\overline{x}$ e x_k é menor que ε_1, ou seja, $x_k - \varepsilon_1 < \overline{x} < x_k + \varepsilon_1$.

$|f(x_k)| < \varepsilon_2$: o valor da função em x_k dista no máximo ε_2 do valor 0, ou seja, $-\varepsilon_2 < f(x_k) < \varepsilon_2$.

2. $|x_k - x_{k-1}| < \varepsilon_1$: a distância entre dois iterados (aproximação calculada em uma iteração) consecutivos é menor que ε_1, ou seja, $x_{k-1} - \varepsilon_1 < x_k < x_{k-1} + \varepsilon_1$.
3. $k = N$: o número de iterações atingiu um limite máximo N preestabelecido.

Devemos fazer algumas observações:

Observação 1

Como efetuar o teste 1 se não conhecemos $\overline{x}$? Uma forma é reduzir o intervalo que contém o zero a cada iteração (RUGGIERO e LOPES, 1996, p. 39). Se obtivermos um intervalo $[a, b]$ de tamanho menor que ε_1 contendo $\overline{x}$, então qualquer ponto nesse intervalo pode ser tomado como zero aproximado. Assim, basta exigir que x_k esteja no intervalo $[a, b]$. Perceba que a distância entre $\overline{x}$ e x_k é menor que a distância entre a e b. A figura 2 ilustra esta situação. Simbolicamente, temos

Se $[a, b]$ é tal que $b - a < \varepsilon_1$ e $\overline{x} \in [a, b]$, então $|\overline{x} - x_k| < \varepsilon_1$, $\forall x_k \in [a, b]$.

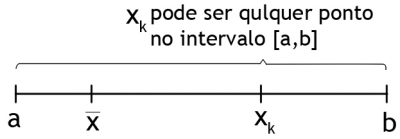


Figura 2: Critério de parada $|\overline{x} - x_k| < \varepsilon_1$

Observação 2

Devemos tomar cuidado com o teste de parada $|f(x_k)| < \varepsilon_2$ dado em 2, pois, a menos que conheçamos bem o comportamento de f, o fato de ele ser satisfeito não implica necessariamente que x_k esteja próximo do zero procurado. A função $f : (0, +\infty) \to \mathbb{R}$, dada por $f(x) = \dfrac{\text{Log}\, x}{x}$, por exemplo, possui um único zero $\overline{x} = 1$. Entretanto, calculando f para x = 10, 100, 1000, 10000, 100000, ..., obteremos, respectivamente: 0.1, 0.02, 0.003, 0.0004, 0.00005, ..., isto é, quanto mais distante estamos de $\overline{x}$, menor é o valor de $f(x)$.

Observação 3

O teste de parada em 3 também devemos ser visto com cautela, pois $|x_k - x_{k-1}| < \varepsilon_1$ não implica necessariamente que $|\overline{x} - x_k| < \varepsilon_1$. Isso é ilustrado na

figura 3, em que x_k e x_{k-1} são próximos sem que $\overline{x}$ e x_k também sejam próximos.

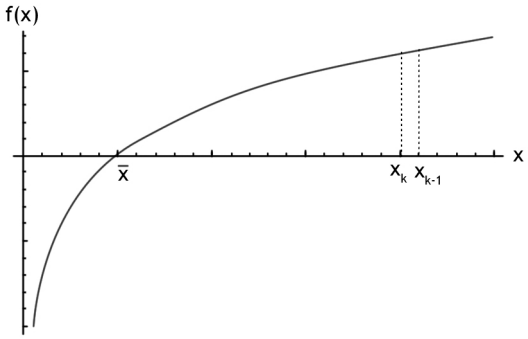


Figura 3 - Critério de parada $|x_k - x_{k-1}| < \varepsilon_1$

Observação 4

Dependendo da ordem de grandeza dos números envolvidos, devemos usar o teste do erro relativo, quando as desigualdades em 1, 2 e 3 seriam, respectivamente:

1. $\dfrac{|\overline{x} - x_k|}{|x_k|} < \varepsilon_1$.
2. $\dfrac{|f(x_k)|}{L} < \varepsilon_1$, em que $L = |f(x)|$ para algum x em uma vizinhança de $\overline{x}$ (RUGGIERO e LOPES, 1996, p. 40).
3. $\dfrac{|x_k - x_{k-1}|}{|x_k|} < \varepsilon_1$.

Observação 5

Ao contrário do que ocorre com os outro três, o teste de parada em 4 ($k = N$) que estipula um número máximo de iterações, não pode ser visto como um critério de proximidade propriamente dito. Ele é usado para evitar que o processo iterativo entre em looping, ou seja, ficar se repetindo ciclicamente sem parar. O looping pode ocorrer devido a vários fatores: erros de arredondamento, erros no processo iterativo, inadequação do processo iterativo ao problema, dente outros.

Observação 6

O ideal seria parar o processo com uma aproximação x_k que satisfizesse os critérios 1 e 2 simultaneamente. Isso significaria estar próximo do zero exato $\overline{x}$ pela distância e ter também o valor da função na aproximação próximo de zero. Entretanto, pode ocorrer que um critério seja satisfeito sem que os outros sejam. Esse procedimento será ilustrado nas figuras 4a e 4b. Na figura 4a, temos uma

situação em que o critério 1 é satisfeito, mas o 2 não. Na figura 4b, ocorre a situação inversa.

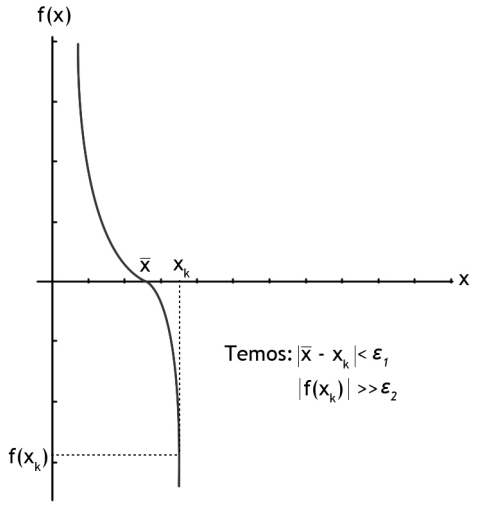


Figura 4a - Critério 1 é satisfeito, mas critério 2 não

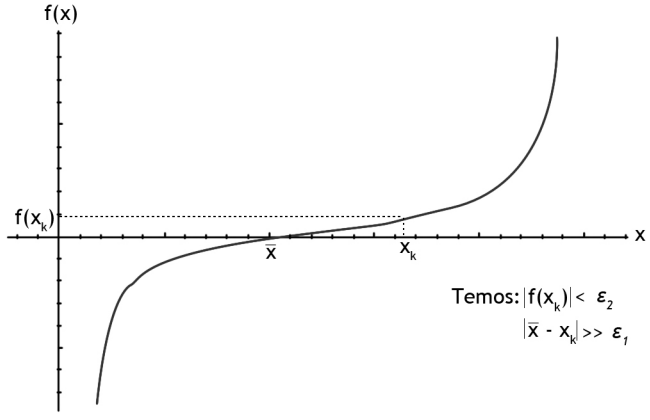


Figura 4b - Critério 2 é satisfeito, mas critério 1 não

Vimos a forma como os métodos iterativos operam para calcular zeros de funções e estabelecemos os principais critérios de parada para estes processos. Agora você está preparado para a parte central desta aula: o modo como cada método iterativo faz o cálculo de uma nova aproximação. Então, vamos ao primeiro método.

TÓPICO 2

Método da bissecção e método da posição falsa

OBJETIVOS

- Compreender o funcionamento do método da bissecção e da posição falsa
- Calcular aproximações para zeros de funções
- Fazer estimativas do número de iterações

A partir deste tópico, estudaremos o módulo de atualização, ou seja, a forma como cada método iterativo específico faz o refinamento dos zeros. Este módulo é o que caracteriza e dá nome a cada método, correspondendo ao cálculo, a partir de iterações anteriores, de uma nova iteração. Iniciamos com o *método da bissecção*, também chamado de *método da dicotomia*.

O método da bissecção está na categoria dos métodos de quebra (reveja as categorias de métodos vista no final da aula 2). Portanto, para determinar uma aproximação para o zero de uma função f:

Satisfeitas as condições requeridas, o *método da bissecção* opera reduzindo a amplitude do intervalo que contém o zero até obter um intervalo $[a, b]$ de tamanho menor que ε, ou seja, tal que $b - a < \varepsilon$, em que ε é uma precisão prefixada. Desse modo, conforme indicado na

VOCÊ SABIA?

Inspirado no teorema de Bolzano, o método da bissecção é um método bem intuitivo para achar o zero de uma função f em um intervalo que contém um único zero de f. A cada iteração, o método da bissecção obtém um novo intervalo com um tamanho igual à metade do tamanho do intervalo anterior.

GUARDE BEM ISSO!

O método da bissecção requer um intervalo fechado $[a, b]$ em que f seja contínua tal que $f(a) \cdot f(b) < 0$ (a função troca de sinal nos extremos do intervalo). Por questões de simplicidade, exigi-se ainda que o zero de f em $[a, b]$ seja único.

observação 1, podemos escolher um ponto qualquer x_k no intervalo final $[a, b]$ para ser a aproximação do zero exato $\overline{x}$ que teremos o critério de parada 1 satisfeito.

Tecnicamente, a redução da amplitude do intervalo faz-se pela sucessiva divisão de $[a, b]$ ao meio, ou seja, pelo ponto médio $x_M = \frac{a+b}{2}$, mantendo a cada iteração o subintervalo que contém o zero desejado e desprezando o outro subintervalo. A escolha do subintervalo que será mantido é feita de modo simples: calculamos o valor da função f no ponto médio $x_M = \frac{a+b}{2}$. Temos, assim, três possiblidades:

1. $f(x_M) = 0$. Nesse caso x_M é o zero (exato) de f e não temos mais nada a fazer. Em geral, não é isso que ocorre.
2. $f(a) \cdot f(x_M) < 0$. Aqui o zero de f está entre a e x_M. O intervalo a ser mantido será, então, $[a, x_M]$.
3. $f(a) \cdot f(x_M) > 0$. Nesse caso, desde que $f(a)$ e $f(b)$ têm sinais opostos, teremos também $f(x_M) \cdot f(b) < 0$. Assim, o zero de f está entre x_M e b, e o intervalo a ser mantido será, então, $[x_M, b]$.

De modo mais simplificado, temos o esquema seguinte:

Se $f(x_M) = 0$, então $\overline{x} = x_M$

Se $f(a) \cdot f(x_M)$ $\begin{cases} < 0, \text{ então } b = x_M \\ > 0, \text{ então } a = x_M \end{cases}$

Em termos de algoritmo, o método da bissecção pode ser descrito como

Dados um intervalo $[a_0, b_0]$, uma função real de uma variável real f contínua em $[a_0, b_0]$ tal que $f(a_0) \cdot f(b_0) < 0$,uma precisão ε e $N \in \mathbb{N}$.

$k = 0$.

Enquanto $b_k - a_k > \varepsilon$ e $k < N$, faça

$x_k = \frac{a_k + b_k}{2}$.

Se $f(a_k) \cdot f(x_k) = 0$, faça $\tilde{x} = x_k$. PARE.

Se $f(a_k) \cdot f(x_k) < 0$, faça $a_{k+1} = a_k$ e $b_{k+1} = x_k$.

Caso contrário, faça $a_{k+1} = x_k$ e $b_{k+1} = b_k$.

$k = k + 1$.

Faça $\tilde{x} = \frac{a_k + b_k}{2}$. PARE.

Terminado o processo iterativo, teremos um intervalo $[a, b]$ que contém o

zero $\overline{x}$ de f e, caso $k < N$, encontraremos também uma aproximação $\tilde{x}$ de $\overline{x}$ que satisfaz o critério de parada 1, ou seja, tal que $|\overline{x} - \tilde{x}| < \varepsilon$. Uma interpretação geométrica do método da bissecção é dada na figura seguinte.

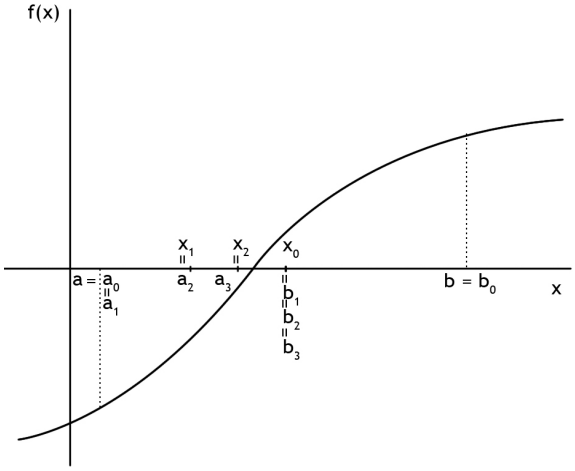


Figura 5- Método da bissecção. Fonte: Adaptado de Ruggiero e Lopes (1996, p. 41).

Para exemplificar, vamos usar o método da bissecção para obter uma aproximação para $\sqrt{2}$ com erro inferior a 10^{-2}.

Exercício resolvido 1:

Encontre uma aproximação para $\sqrt{2}$ com erro inferior a 10^{-2} pelo método da bissecção.

Solução:

Este problema é equivalente a determinar uma aproximação para o zero de $f(x) = x^2 - 2$ com erro inferior a 10^{-2}.

Temos $f(1) = -1$ e $f(2) = 2$. Assim, $f(1) \cdot f(2) = -2 < 0$ e, uma vez que f é contínua no intervalo $[1, 2]$, podemos garantir f tem zeros nesse intervalo. Como $f'(x) = 2x$, o que implica que $f'(x) > 0$ para todo $x \in (1, 2)$, temos que o zero de f no intervalo $[1, 2]$ é único.

k	a_k	b_k	$b_k - a_k$	x_k	$f(x_k)$
0	1	2	1	1,5	0,25
1	1	1,5	0,5	1,25	-0,43
2	1,25	1,5	0,25	1,375	-0,109375
3	1,375	1,5	0,125	1,4375	0,06640625
4	1,375	1,4375	0,0625	1,40625	-0,0224609375
5	1,40625	1,4375	0,03125	1,421875	0,021728515625

6	1,40625	1,421875	0,015625	1,4140625	-0,00042724609375
7	1,4140625	1,421875	0,0078125		

Tabela 1: Método da bissecção para calcular $\sqrt{2}$ com erro inferior a 10^{-2}.

Portanto, depois de 7 iterações ($k = 0, 1, 2, ..., 6$), teremos um intervalo $[a_7, b_7] = [1,4140625;\ 1,421875]$ com tamanho $b_7 - a_7 = 0,0078125 < 10^{-2}$. Assim, como indicado no algoritmo, fazendo

$$\tilde{x} = \frac{a_7 + b_7}{2} = \frac{1,4140625 + 1,421875}{2} = 1,41796875,$$

obteremos uma aproximação $\tilde{x}$ de $\sqrt{2}$ com erro inferior a 10^{-2}, ou seja, coincidindo com o valor de $\sqrt{2}$ até pelo menos duas casas decimais (casas depois da vírgula). Compare com o valor de $\sqrt{2}$ exibido a seguir com 10 casas decimais.

$\sqrt{2} = 1,41421356237...$.

Para uma melhor visualização dos intervalos obtidos a cada iteração, observe o esquema seguinte:

$$\begin{array}{lcll}
 & & f(a_0) < 0 & \overline{x} \in [a_0, x_0] \\
k = 0 & \Rightarrow & f(b_0) > 0 & \Rightarrow \quad a_1 = a_0 \\
 & & f(x_0) > 0 & b_1 = x_0 \\
 & & f(a_1) < 0 & \overline{x} \in [x_1, b_1] \\
k = 1 & \Rightarrow & f(b_1) > 0 & \Rightarrow \quad a_2 = x_1 \\
 & & f(x_1) < 0 & b_2 = b_1 \\
 & & f(a_2) < 0 & \overline{x} \in [x_2, b_2] \\
k = 2 & \Rightarrow & f(b_2) > 0 & \Rightarrow \quad a_3 = x_2 \\
 & & f(x_2) < 0 & b_3 = b_2 \\
 & & f(a_3) < 0 & \overline{x} \in [a_3, x_3] \\
k = 3 & \Rightarrow & f(b_3) > 0 & \Rightarrow \quad a_4 = a_3 \\
 & & f(x_3) > 0 & b_4 = x_3 \\
 & & f(a_4) < 0 & \overline{x} \in [x_4, b_4] \\
k = 4 & \Rightarrow & f(b_4) > 0 & \Rightarrow \quad a_5 = x_4 \\
 & & f(x_4) < 0 & b_5 = b_4 \\
 & & f(a_5) < 0 & \overline{x} \in [a_5, x_5] \\
k = 5 & \Rightarrow & f(b_5) > 0 & \Rightarrow \quad a_6 = a_5 \\
 & & f(x_5) > 0 & b_6 = x_5 \\
 & & f(a_6) < 0 & \overline{x} \in [x_6, b_6] \\
k = 6 & \Rightarrow & f(b_6) > 0 & \Rightarrow \quad a_7 = x_6 \\
 & & f(x_6) < 0 & b_7 = b_6
\end{array}$$

E se desejássemos uma aproximação para $\sqrt{2}$ com erro inferior a 10^{-5}, ou seja, coincidindo com o valor de $\sqrt{2}$ até pelo menos cinco casas decimais? Seria possível dizer quantas iterações precisaríamos executar?

Evidentemente, para uma maior precisão, o processo de redução dos intervalos deverá prosseguir. Felizmente, é possível precisar a *priori* (sem precisar realizar a experiência) quantas iterações serão executadas pelo método da bissecção até obter uma aproximação para o zero de uma função com uma precisão prefixada.

Teorema 1: *Dado um intervalo* $I_0 = [a_0, b_0]$ *que contém um único zero* $\overline{x}$ *de uma função contínua* $f : \mathbb{R} \to \mathbb{R}$ *e uma precisão prefixada* $\varepsilon > 0$, *após k iterações, k satisfazendo* $k > \dfrac{\text{Log}(b_0 - a_0) - \text{Log}(\varepsilon)}{\text{Log}(2)}$, *o método da bissecção obtém um intervalo* $I_k = [a_k, b_k]$ *contendo o zero* $\overline{x}$ *de f e tal que qualquer que seja a aproximação* $\tilde{x}$ *escolhida em* I_k, $|\overline{x} - \tilde{x}| < \varepsilon$.

De fato, uma vez que a amplitude de cada novo intervalo é igual à metade da amplitude do intervalo anterior, temos

$$b_k - a_k = \frac{b_{k-1} - a_{k-1}}{2} = \frac{b_{k-2} - a_{k-2}}{2^2} = \cdots = \frac{b_0 - a_0}{2^k}$$

Assim,

$$\begin{aligned} b_k - a_k < \varepsilon \quad &\Leftrightarrow \quad \frac{b_0 - a_0}{2^k} < \varepsilon \\ &\Leftrightarrow \quad 2^k > \frac{b_0 - a_0}{\varepsilon} \\ &\Leftrightarrow \quad k \cdot \text{Log}(2) > \text{Log}\left(\frac{b_0 - a_0}{\varepsilon}\right) \\ &\Leftrightarrow \quad k > \frac{\text{Log}(b_0 - a_0) - \text{Log}(\varepsilon)}{\text{Log}(2)}. \end{aligned}$$

Agora, voltando ao nosso exemplo, podemos calcular o número de mínimo de iterações para ter a garantia de uma aproximação para $\sqrt{2}$ no intervalo $[1, 2]$ com erro inferior a 10^{-5}. Temos

$$k > \frac{\text{Log}(2-1) - \text{Log}(10^{-5})}{\text{Log}(2)} = \frac{5}{\text{Log}(2)} \cong \frac{5}{0{,}3010} \cong 16{,}61.$$

Portanto, serão necessárias pelo menos 17 iterações para garantir uma aproximação para $\sqrt{2}$ com erro inferior a 10^{-5}.

Calcular todas essas iterações daria um trabalhão, você não acha?

Você já sabe que outra preocupação que devemos ter é com a convergência do método. No caso do método da bissecção, uma vez que a amplitude do intervalo que contém o zero é reduzida pela metade a cada iteração, pode parecer bem intuitivo que a sequência (x_k) gerada convirja para o zero exato $\overline{x}$.

Entretanto, para termos a garantia da eficácia do método da bissecção, a prova analítica de sua convergência é imprescindível. Você pode ver tal prova em Ruggiero e Lopes (1996, p. 44-46).

Na mesma categoria dos métodos de quebra, está o método da posição falsa ou método das cordas. Como o método da bisecção, este método também requer um intervalo fechado $[a, b]$, em que f seja contínua tal que $f(a) \cdot f(b) < 0$. Sob estas condições, para determinar uma aproximação para o zero de f, o método da posição falsa particiona (quebra) o intervalo $[a, b]$ de um modo diferente.

Enquanto no método da bissecção é feita uma média aritmética simples (sem ponderação) dos valores a e b, o método da posição falsa faz uma média ponderada desses valores com pesos $|f(b)|$ e $|f(a)|$, respectivamente, ou seja, o ponto x que divide o intervalo $[a, b]$ de certa iteração é dado por

$$x = \frac{a \cdot |f(b)| + b \cdot |f(a)|}{|f(b)| + |f(a)|} = \frac{a \cdot f(b) - b \cdot f(a)}{f(b) - f(a)}.$$

A segunda igualdade segue do fato que $f(a)$ e $f(b)$ têm sinais contrários. Há uma interpretação geométrica para o ponto x. Ele é o ponto de intersecção da reta que passa pelos pontos $(a, f(a))$ e $(b, f(b))$ com o eixo das abscissas, como ilustra a figura seguinte.

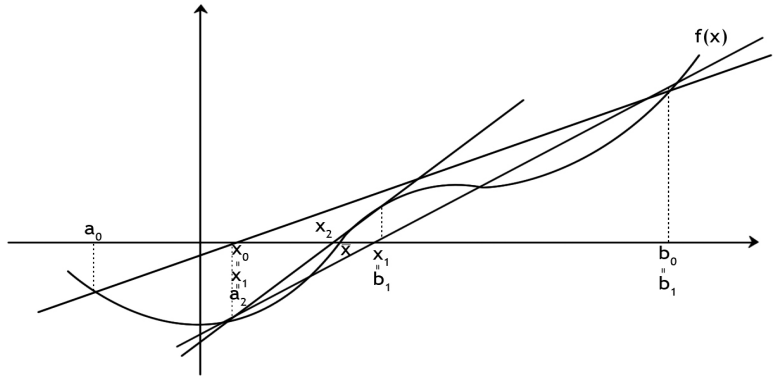


Figura 6 - Método da posição falsa. Fonte: Adaptado de Ruggiero e Lopes (1996, p. 49)

Desse modo, o método da *posição falsa* leva em conta as informações dos valores da função. Isso parece lógico, uma vez que, se $f(a)$ estiver mais próximo de zero do que $f(b)$, é de se esperar que o zero de *f* esteja mais próximo de a do que de b, e vice-versa. Isso é o que ocorre, por exemplo, para funções afins. Na verdade, o que se faz no método da posição falsa é substituir *f* no intervalo $[a, b]$ de cada iteração por uma reta.

GUARDE BEM ISSO!

A diferença ente os métodos da bissecção e da posição falsa é a forma de dividir o intervalo $[a, b]$ a cada iteração. No método da bissecção, quebra-se o intervalo ao meio, enquanto no método da posição falsa se toma o ponto de intersecção da reta que une os pontos $(a, f(a))$ e $(b, f(b))$ com o eixo x.

Quanto ao critério de parada, no método da posição falsa, além da parada pelo critério 1, $b_k - a_k < \varepsilon_1$, paramos também se $|f(x_k)| < \varepsilon_2$, pois isso pode ocorrer sem que o intervalo seja suficientemente pequeno. Finalizamos este tópico com um exemplo:

Exercício resolvido 2:

Aplicar o método da posição falsa para encontrar uma aproximação para o zero de $f(x) = 3\sqrt{x} + \ln x - 4$ no intervalo $[1, 2]$ com precisões $\varepsilon_1 = \varepsilon_2 = 10^{-4}$. Fazer arredondamentos e usar 5 casas decimais.

Solução:

$f(1) = -1$ e $f(2) = 3\sqrt{2} + \ln 2 - 4 \cong 0,93579$. Assim, $f(1) \cdot f(2) < 0$ e, uma vez que f é contínua no intervalo $[1, 2]$, podemos garantir f tem zeros nesse intervalo. Como $f'(x) = \frac{1}{x} + \frac{3}{2\sqrt{x}}$, o que implica que $f'(x) > 0$ para todo $x \in (1, 2)$, temos que o zero de *f* no intervalo $[1, 2]$ é único.

k	a_k	b_k	$b_k - a_k$	x_k	$f(x_k)$
0	1,00000	2,00000	1,00000	1,51658	0,11094
1	1,00000	1,51658	0,51658	1,46499	0,01295
2	1,00000	1,46499	0,46499	1,45905	0,00152
3	1,00000	1,45905	0,45905	1,45835	0,00017
4	1,00000	1,45835	0,45835	1,45827	0,00002

Tabela 1: Método da posição falsa para o zero de $f(x) = 3\sqrt{x} + \ln x - 4$ em $[1, 2]$ com precisões $\varepsilon_1 = \varepsilon_2 = 10^{-4}$

Observe o cálculo de x_k e de $f(x_k)$ em cada iteração:

$$
\begin{aligned}
& & x_0 &= \frac{1{,}00000 \cdot f(2{,}00000) - 2{,}00000 \cdot f(1{,}00000)}{f(2{,}00000) - f(1{,}00000)} & \\
k = 0 \;\Rightarrow\; & & &\cong \frac{1{,}00000 \cdot 0{,}93579 - 2{,}00000 \cdot (-1{,}00000)}{0{,}93579 - (-1{,}00000)} & \Rightarrow f(x_0) \cong 0{,}11094 \\
& & &\cong 1{,}51658 &
\end{aligned}
$$

$$
\begin{aligned}
& & x_1 &= \frac{1{,}00000 \cdot f(1{,}51658) - 1{,}51658 \cdot f(1{,}00000)}{f(1{,}51658) - f(1{,}00000)} & \\
k = 1 \;\Rightarrow\; & & &\cong \frac{1{,}00000 \cdot 0{,}11094 - 1{,}51658 \cdot (-1{,}00000)}{0{,}11094 - (-1{,}00000)} & \Rightarrow f(x_1) \cong 0{,}01295 \\
& & &\cong 1{,}46499 &
\end{aligned}
$$

$$
\begin{aligned}
& & x_2 &= \frac{1{,}00000 \cdot f(1{,}46499) - 1{,}46499 \cdot f(1{,}00000)}{f(1{,}46499) - f(1{,}00000)} & \\
k = 2 \;\Rightarrow\; & & &\cong \frac{1{,}00000 \cdot 0{,}01295 - 1{,}46499 \cdot (-1{,}00000)}{0{,}01295 - (-1{,}00000)} & \Rightarrow f(x_2) \cong 0{,}00152 \\
& & &\cong 1{,}45905 &
\end{aligned}
$$

$$
\begin{aligned}
& & x_3 &= \frac{1{,}00000 \cdot f(1{,}45905) - 1{,}45905 \cdot f(1{,}00000)}{f(1{,}45905) - f(1{,}00000)} & \\
k = 3 \;\Rightarrow\; & & &\cong \frac{1{,}00000 \cdot 0{,}00152 - 1{,}45905 \cdot (-1{,}00000)}{0{,}00152 - (-1{,}00000)} & \Rightarrow f(x_3) \cong 0{,}00017 \\
& & &\cong 1{,}45835 &
\end{aligned}
$$

$$
\begin{aligned}
& & x_4 &= \frac{1{,}00000 \cdot f(1{,}45835) - 1{,}45835 \cdot f(1{,}00000)}{f(1{,}45835) - f(1{,}00000)} & \\
k = 4 \;\Rightarrow\; & & &\cong \frac{1{,}00000 \cdot 0{,}00017 - 1{,}45835 \cdot (-1{,}00000)}{0{,}00017 - (-1{,}00000)} & \Rightarrow f(x_4) \cong 0{,}00002 \\
& & &\cong 1{,}45827 &
\end{aligned}
$$

Portanto, depois de 5 iterações ($k = 0, 1, 2, 3, 4$), temos uma aproximação

$$\tilde{x} = x_4 = 1{,}45827$$

que satisfaz a precisão prefixada, pois

$f(x_4) = f(1{,}45827) \cong 0{,}00002 \Rightarrow \left| f(x_4) \right| < \varepsilon_2 = 10^{-4}$.

Neste caso, a parada se deu pelo valor da função em x_4 ser próximo de 0 e não pela distância entre $\overline{x}$ e x_4 ser suficientemente pequena.

Em termos de comparação, para obter uma aproximação com a precisão requerida pelo método da bissecção para este exemplo, seriam necessárias:

$$k > \frac{\text{Log}(2-1) - \text{Log}(10^{-4})}{\text{Log}(2)} = \frac{4}{\text{Log}(2)} \cong \frac{4}{0,3010} \cong 13,29 \text{ iterações,}$$

ou seja, pelo menos 14 iterações, bem mais que pelo método da posição falsa.

Vimos o funcionamento dos métodos da bissecção e da posição falsa. Métodos mais sofisticados serão estudados no próximo tópico.

TÓPICO 3

Métodos de ponto fixo: método de Newton-Raphson

OBJETIVOS

- Compreender o funcionamento dos métodos de ponto fixo
- Conhecer o método de Newton-Raphson
- Calcular aproximações para zeros de funções

Neste tópico, discutiremos a determinação de aproximações para zeros de funções através dos *métodos de ponto fixo,* denominados também *métodos de iteração linear*. Sabemos que os métodos de quebra, como o método da bissecção e o método da posição falsa, necessitam da existência de um intervalo no qual a função troca de sinal. Entretanto nem sempre é possível satisfazer este requisito.

Imagine uma função *f* tal que para todo x do seu domínio $f(x) \geq 0$ ou $f(x) \leq 0$. Evidentemente f pode possuir zeros reais, entretanto não existem intervalos em que f troque de sinal. Nesses casos, aproximações para os possíveis zeros de f não poderiam ser obtidas por meio do método da bissecção ou do método da posição falsa, sendo necessários outros métodos. Uma boa saída nesses casos ou mesmo em qualquer situação que satisfaça certas restrições que veremos são os métodos de ponto fixo. Basicamente, estes métodos funcionam da seguinte maneira (ASANO e COLLI, 2007):

1. Dada a função f da qual se procura um zero $\overline{x}$, "arranja-se" uma função auxiliar g que deve satisfazer certas características (veremos como achar uma tal função).
2. Arrisca-se um "palpite" de uma aproximação inicial x_0 e, a partir desse palpite, constrói-se uma sequência de aproximações x_0, x_1, x_2, ..., na qual a aproximação x_{k+1} depende da aproximação x_k pela relação $x_{k+1} = g(x_k)$.

3. Para-se o processo, tomando algum dos x_k como aproximação de $\overline{x}$, quando algum critério de parada para alguma precisão prefixada for satisfeito.

A função g é chamada *função de iteração* para a equação $f(x)=0$. Como obter uma função de iteração?

Pela forma como é construída a sequência x_k, uma condição necessária para que o método funcione é que $\overline{x}$ seja um ponto fixo de g, ou seja,

$$g(\overline{x})=\overline{x}\,.$$

Dada uma função $\varphi:\mathbb{R}\to\mathbb{R}$, um número real a tal que $\varphi(a)=a$ é chamado ponto fixo de φ. Geometricamente, um ponto fixo de φ corresponde à abscissa de um ponto de intersecção do gráfico de φ com a reta $y=x$ (diagonal dos quadrantes ímpares). Na figura abaixo, por exemplo, vemos 2 pontos fixos da função φ (aqui, as raízes de φ não nos interessam).

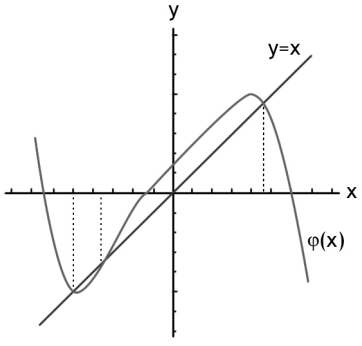


Figura 7: Pontos fixos de uma função φ.

Os métodos de ponto fixo transformam o problema de obter zeros de f em obter pontos fixos de g, com g sendo uma função de iteração para a equação $f(x)=0$, pela equivalência

$$f(x)=0 \Leftrightarrow x=g(x)\,.$$

Não é difícil introduzir uma função de iteração g para a equação $f(x)=0$. Vejamos um exemplo:

Exemplo 1:

Considere a equação $x^2-2x-3=0$, ou seja, $f(x)=0$ com $f(x)=x^2-2x-3$. Vamos obter algumas funções de iteração para $f(x)=0$. Para isso, basta obtermos uma equação equivalente do tipo $x=g(x)$. Temos:

$$x^2 - 2x - 3 = 0 \Leftrightarrow x = x^2 - x - 3 \quad \Rightarrow \quad g_1(x) = x^2 - x - 3$$

$$x^2 - 2x - 3 = 0 \Leftrightarrow x = \pm\sqrt{2x+3} \ (\text{se } 2x+3 \geq 0) \quad \Rightarrow \quad g_2(x) = \pm\sqrt{2x+3}$$

$$x^2 - 2x - 3 = 0 \Leftrightarrow x = 2 + \frac{3}{x} \ (\text{se } x \neq 0) \quad \Rightarrow \quad g_3(x) = 2 + \frac{3}{x}$$

$$x^2 - 2x - 3 = 0 \Leftrightarrow x = \frac{3}{x-2} \ (\text{se } x - 2 \neq 0) \quad \Rightarrow \quad g_4(x) = \frac{3}{x-2}$$

Em geral, há muitos modos de expressar $f(x) = 0$ na forma. Basta considerarmos $g(x) = x + A(x)f(x)$, para qualquer $A(x)$ que satisfaça $A(\overline{x}) \neq 0$, em que $\overline{x}$ é um ponto fixo de g ou, equivalentemente, um zero de *f*.

Exemplo 2:

Voltemos à equação $x^2 - 2x - 3 = 0$ do exemplo 1. Por ser uma equação quadrática, suas raízes podem ser obtidas analiticamente pela fórmula de Bhaskara e valem –1 e 3. Entretanto, para exercitarmos a aplicação dos métodos de ponto fixo, vamos tentar obter a raiz 3, usando duas das funções de iteração obtidas no exemplo 1 e partindo de uma aproximação inicial $x_0 = 1{,}5$.

Para $g_3(x) = 2 + \frac{3}{x}$, temos

$x_0 = 1{,}5$.

$x_1 = g_3(x_0) = 2 + \frac{3}{1{,}5} = 4$.

$x_2 = g_3(x_1) = 2 + \frac{3}{4} = 2{,}75$.

$x_3 = g_3(x_2) = 2 + \frac{3}{2{,}75} \cong 3{,}0909090909$.

$x_4 = g_3(x_3) = 2 + \frac{3}{3{,}0909090909} \cong 2{,}9705882353$.

$x_5 = g_3(x_4) = 2 + \frac{3}{2{,}9705882353} \cong 3{,}0099009901$.

$x_6 = g_3(x_5) = 2 + \frac{3}{3{,}0099009901} \cong 2{,}9967105263$.

$x_7 = g_3(x_6) = 2 + \frac{3}{2{,}9967105263} \cong 3{,}0010976948$.

$\vdots$

Vemos que o processo parece convergir para a raiz 3. Agora, para $g_1(x) = x^2 - x - 3$, temos:

$x_0 = 1,5$.

$x_1 = g_1(x_0) = 1,5^2 - 1,5 - 3 = -2,25$.

$x_2 = g_1(x_1) = (-2,25)^2 - (-2,25) - 3 = 4,3125$.

$x_3 = g_1(x_2) = 4,3125^2 - 4,3125 - 3 = 11,28515625$.

$x_4 = g_1(x_3) = 11,28515625^2 - 11,28515625 - 3 \cong 113,0695953369$.

$x_5 = g_3(x_4) = 113,0695953369^2 - 113,0695953369 - 3 \cong 12668,6637943134$.

$x_6 = g_3(x_5) = 12668,6637943134^2 - 12668,6637943134 - 3 \cong 1,6048237067 \times 10^8$.

$x_7 = g_3(x_6) = (1,6048237067 \times 10^8)^2 - 1,6048237067 \times 10^8 - 3 \cong 2,5754591135 \times 10^{16}$

$\vdots$

Vemos que o processo parece divergir (não convergir) da raiz 3.

O exemplo 2 mostra que não é para qualquer escolha da função de iteração para $f(x) = 0$ e da aproximação inicial x_0 que o processo gerado pelo método do ponto fixo convergirá para um zero $\overline{x}$ de *f*. Em Ruggiero e Lopes (1996, p. 58-60), você pode encontrar a demonstração do teorema seguinte que estabelece condições suficientes para que o processo seja convergente.

Teorema 2: *Seja* $\overline{x}$ *uma raiz da equação* $f(x) = 0$*, isolada em um intervalo I centrado em* $\overline{x}$ *e seja g uma função de iteração para a equação* $f(x) = 0$*. Se*

i) g *e sua derivada,* g'*, são contínuas em I*

ii) $|g'(x)| \leq M < 1, \forall x \in I$

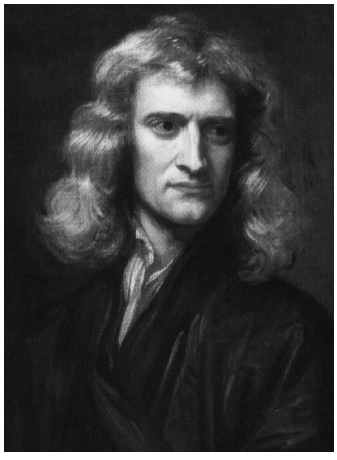
iii) $x_0 \in I$

então a sequência $(x_k)_{k \in \mathbb{N}}$ *gerada pelo processo iterativo* $x_{k+1} = g(x_k)$ *converge para* $\overline{x}$ *.seja a aproximação* $\tilde{x}$ *escolhida em* I_k*,* $|\overline{x} - \tilde{x}| < \varepsilon$.

Quanto ao critério de parada, nos métodos de ponto fixo, adotamos os critérios 2 e 3 apresentados no tópico 1, ou seja, para em um ponto x_k se $|x_k - x_{k-1}| < \varepsilon_1$ ou $|f(x_k)| < \varepsilon_2$.

Dependendo das propriedades de *g*, surgem diferentes tipos de métodos de ponto fixo. Finalizaremos esta aula, destacando um particular método de ponto fixo, o *Método de Newton-Raphson* que é bem conhecido e bastante utilizado.

O método de Newton-Raphson é um método de ponto fixo em que a escolha da função de iteração é feita visando acelerar a



Fonte: http://http://pt.wikipedia.org/

Figura 8: Isaac Newton

convergência, ou seja, tentando tornar o processo mais rápido. A condição (ii) no teorema 2 estabelece que $|g'(x)|<1$. Na verdade, é possível mostrar que a convergência será tanto mais rápida quanto menor for o fator $|g'(\overline{x})|$. Portanto, para acelerar a convergência, o método de Newton-Raphson escolhe g tal que $g'(\overline{x})=0$.

Olhando para a forma geral $g(x)=x+A(x)f(x)$, a condição $g'(\overline{x})=0$ será atingida se tomarmos $A(x)=-\dfrac{1}{f'(x)}$. Portanto, a função de iteração para o método de Newton-Raphson é

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

Verifique, como forma de exercício, que $g'(\overline{x})=0$ (evidentemente, devemos impor $f'(\overline{x})\neq 0$).

Assim, partindo de uma aproximação inicial x_0, a aproximação x_k é dada pela relação

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

Exemplo 3:

Voltemos mais uma vez à equação $x^2-2x-3=0$ do exemplo 1. Aqui, $f(x)=x^2-2x-3$, o que implica que $f'(x)=2x-2$. Portanto, a função de iteração é $g(x)=x-\dfrac{x^2-2x-3}{2x-2}$ e o processo iterativo é dado por

$$x_{k+1}=x_k-\frac{x_k^2-2x_k-3}{2x_k-2}\Rightarrow x_{k+1}=\frac{x_k^2+3}{2x_k-2}.$$

Partindo, novamente, da aproximação inicial $x_0=1,5$, obtemos

$x_0=1,5$.

$x_1=\dfrac{1,5^2+3}{2\cdot 1,5-2}=5,25$.

$x_2=\dfrac{5,25^2+3}{2\cdot 5,25-2}\cong 3,5955882353$.

$x_3=\dfrac{3,5955882353^2+3}{2\cdot 3,5955882353-2}\cong 3,0683323613$.

$x_4=\dfrac{3,0683323613^2+3}{2\cdot 3,0683323613-2}\cong 3,0011287624$.

$x_5=\dfrac{3,0011287624^2+3}{2\cdot 3,0011287624-2}\cong 3,0000003183$.

$\vdots$

Perceba que, em 5 iterações, obtivemos uma aproximação $x_5 = 3,0000003183$ para a raiz $\overline{x} = 3$ bem mais precisa que a aproximação $x_7 = 3,0010976948$ obtida em 7 iterações no exemplo 2 com a função de iteração g_3 dada por $g_3(x) = 2 + \frac{3}{x}$.

Há uma interpretação geométrica para o método de Newton-Raphson. A partir da aproximação x_k, a aproximação x_{k+1} é obtida graficamente traçando-se a reta t tangente ao gráfico de *f* pelo ponto passando pelo ponto de abscissa x_k. O valor x_{k+1} é, então, dado pela abscissa do ponto de interseção da tangente com o eixo das abscissas (eixo x). Isso justifica que o método de Newton-Raphson seja também chamado de *Método das Tangentes*.

Conforme indicado na figura 9, por um lado, a tangente do ângulo α que a reta *t* forma com o eixo *x* é igual a $f'(x_k)$ e, por outro, dá-se pela razão $\frac{f(x_k)}{x_k - x_{k+1}}$. Assim,

$$f'(x_k) = \frac{f(x_k)}{x_k - x_{k+1}} \quad \Rightarrow \quad x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

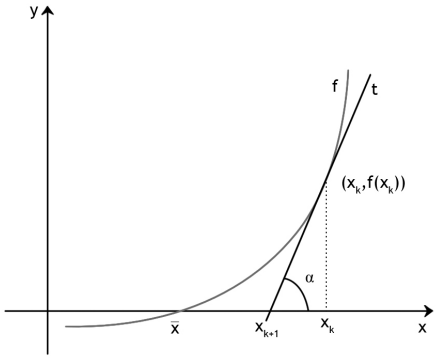


Figura 9: Interpretação geométrica do Método de Newton-Raphson

A convergência do método de Newton-Raphson é assegurada no teorema seguinte. Sua demonstração segue a demonstração do teorema 2 com a especificidade da função de iteração para o método de Newton-Raphson e também pode ser encontrada em Ruggiero e Lopes (1996, p. 69-70).

Teorema 3: *Sejam* f, f' *e* f'' *contínuas em um intervalo I que contém a raiz* $\overline{x}$ *da equação* $f(x) = 0$. *Suponha que* $f'(\overline{x}) \neq 0$. *Então, existe um intervalo* $\overline{I} \subset I$, *contendo* $\overline{x}$, *tal que, se* $x_0 \in \overline{I}$, *a sequência* $(x_k)_{k \in \mathbb{N}}$ *gerada pelo processo iterativo* $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$ *converge para* $\overline{x}$.

Os critérios de parada para o método de Newton-Raphson são os mesmos

adotados para os métodos de ponto fixo de modo geral. Para finalizar, vamos a mais um exemplo.

Exercício resolvido 2:

Determinar, usando o método de Newton-Raphson, uma aproximação para o zero de $f(x) = x \cdot \ln x - 1$, com erro inferior a 10^{-3}.

Solução:

Temos $f'(x) = 1 \cdot \ln x + x \cdot \frac{1}{x} - 0 = \ln x + 1$. Portanto, o processo iterativo é dado por

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_k \cdot \ln x_k - 1}{\ln x_k + 1} \quad \Rightarrow \quad x_{k+1} = \frac{x_k + 1}{\ln x_k + 1}.$$

Precisamos obter uma aproximação inicial x_0. Para tanto, recorremos ao método gráfico. Da equivalência

$x \cdot \ln x - 1 = 0 \quad \Leftrightarrow \quad \ln x = \frac{1}{x}$, fazemos $f_1(x) = \ln x$ e $f_2(x) = \frac{1}{x}$ e esboçamos os gráficos de f_1 e f_2 no mesmo sistema de coordenadas, observando seus pontos de intersecção (figura 10). Como você já sabe, as abscissas dos pontos de interseção das duas curvas correspondem aos zeros de *f*.

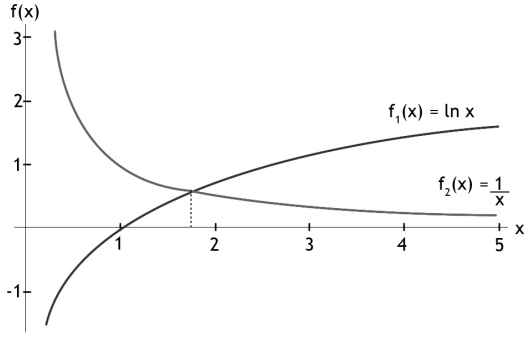


Figura 10: Gráficos de $f_1(x) = \ln x$ e $f_2(x) = \frac{1}{x}$ no intervalo $(0, 5]$.

Analisando a figura 10, vemos que há um zero de *f* no intervalo $[1, 2]$ e, portanto, tomaremos $x_0 = 1{,}5$. Trabalharemos com a representação em ponto fixo e 4 (quatro) casas decimais e usando arredondamentos, obtemos

k	x_k	$\lvert f(x_k) \rvert$	$\lvert x_{k+1} - x_k \rvert$
0	$x_0 = 1{,}5000$	0,3918	

1	$x_1 = \frac{1,5000+1}{\ln 1,5000+1} = \frac{1,5000+1}{0,4055+1} = 1,7787$	0,0244	0,3674
2	$x_2 = \frac{1,7787+1}{\ln 1,7787+1} = \frac{1,7787+1}{0,5759+1} = 1,7632$	0,0000	0,0155

Assim, em apenas duas iterações, obtemos uma aproximação $x_2 = 1,7632$ que satisfaz a precisão requerida.

Nesta aula, conhecemos os principais métodos numéricos iterativos para obter aproximações para zeros reais de funções reais e os aplicamos para a solução de alguns problemas. Vimos também condições para a garantia da convergência destes métodos e estabelecemos critérios de parada dos processos.

SAIBA MAIS!

Consulte as referências que citamos ou outras da área e acesse páginas da internet relacionadas ao tema estudado nessa aula para complementar seus conhecimentos. Abaixo, listamos algumas páginas que poderão ajudá-lo. Bons estudos!

1. http://www.profwillian.com/_diversos/download/livro_metodos.pdf
2. www.ime.usp.br/~asano/LivroNumerico/LivroNumerico.pdf
3. http://www.das.ufsc.br/~camponog/Disciplinas/DAS-5103/LN.pdf

AULA 4

Resolução de sistemas lineares: Métodos Diretos

Caro(a) aluno(a),

Olá! Nesta aula, iniciaremos nossos estudos sobre o problema de resolver sistemas lineares. Faremos uma breve introdução mostrando a importância do problema e apresentando alguns conceitos e a notação utilizada. Teremos ainda a oportunidade de conhecer e trabalhar com alguns dos chamados métodos diretos para resolver o problema, como o *método de eliminação de Gauss* e o *método da fatoração de Cholesky*.

Objetivos

- Contextualizar o problema de resolver sistemas lineares
- Caracterizar métodos numéricos diretos e iterativos para resolver o problema
- Conhecer alguns dos principais métodos diretos

TÓPICO 1

Introdução aos Sistemas Lineares

OBJETIVOS

- Conhecer o problema de resolver sistemas lineares e a sua importância
- Rever conceitos básicos
- Estabelecer a notação utilizada

Você já tem uma boa noção sobre o problema de resolver sistemas lineares. Este tema foi discutido na disciplina de *Fundamentos de Álgebra* do segundo semestre. Nela, foram apresentados, inclusive, alguns métodos diretos de resolução de sistemas lineares. Portanto, usaremos esta aula para revisitar alguns dos métodos que vocês já conhecem, dando-lhes um maior aprofundamento e para introduzir outros métodos diretos ainda não trabalhados.

O tema de sistemas lineares é um dos principais objetos de estudo da Álgebra Linear e desempenha um papel fundamental na Matemática, bem como em outras ciências, em especial nas exatas e nas engenharias. Aplicações de sistemas lineares a situações concretas ocorrem em diversas situações, como *"nas engenharias, na análise econômica, nas imagens de ressonância magnética, na análise de fluxo de tráfego, na previsão do tempo e na formulação de decisões ou de estratégias comerciais"* (ANTON E BUSBY, 2006, p.59), e podem ter milhares ou até milhões de incógnitas.

Encontraremos aplicações dos sistemas lineares em vários problemas que são tratados por métodos numéricos como na interpolação polinomial, no ajuste de curvas, na solução de sistemas de equações não lineares, na solução de equações diferenciais parciais e no cálculo de autovalores e autovetores.

Nesta aula, faremos uma breve revisão do estudo de sistemas lineares, destacando as possibilidades para as soluções de um sistema linear, apresentando a notação utilizada e descrevendo alguns dos métodos diretos para resolvê-los.

ATENÇÃO!

Nas equações lineares com poucas incógnitas (quando n é igual a 2, 3 ou 4, por exemplo), costumamos indicar as incógnitas sem índices. As incógnitas de uma equação linear costumam ser chamadas também de variáveis. Entretanto esta terminologia é mais indicada para funções.

Desde que um sistema de *equações lineares* é um conjunto de equações lineares, devemos relembrar que uma equação é linear se cada termo contém não mais do que uma incógnita e cada incógnita aparece na primeira potência.

Definição 1: *Uma equação linear nas incógnitas* $x_1, x_2, \ldots, x_n$ *é uma equação que pode ser expressa na forma padrão*

$$a_1x_1 + a_2x_2 + \ldots + a_nx_n = b, \qquad (1)$$

em que $a_1, a_2, \ldots, a_n$ *e b são constantes reais. A constante* a_i *é chamada coeficiente da incógnita* x_i *e a constante b é chamada constante ou termo independente da equação.*

São, portanto, lineares as equações $2x - 3y + 5z = 1$ e $x_1 - 3x_2 + 4x_3 = 5 - x_4 + 2x_5$. Observe que a segunda equação pode ser escrita na forma $x_1 - 3x_2 + 4x_3 + x_4 - 2x_5 = 5$. Entretanto as equações $2x - 3yz = 4$ e $x^3 + 4y - z = 7$ não são lineares, pois, na primeira equação, o segundo termo contém duas incógnitas e, na segunda equação, o primeiro termo contém uma incógnita elevada ao cubo.

A seguir, formalizamos a definição de sistema linear e apresentamos a forma comumente utilizada para descrevê-lo.

Definição 2: *Uma coleção finita de equações lineares é denominada um sistema de equações lineares ou, simplesmente, um sistema linear. Um sistema linear de m equações a n incógnitas* $x_1, x_2, \ldots, x_n$ *pode ser descrito na forma*

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots \quad\quad \vdots \quad\quad\quad\quad \vdots &\quad\quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned} \qquad (2)$$

em que a_{ij} *e* b_i *são constantes reais. A constante* a_{ij} *é chamada coeficiente da incógnita* x_j *na equação i e a constante* b_i *é chamada constante ou termo independente da equação i.*

Uma *solução* do sistema linear (2) é uma n-upla de números $(s_1, s_2, \ldots, s_n)$

tais que, sendo substituídos nos lugares de $x_1, x_2, \ldots, x_n$, respectivamente, tornam cada equação uma identidade. Ou seja, uma solução para o sistema linear (2) é um vetor $(s_1, s_2, \ldots, s_n)$, cujos componentes satisfazem simultaneamente a todas as equações do sistema.

Exemplo 1:

$$\begin{aligned} 2x_1 - x_2 + 3x_3 + x_4 &= -1 \\ x_1 + 5x_2 + 2x_3 - 7x_4 &= 8 \end{aligned}$$

Este exemplo se trata de um sistema linear de duas equações a quatro incógnitas. A quádrupla $s = (2, 3, -1, 1)$ é uma solução do sistema linear (3), porque, quando substituímos $x_1 = 2$, $x_2 = 3$, $x_3 = -1$ e $x_4 = 1$, as duas equações são satisfeitas. Verifique isso! Já o vetor $v = (1, 2, -1, 2)$ não é uma solução deste sistema linear, pois, apesar de satisfazer a primeira equação, não satisfaz a segunda, uma vez que $1 + 5 \cdot 2 + 2 \cdot (-1) - 7 \cdot 2 = 8$ ou $-5 = 8$ não é uma verdade.

O conjunto de todas as soluções de um sistema linear é denominado c*onjunto solução* ou *solução geral* do sistema linear. Referimos-nos ao processo de encontrar o conjunto solução de um sistema linear como *resolver o sistema*.

VOCÊ SABIA?

A determinação do conjunto solução dos sistemas lineares é um tema de estudo relevante dentro da Matemática Aplicada e, particularmente, em muitos tópicos de Engenharia. A complexidade de muitos sistemas, com elevado número de equações e de incógnitas, requer, muitas vezes, o auxílio de um computador para resolvê-los. Existem diversos algoritmos que permitem encontrar, caso existam, soluções de um sistema, recorrendo eventualmente a *métodos numéricos de aproximação*.

Quanto ao número de soluções, você já sabe da disciplina de Fundamentos de Álgebra que um sistema linear geral de *m* equações a *n* incógnitas pode ter *nenhuma*, *uma* ou *uma infinidade* de soluções, não havendo outras possibilidades. Um sistema linear é chamado *possível* quando tem pelo menos uma solução e *impossível* quando não tem solução. Assim, um sistema linear *possível* tem ou uma solução ou uma infinidade de soluções, não havendo outras possibilidades. Quando tem uma única solução, dizemos ainda que o sistema é *possível determinado*. Quando tem uma infinidade de soluções, dizemos também que o sistema é *possível indeterminado*. A figura 1 ilustra todas as possibilidades para o número de soluções de um sistema linear.

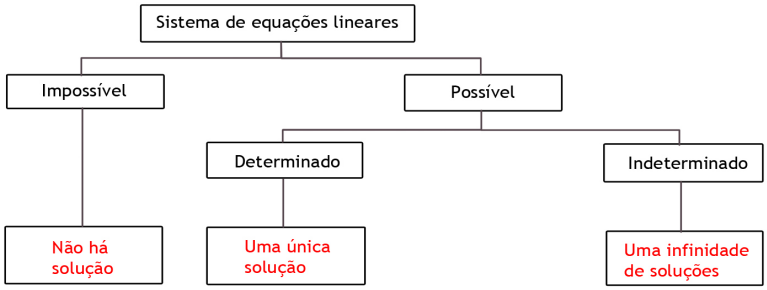


Figura 1: Classificação de um sistema linear quanto ao número de soluções

Recorrendo à notação matricial, o sistema linear (2) acima é equivalente à equação matricial

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \qquad (3)$$

ou, simplesmente, $AX = B$, em que

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \; X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \text{ e } B = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}.$$

A matriz $A = [a_{ij}]$ é a *matriz dos coeficientes das incógnitas*, também chamada *matriz do sistema*; $X = [x_j]$ é a *matriz (vetor) das incógnitas* e $B = [b_i]$ é a *matriz (vetor) das constantes* ou *matriz (vetor) dos termos independentes*.

A afirmação de equivalência significa que toda solução do sistema linear (2) é também solução da equação matricial (3) e vice-versa.

Outra matriz associada ao sistema linear é a matriz

ATENÇÃO!

Os termos *consistente* e *compatível* também são usados para nos referirmos a um sistema linear possível. Um sistema linear impossível é também chamado de *inconsistente* ou *incompatível*.

GUARDE BEM ISSO!

À medida que aumenta o número de equações e de incógnitas dos sistemas lineares, a complexidade da álgebra envolvida na obtenção de soluções também aumenta. Entretanto os cálculos necessários podem ficar mais tratáveis pela simplificação da notação e pela padronização dos procedimentos. Desse modo, ao estudar sistemas de equações lineares, é, em geral, mais simples utilizar a linguagem e a teoria das matrizes.

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{bmatrix},$$

chamada *matriz aumentada do sistema* ou *matriz completa do sistema*. Ela é a matriz *A* do sistema linear aumentada de uma coluna correspondente ao vetor *B* das constantes.

Exemplo 2:

O sistema linear de duas equações a três incógnitas

$$\begin{aligned} 2x - 3y + 4z &= -8 \\ x + 2y - 5z &= 10 \end{aligned}$$

pode ser escrito como

$$\begin{bmatrix} 2 & -3 & 4 \\ 1 & 2 & -5 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -8 \\ 10 \end{bmatrix}.$$

A matriz aumentada do sistema é

$$\begin{bmatrix} 2 & -3 & 4 & -8 \\ 1 & 2 & -5 & 10 \end{bmatrix}.$$

Consideraremos apenas os sistemas lineares em que o número de equações seja igual ao número de incógnitas, ou seja, em que $m = n$ e nos referiremos a eles como um sistema linear de ordem *n*. Tais sistemas aparecem com frequência em aplicações de diversas áreas.

Antes de descrevermos detalhadamente alguns dos métodos de solução de sistemas lineares, devemos deixar claro que eles são divididos em dois grupos (RUGGIERO e LOPES, 1996):

- *Métodos diretos:* também chamados *métodos exatos*, são aqueles que, a menos de erros de arredondamento, fornecem uma solução exata (caso uma exista) em um número finito de operações aritméticas.
- *Métodos iterativos:* são aqueles que, partindo de uma aproximação inicial, geram uma sequência de aproximações da solução exata que, sob certas condições, converge para uma solução exata (caso uma exista).

Nessa aula, abordaremos apenas métodos diretos. Estudaremos métodos iterativos na aula seguinte.

Nosso objetivo será o de estudar métodos numéricos para resolver sistemas lineares de ordem *n*, que tenham solução única. Vale destacar que para esses sistemas a matriz *A* dos coeficientes é não singular, ou seja, é tal que $\det(A) \neq 0$. Mais ainda, nesses casos, a matriz A é invertível, ou seja, existe a matriz A^{-1} tal que $AA^{-1} = A^{-1}A = I$. Portanto, temos

$$AX = B \Leftrightarrow X = A^{-1}B$$

e, então, $A^{-1}B$ é a solução do sistema linear.

Desse modo, o problema estaria resolvido por um método direto. Na prática, necessitaríamos apenas de calcular a inversa A^{-1} e, em seguida, efetuar o produto $A^{-1}B$. Entretanto, computacionalmente, a tarefa de determinar a inversa de uma matriz não é das mais fáceis.

Além da solução por inversão da matriz dos coeficientes, outro método direto é a regra de Cramer, comumente utilizada no ensino médio para a resolução de um sistema linear de ordem n. Esse método envolve o cálculo de $n+1$ determinantes de matrizes de ordem *n*, demandando também um enorme esforço computacional, especialmente para sistemas lineares de porte maior. Para se ter uma ideia da ineficiência da Regra de Cramer frente ao método do escalonamento (método que estudaremos a seguir), Lima et al. (2001, p. 289) apresenta a seguinte comparação

[...] imaginemos um computador (um tanto ultrapassado) capaz de efetuar um milhão de multiplicações ou divisões por segundo. Para resolver um sistema de 15 equações lineares com 15 incógnitas, usando a Regra de Cramer, tal computador demoraria 1 ano, 1 mês e 16 dias. O mesmo computador, usando o método de escalonamento (que é bem elementar e não requer determinantes) levaria $2\frac{1}{2}$ milésimos de segundo para resolver dito sistema. Se tivéssemos um sistema 20×20, a Regra de Cramer requereria 2 milhões, 745 mil e 140 anos para obter a solução! O método de escalonamento usaria apenas 6 milésimos de segundo para resolver o sistema.

Nos dias de hoje, a Regra de Cramer deve ser tratada como um fato teórico interessante, útil em algumas situações. Entretanto, pelas desvantagens e limitações que apontamos, não pode ser considerada uma técnica computacional eficiente para resolver sistemas lineares. Desse modo, precisamos buscar métodos mais eficientes para resolvê-los. É o que faremos no próximo tópico.

TÓPICO 2

Método de eliminação de Gauss

OBJETIVOS

- Resolver sistemas lineares triangulares
- Compreender o funcionamento do método de eliminação de Gauss
- Usar estratégias de pivoteamento

Mesmo quando se trata de sistemas lineares pequenos e, especialmente, quando o número de equações e/ou incógnitas cresce, o excesso de trabalho (cálculos) que se apresenta justifica a utilização de alguma técnica que sistematize e simplifique seu processo de resolução. Uma técnica muito utilizada e bastante eficiente e conveniente é o método de *eliminação de Gauss* ou *método de eliminação gaussiana*, também conhecido como *método do escalonamento*, que apresentaremos neste tópico. Esta técnica se baseia em combinações lineares das equações do sistema.

Fonte: http://upload.wikimedia.org/wikipedia/commons/9/9b/Carl_Friedrich_Gauss.jpg

Figura 2: Carl Friedrich Gauss

Para se ter uma ideia da importância do método de eliminação de Gauss, inclusive para a Educação Básica, destacamos o que dizem a esse respeito as orientações curriculares para o Ensino Médio:

A resolução de sistemas 2×3 ou 3×3 também deve ser feita via operações elementares (o processo de escalonamento), com discussão das diferentes situações (sistemas com uma única solução, com infinitas soluções e sem solução). Quanto à resolução de sistemas de equação 3×3, a regra de Cramer deve ser abandonada, pois é um procedimento custoso (no geral, apresentado sem demonstração, e, portanto de pouco significado para o aluno), que só permite resolver os sistemas quadrados com solução única. Dessa forma, fica também dispensado o estudo de determinantes. (BRASIL, 2006, p. 78).

De um modo simplificado, uma forma de resolver um sistema linear é substituir o sistema inicial por outro equivalente (que tenha o mesmo conjunto solução) ao primeiro, porém que seja mais fácil de resolver.

O método de eliminação de Gauss aplicado a um sistema linear de ordem n consiste em transformar o sistema original em um sistema equivalente com matriz dos coeficientes triangular superior. O método de Gauss se baseia no fato de um sistema linear de ordem n triangularizado

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \ddots \quad \vdots &\quad \vdots \\ a_{nn}x_n &= b_n \end{aligned}, \qquad (4)$$

ou seja, um sistema AX = B cuja matriz dos coeficientes é triangular superior e tal que os elementos da diagonal são não nulos ($a_{ii} \neq 0$, $i = 1, 2, ..., n$) ter solução obtida facilmente por retrossubstituição (substituição de trás para frente) dos valores das incógnitas encontrados a partir da última equação na equação anterior.

De fato, da última equação do sistema (4), temos que $x_n = \frac{b_n}{a_{nn}}$. Substituindo o valor de x_n na penúltima equação, obtemos $x_{n-1} = \frac{b_{n-1} - a_{n-1,n}x_n}{a_{n-1,n-1}}$. Prosseguindo desse modo, obtemos, sucessivamente, x_{n-2}, x_{n-3}, ..., x_2 e, finalmente, x_1 que é dado por $x_1 = \frac{b_1 - a_{12}x_2 - a_{13}x_3 - \ldots - a_{1n}x_n}{a_{11}}$. De uma forma mais resumida, x_i é dado por

$$x_i = \frac{1}{a_{ii}}(b_i - \sum_{k=i+1}^{n} a_{ik}x_k) , \ i = n, n-1, ..., 1 .$$

Exemplo 3:

O sistema linear

$$\begin{aligned} 2x + 4y - z &= 11 \\ 5y + z &= 2 \\ 3z &= -9 \end{aligned}$$

é triangular. Podemos resolvê-lo por retrossubstituição:

i. A última equação dá $z = -3$.

ii. Levando o valor de z na segunda equação, obtemos $5y + (-3) = 2$, ou $5y = 5$, ou $y = 1$.

iii. Levando os valores de z e de y na primeira equação, obtemos $2x+4\cdot(1)-(-3)=11$, ou $2x+4+3=11$, ou $2x=4$, ou $x=2$.

Portanto, o vetor $s=(2,1,-3)$ é a solução única do sistema.

Uma forma de obter um sistema equivalente a um sistema dado é aplicar sucessivamente uma série de operações (que não alterem a solução do sistema) sobre as suas equações. Desse modo, uma sucessão de sistemas cada vez mais simples pode ser obtida eliminando incógnitas de maneira sistemática usando três tipos de operações:

1. Trocar duas equações de posição.
2. Multiplicar uma equação por uma constante não-nula.
3. Somar a uma equação outra equação multiplicada por uma constante.

Tais operações são chamadas operações elementares com as equações de um sistema linear e, formalmente, temos o seguinte teorema:

Teorema 1: *Seja um sistema S' de equações lineares, obtido de outro sistema S de equações lineares por uma sequência finita de operações elementares. Então S' e S têm o mesmo conjunto solução.*

A prova deste teorema pode ser vista em Lipschutz (1994) ou nos outros livros de Álgebra Linear citados em nossas referências. As ideias centrais por trás da prova são

→ Se $\overline{x}$ é solução de um sistema linear, então $\overline{x}$ também é solução do sistema linear obtido aplicando-se uma operação elementar sobre suas equações.

→ Se o sistema S', é obtido de S aplicando-se *uma* operação elementar às suas equações, então o sistema S também pode ser obtido de S' aplicando-se uma operação elementar às suas equações, pois cada operação elementar possui uma operação elementar inversa do mesmo tipo, que desfaz o que a anterior fez.

Usaremos a seguinte notação para as três operações elementares com as equações de um sistema linear com equações $E_1, E_2, \ldots, E_m$:

1. $E_i \leftrightarrow E_j$ significa trocar as equações i e j.
2. $E_i \leftarrow kE_i$ significa multiplicar a equação i pela constante k.
3. $E_i \leftarrow E_i + kE_j$ significa somar k vezes a equação i à equação.

Já vimos como é fácil resolver um sistema linear triangular. Para completar o processo todo do método de eliminação de Gauss, resta-nos apresentar o algoritmo para "reduzir" ou "transformar" um sistema linear de ordem n para um sistema triangular equivalente. Chamaremos esse algoritmo de *algoritmo da redução.*

Algoritmo da Redução:

Passo 1: Seja $k=1$.

Passo 2: Permute a primeira equação com outra, se necessário, de modo que a incógnita x_k apareça como a primeira incógnita com coeficiente diferente de zero na primeira equação.

Passo 3: Some múltiplos convenientes da primeira equação a cada uma das equações seguintes de modo a ter todos os coeficientes da incógnita x_k abaixo da primeira equação iguais a zero.

Passo 4: Se $k=n-1$, pare. Se não, oculte a primeira equação, faça $k=k+1$ e repita todos os passos, a partir do passo 2, ao sistema linear que restou.

Na etapa j do processo, o passo 3 consiste em eliminar a incógnita x_k de todas as equações ainda envolvidas no processo, exceto da primeira. Para isso, devem-se somar múltiplos convenientes da primeira equação a cada uma das equações seguintes. Se no Passo 3 a é o coeficiente de x_k na primeira equação envolvida no processo e b é o coeficiente de x_k em uma equação l seguinte, então o múltiplo conveniente é $-\frac{b}{a}$. Nesse caso, dizemos que a é o *pivô* da etapa k e que o número $\frac{b}{a}$, denotado por m_{lk} é o multiplicador da equação l na etapa k.

VOCÊ SABIA?

Uma vez que estamos interessados apenas em sistemas lineares de ordem n que tenha solução única, é possível mostrar que o pivô em cada etapa será não-nulo.

Exemplo 4:

Vamos aplicar o algoritmo da redução ao sistema linear

$$\begin{aligned} 2x &+ y - 2z = 10 \\ -4x &+ 2y + z = -3 \\ 5x &+ \frac{11}{2}y - 3z = 25 \end{aligned}$$

Etapa 1 $(k=1)$:

Aqui, x já é a primeira incógnita com coeficiente diferente de zero da

primeira equação. O pivô da etapa 1 é $a_{11}=2$. Os multiplicadores da etapa 1 são $m_{21}=\frac{a_{21}}{a_{11}}=\frac{-4}{2}=-2$, multiplicador da equação 2, e $m_{31}=\frac{a_{31}}{a_{11}}=\frac{5}{2}$, multiplicador da equação 3. Vamos agora eliminar a incógnita x da segunda e terceira equações. Para isso, vamos somar $-m_{21}=2$ vezes a primeira equação à segunda equação e somar $-m_{31}=-\frac{5}{2}$ vezes a primeira equação à terceira equação para obter

$$\begin{aligned} 2x &+ y - 2z = 10 \\ & 4y - 3z = 17 \\ & 3y + 2z = 0 \end{aligned}$$

Uma vez que esse sistema ainda não é triangular, ocultaremos a primeira equação e repetiremos o procedimento considerando apenas as duas últimas equações.

Etapa 2 ($k=2$):

Aqui, y já é a primeira incógnita com coeficiente diferente de zero da primeira equação restante. O pivô da etapa 2 é $a_{22}=4$. A etapa 2 tem apenas um multiplicador: $m_{32}=\frac{a_{32}}{a_{22}}=\frac{3}{4}$, multiplicador da equação 3. Vamos agora eliminar a incógnita y da terceira equação. Para isso, vamos somar $-m_{32}=-\frac{3}{4}$ vezes a primeira equação à terceira equação para obter

$$\begin{aligned} 2x &+ y - 2z = 10 \\ & 4y - 3z = 17 \\ & \frac{17}{4}z = -\frac{51}{4} \end{aligned}$$

ATENÇÃO!

Alternativamente, temos ainda um método de eliminação que evita a etapa de retrossubstituição. Esse método, denominado método de *eliminação de Gauss-Jordan*, consiste em uma modificação do método de eliminação de Gauss e exige que o sistema seja transformado para um sistema linear em uma forma denominada "escalonada reduzida". No caso de o sistema original ser de ordem n e ter solução única, o sistema obtido será triangular superior com a matriz dos coeficientes tendo diagonal unitária.

Este último sistema linear é triangular. Resolvendo-o por retrossubstituição, temos $z=-3$, $y=2$ e $x=1$. Portanto, a única solução do sistema linear original é o vetor $s=(1,2,-3)$.

Conforme vimos, o método de eliminação de Gauss requer o cálculo dos multiplicadores em cada etapa, ou seja, na etapa k, dos números $m_{lk}=\frac{a_{lk}}{a_{kk}}$, multiplicador da equação l na etapa k, com a_{kk} e a_{lk} sendo os coeficientes de x_k nas equações k e l. Já sabemos que o pivô em cada etapa será não-nulo. Mas, o que

ocorrerá se tivermos um pivô próximo de zero? De acordo com Ruggiero e Lopes (1996, p. 127),

> ... trabalhar com um pivô próximo de zero pode conduzir a resultados totalmente imprecisos. Isto porque em qualquer calculadora ou computador os cálculos são efetuados com aritmética de precisão finita, e pivôs próximos de zero dão origem a multiplicadores bem maiores que a unidade que, por sua vez, origina uma ampliação dos erros de arredondamento.

O uso de estratégias de pivoteamento, ou seja, de processos para a escolha da linha e/ou coluna do pivô, é indicado para evitar (ou pelo menos minimizar) este tipo de problema. As estratégias de pivoteamento podem ser de

→ *Pivoteamento parcial:* o pivô para a etapa *k* é escolhido como o elemento de maior módulo entre os coeficientes a_{lk}, $l = k, k+1, ..., n$ (coeficientes da incógnita x_k nas equações ainda restantes no processo), ou seja, o pivô será o elemento a_{rk} tal que

$$|a_{rk}| = \max\{|a_{lk}| : l = k, k+1, ..., n\}.$$

Se $r \neq k$, trocam-se as linhas k e r.

→ *Pivoteamento total:* o pivô para a etapa k é escolhido como o elemento de maior módulo entre os coeficientes a_{ij}, tais que $i = k, k+1, ..., n$ e $j = k, k+1, ..., n$ (coeficientes ainda restantes no processo), ou seja, o pivô será o elemento a_{rs} tal que

$$|a_{rs}| = \max\{|a_{ij}| : i = k, k+1, ..., n \text{ e } j = k, k+1, ..., n\}.$$

Se necessário, são feitas trocas de linhas e/ou colunas de modo que o pivô passe a ser o elemento a_{kk}.

O exemplo seguinte, adaptado de Ruggiero e Lopes (1996, p. 129-131), mostra a importância do uso de estratégias de pivoteamento. Ele servirá também para ilustrar possíveis erros de arredondamento causados pelo número limitado de algarismos significativos. Lembramos que os arredondamentos devem ser feitos após cada operação.

Exercício resolvido 1:

Resolver pelo método de eliminação de Gauss e pelo método de eliminação de Gauss com estratégia de pivoteamento parcial o sistema linear abaixo. Usar representação em ponto flutuante com 4 algarismos significativos

$$\begin{aligned} 0,0002x_1 &+ 2x_2 = 5 \\ 2x_1 &+ 2x_2 = 6 \end{aligned}$$

Solução:

Vamos resolver inicialmente pelo método de eliminação de Gauss sem adotar qualquer estratégia de pivoteamento.

Etapa 1 ($k=1$):

Pivô: $a_{11} = 0,2000\times10^{-3}$.

Multiplicadores: $m_{21} = \dfrac{a_{21}}{a_{11}} = \dfrac{0,2000\times10^{1}}{0,2000\times10^{-3}} = 1,000\times10^{4} = 0,1000\times10^{5}$.

Vamos agora eliminar a incógnita x da segunda. Para isso, vamos somar $-m_{21} = -0,1000\times10^{5}$ vezes a primeira equação à segunda. Temos

$$\begin{aligned} a_{22} &= a_{22} - m_{21}\times a_{12} = 0,2000\times10^{1} - (0,1000\times10^{5})\times(0,2000\times10^{1}) \\ &= 0,2000\times10^{1} - 0,2000\times10^{5} = -0,2000\times10^{5} \end{aligned}$$

$$\begin{aligned} b_2 &= b_2 - m_{21}\times b_1 = 0,6000\times10^{1} - (0,1000\times10^{5})\times(0,5000\times10^{1}) \\ &= 0,6000\times10^{1} - 0,5000\times10^{5} = -0,5000\times10^{5} \end{aligned}$$

O sistema obtido é então,

$$\begin{aligned} 0,2000\times10^{-3}x_1 &+ 0,2000\times10^{1}x_2 = 0,5000\times10^{1}, \\ &- 0,2000\times10^{5}x_2 = -0,5000\times10^{5} \end{aligned}$$

que é triangular. Resolvendo-o por retrossubstituição, obtemos

$$x_2 = \frac{-0,5000\times10^{5}}{-0,2000\times10^{5}} = 2,500\times10^{0} = 0,2500\times10^{1}$$

e

$$0,2000\times10^{-3}x_1 + 0,2000\times10^{1}\times0,2500\times10^{1} = 0,5000\times10^{1} \Rightarrow$$

$$0,2000\times10^{-3}x_1 + 0,0500\times10^{2} = 0,5000\times10^{1} \Rightarrow$$

$$x_1 = \frac{0,5000\times10^{1} - 0,5000\times10^{1}}{0,2000\times10^{-3}} = \frac{0,0000\times10^{1}}{0,2000\times10^{-3}} = 0,0000\times10^{4}.$$

Portanto, $\overline{x} = (0,0000\times10^{4}; 0,2500\times10^{1}) = (0; 2,5)$. Entretanto é fácil verificar que $\overline{x}$ não satisfaz a segunda equação, pois

$2\times0 + 2\times2,5 = 5 \neq 6$.

Agora vamos resolver novamente pelo método de eliminação de Gauss, mas, desta vez, adotaremos a estratégia de pivoteamento parcial.

Etapa 1 ($k=1$):

$\max\{|a_{l1}|: l=1,2\}=|0,2000\times10^{1}|=|a_{21}|\Rightarrow$ Pivô: $a_{21}=0,2000\times10^{1}$.

Logo, devemos trocar as equações 1 e 2. Obtemos assim o sistema

$$\begin{aligned} 0,2000\times10^{1}x_1 &+ 0,2000\times10^{1}x_2 = 0,6000\times10^{1}, \\ 0,2000\times10^{-3}x_1 &+ 0,2000\times10^{1}x_2 = 0,5000\times10^{1} \end{aligned}$$

para o qual temos

Pivô: $a_{11}=0,2000\times10^{1}$.

Multiplicadores: $m_{21}=\dfrac{a_{21}}{a_{11}}=\dfrac{0,2000\times10^{-3}}{0,2000\times10^{1}}=1,000\times10^{-4}=0,1000\times10^{-3}$.

Vamos agora eliminar a incógnita x da segunda. Para isso, vamos somar $-m_{21}=-0,1000\times10^{-3}$ vezes a primeira equação à segunda. Encontramos

$$\begin{aligned} a_{22}&=a_{22}-m_{21}\times a_{12}=0,2000\times10^{1}-(0,1000\times10^{-3})\times(0,2000\times10^{1}) \\ &=0,2000\times10^{1}-0,2000\times10^{-3}=0,2000\times10^{1} \\ b_2&=b_2-m_{21}\times b_1=0,5000\times10^{1}-(0,1000\times10^{-3})\times(0,6000\times10^{1}) \\ &=0,5000\times10^{1}-0,6000\times10^{-3}=0,5000\times10^{1} \end{aligned}$$

O sistema obtido é então

$$\begin{aligned} 0,2000\times10^{1}x_1 &+ 0,2000\times10^{1}x_2 = 0,6000\times10^{1} \\ &\quad 0,2000\times10^{1}x_2 = 0,5000\times10^{1} \end{aligned}$$

que é triangular. Resolvendo-o por retrossubstituição, temos

$$x_2=\frac{0,5000\times10^{1}}{0,2000\times10^{1}}=2,500\times10^{0}=0,2500\times10^{1}$$

e

$$0,2000\times10^{1}x_1+0,2000\times10^{1}\times0,2500\times10^{1}=0,6000\times10^{1}\Rightarrow$$

$$0,2000\times10^{1}x_1+0,0500\times10^{2}=0,6000\times10^{1}\Rightarrow$$

$$x_1=\frac{0,6000\times10^{1}-0,5000\times10^{1}}{0,2000\times10^{1}}=\frac{0,1000\times10^{1}}{0,2000\times10^{1}}=0,5000\times10^{0}.$$

Assim, $\overline{x}=(0,5000\times10^{0};0,2500\times10^{1})=(0,5;2,5)$. Podemos verificar que $\overline{x}$ satisfaz cada uma das equações do sistema. De fato,

$$\begin{aligned} &(0,2000\times10^{-3})\times(0,5000\times10^{0})+(0,2000\times10^{1})\times(0,2500\times10^{1})= \\ &\quad 0,1000\times10^{-3}+0,5000\times10^{1}=0,5000\times10^{1}=5 \end{aligned}$$

e

$$(0,2000\times10^{1})\times(0,5000\times10^{0})+(0,2000\times10^{1})\times(0,2500\times10^{1})=$$
$$0,1000\times10^{1}+0,5000\times10^{1}=0,6000\times10^{1}=6$$

Neste tópico, revimos o método de eliminação de Gauss para resolver sistemas lineares. Vimos também que o uso de estratégias de pivoteamento é importante para a redução dos possíveis erros de arredondamentos. No próximo tópico, apresentaremos mais um método que pertence à categoria dos métodos diretos.

TÓPICO 3

Método de fatoração de Cholesky

OBJETIVOS

- Compreender o funcionamento dos métodos de fatoração
- Conceituar matrizes definidas positivas
- Conhecer o método de fatoração de Cholesky

SAIBA MAIS!

A fatoração LU ou decomposição LU é das técnicas mais usadas para resolver sistemas de equações lineares. Ela consiste em decompor a matriz A dos coeficientes do sistema em um produto de duas matrizes L e U, em que L é uma matriz triangular inferior (*lower*) com diagonal unitária e U é uma matriz triangular superior (*upper*).

Em certas situações, necessitamos resolver vários sistemas lineares que têm a mesma matriz dos coeficientes. Nesses casos, as chamadas *técnicas de fatoração* ou de *decomposição* da matriz dos coeficientes se tornam bastante adequadas e eficientes. Dentre essas técnicas, merece destaque a da *fatoração LU*, bastante utilizada. Dela deriva o método de fatoração de Cholesky que abordaremos neste tópico.

Conforme visto em Ruggiero e Lopes (1996, p. 132), a técnica de fatoração para resolver um sistema linear *"consiste em decompor a matriz A dos coeficientes em um produto de dois ou mais fatores e, em seguida, resolver uma sequência de sistemas lineares que nos conduzirá à solução do sistema linear original"*.

Desse modo, se a matriz A de um sistema linear $Ax = b$ puder ser fatorada como $A = MN$, teremos que o sistema poderá ser escrito como

$$(MN)x = b\,.$$

Fazendo $y = Nx$, o problema de resolver $Ax = b$ torna-se equivalente a resolver o sistema linear $My = b$ e, em seguida, o sistema linear $Nx = y$.

Evidentemente, é desejável que, feita a fatoração da matriz A, os sistemas

lineares a serem resolvidos sejam de fácil resolução. Ademais, como deixamos transparecer acima, a vantagem dos métodos de fatoração é a de que, uma vez fatorada a matriz A, fica fácil resolver qualquer sistema linear que tenha A como matriz dos coeficientes, ou seja, se o vetor b for alterado, a resolução do novo sistema linear torna-se bem simples.

O método de fatoração de Cholesky é um método direto que se aplica a certos sistemas lineares particulares, aqueles cuja matriz dos coeficientes é *simétrica e definida positiva*. Boa parte dos problemas que envolvem sistemas de equações lineares nas ciências e engenharias têm a matriz de coeficientes simétrica e definida positiva.

Você já conhece o conceito de matriz simétrica visto na disciplina de Fundamentos de Álgebra. Vamos relembrá-lo com a definição 3 seguinte. Na definição 4, daremos o significado de matriz definida positiva.

Definição 3: *Chama-se matriz simétrica toda matriz quadrada A tal que* $A^T = A$, *ou seja, que é igual à sua transposta. Simbolicamente, uma matriz quadrada de ordem n,* $A = [a_{ij}]$, *é simétrica se, e somente se,*

$a_{ij} = a_{ji}$, $\forall i \in \{1, 2, \ldots, n\}$ e $\forall j \in \{1, 2, \ldots, n\}$.

Definição 4: *Uma matriz quadrada A de ordem n é definida positiva se, e somente se,*

$$x^T Ax > 0, \ \forall x \in \mathbb{R}^n, \ x \neq 0 .$$

Um *sistema linear* $Ax = b$ em que a matriz dos coeficientes é simétrica e definida positiva pode ter a matriz A decomposta como

$$A = MM^T ,$$

na qual M é uma matriz triangular inferior de ordem n com elementos da diagonal estritamente positivos. Tal fatoração é conhecida como *fatoração de Cholesky* e a matriz M é chamada *fator de Cholesky* da matriz A. A existência e unicidade do fator de Cholesky é garantida no teorema seguinte.

GUARDE BEM ISSO!

Uma vez que estamos interessados apenas em sistemas lineares de ordem n que tenha solução única, é possível mostrar que o pivô em cada etapa será não-nulo.

Teorema 2: *Se A for uma matriz quadrada de ordem n definida positiva, então existe uma única matriz triangular inferior M de ordem n com elementos da diagonal positivos tal que* $A = MM^T$.

A obtenção do fator M pode ser feita construtivamente a partir da equação matricial $A = MM^T$. Uma vez que $A = [a_{ij}]$ é $M = [m_{ij}]$ é triangular inferior, essa equação pode ser escrita como

$$\begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{21} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} m_{11} & 0 & \cdots & 0 \\ m_{21} & m_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ m_{n1} & m_{n2} & \cdots & m_{nn} \end{pmatrix} \begin{pmatrix} m_{11} & m_{21} & \cdots & m_{n1} \\ 0 & m_{22} & \cdots & m_{n2} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & m_{nn} \end{pmatrix}.$$

Comparando os elementos, temos

$$\begin{aligned} &a_{11} = m_{11}m_{11}, \\ &a_{21} = m_{21}m_{11}, \quad a_{22} = m_{21}m_{21} + m_{22}m_{22} \\ &\qquad\vdots \qquad\qquad\qquad\quad \vdots \\ &a_{n1} = m_{n1}m_{11}, \quad a_{n2} = m_{n1}m_{21} + m_{n2}m_{22} \quad \cdots \quad a_{nn} = m_{n1}m_{n1} + m_{n2}m_{n2} + \cdots + m_{nn}m_{nn} \end{aligned}.$$

Rearranjado as equações acima, obtemos

$$m_{jj} = a_{jj} - \sum_{k=1}^{j-1} m_{jk}^2$$

$$m_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} m_{ik}m_{jk}}{m_{jj}}, \text{ para } i > j.$$

GUARDE BEM ISSO!

Quando decompostas, as matrizes definidas positivas apresentam uma grande estabilidade numérica. O método de Cholesky aplicado a uma matriz simétrica e definida positiva não necessita de estratégias de pivoteamento (troca de linhas e/ou colunas) para manter a estabilidade numérica, o que não acontece com matrizes indefinidas.

Obtido o fator M, a solução do sistema linear original $Ax = b$ vem da resolução de dois sistemas lineares triangulares. De fato, desde que $A = MM^T$, temos

$$Ax = b \Leftrightarrow (MM^T)x = b \Leftrightarrow \begin{cases} My = b \\ M^T x = y \end{cases},$$

ou seja, devemos resolver dois sistemas lineares:

$My = b$: triangular inferior

$M^T x = y$: triangular superior

Você pode estar achando complexo trabalhar com todos esses símbolos e índices. Então, vamos a um exemplo.

Exercício resolvido 2:

Resolva pelo método de fatoração de Cholesky o sistema linear abaixo.

$$\begin{aligned} 4x_1 &+ 2x_2 &+ 14x_3 &= -6 \\ 2x_1 &+ 17x_2 &- 5x_3 &= 9 \\ 14x_1 &- 5x_2 &+ 83x_3 &= -55 \end{aligned}$$

Solução:

Devemos encontrar os coeficientes m_{ij} tais que

$$\underbrace{\begin{pmatrix} 4 & 2 & 14 \\ 2 & 17 & -5 \\ 14 & -5 & 83 \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} m_{11} & 0 & 0 \\ m_{21} & m_{22} & 0 \\ m_{31} & m_{32} & m_{33} \end{pmatrix}}_{M} \underbrace{\begin{pmatrix} m_{11} & m_{21} & m_{31} \\ 0 & m_{22} & m_{32} \\ 0 & 0 & m_{33} \end{pmatrix}}_{M^T}.$$

Dessa equação matricial, igualando coluna a coluna, obtemos

Da coluna 1:

$$4 = m_{11}^2 \Rightarrow m_{11} = \sqrt{4} = 2$$

$$2 = m_{21}m_{11} \Rightarrow m_{21} = \frac{2}{m_{11}} = \frac{2}{2} = 1$$

$$14 = m_{31}m_{11} \Rightarrow m_{31} = \frac{14}{m_{11}} = \frac{14}{2} = 7\,.$$

Da coluna 2:

$$17 = m_{21}^2 + m_{22}^2 \Rightarrow m_{22} = \sqrt{17 - m_{21}^2} = \sqrt{17 - 1^2} = \sqrt{16} = 4$$

$$-5 = m_{31}m_{21} + m_{32}m_{22} \Rightarrow m_{32} = \frac{-5 - m_{31}m_{21}}{m_{22}} = \frac{-5 - 7 \times 1}{4} = \frac{-12}{4} = -3\,.$$

Da coluna 3:

$$83 = m_{31}^2 + m_{32}^2 + m_{33}^2 \Rightarrow m_{33} = \sqrt{83 - m_{31}^2 - m_{32}^2} = \sqrt{83 - 7^2 - (-3)^2} = \sqrt{25} = 5$$

Logo,

$$M = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 4 & 0 \\ 7 & -3 & 5 \end{pmatrix} \text{ e } M^T = \begin{pmatrix} 2 & 1 & 7 \\ 0 & 4 & -3 \\ 0 & 0 & 5 \end{pmatrix}.$$

Vamos agora resolver os sistemas lineares $My = b$ e $M^T x = y$. O sistema $My = b$ é

$$\begin{aligned} 2y_1 &&&= -6 \\ 1y_1 &+ 4y_2 &&= 9\,, \\ 7y_1 &- 3y_2 &+ 5y_3 &= -55 \end{aligned}$$

cuja solução é o vetor $\overline{y} = (-3, 3, -5)$. Assim, o sistema $M^T x = y$ é

$$\begin{aligned} 2x_1 &+ 1x_2 &+ 7x_3 &= -3 \\ &4x_2 &- 3x_3 &= 3\,, \\ &&5x_3 &= -5 \end{aligned}$$

cuja solução é o vetor $\overline{x} = (2, 0, -1)$.

Nesta aula, revimos o método de eliminação de Gauss, aplicando-o para a resolução de sistemas lineares de ordem n e, visando minimizar os possíveis erros de arredondamentos, utilizamos técnicas de pivoteamento. Conhecemos ainda o método de fatoração de Cholesky que se aplica para o caso de o sistema ter matriz dos coeficientes simétrica e definida positiva. Na próxima aula, estudaremos alguns dos métodos para o problema de resolver sistemas lineares.

SAIBA MAIS!

Você pode complementar seus estudos examinando outros métodos diretos para resolver sistemas lineares, como o método de fatoração LU. Para isso, consulte as referências que citamos ou outras da área e acesse páginas da internet relacionadas ao tema. Abaixo, listamos algumas páginas que poderão ajudá-lo. Bons estudos!

http://www.profwillian.com/_diversos/download/livro_metodos.pdf
http://www.das.ufsc.br/~camponog/Disciplinas/DAS-5103/LN.pdf
http://www-di.inf.puc-rio.br/~tcosta/cap2.htm

AULA 5

Resolução de sistemas lineares: Métodos Iterativos

Olá, nesta aula, daremos continuidade aos nossos estudos sobre o problema de resolver sistemas lineares. Desta vez, abordaremos métodos iterativos para resolver o problema e enfocaremos o *método de Gauss-Jacobi* e o *método de Gauss-Seidel.*

Objetivos

- Entender o funcionamento de métodos numéricos iterativos para o problema
- Calcular aproximações para a solução de sistemas lineares
- Estudar a convergência dos métodos apresentados
- Conhecer critérios de parada dos algoritmos

TÓPICO 1

Métodos iterativos para resolução de sistemas lineares: Funcionamento e critérios de parada

OBJETIVOS

- Conhecer a ideia geral dos métodos iterativos para resolução de sistemas lineares
- Apresentar fluxograma de funcionamento dos métodos iterativos
- Estabelecer critérios de parada

Neste tópico, conheceremos, em linhas gerais, o funcionamento dos métodos iterativos para resolver sistemas de equações lineares. Compreenderemos que a ideia central por trás dos métodos que abordaremos é generalizar os métodos de ponto fixo para o cálculo de zeros de funções estudados na aula 3. Apresentaremos ainda os principais critérios de parada para estes processos.

Na aula anterior, apresentamos o problema de resolver sistemas lineares e vimos sua importância para a Matemática e para outras áreas, especialmente para as Ciências Exatas e Engenharias. Nela, você conheceu alguns dos principais métodos diretos para resolver o problema, merecendo destaque o método de eliminação de Gauss.

Além dos métodos exatos para resolver sistemas lineares, existem *os métodos iterativos* e, em certos casos, tais métodos são melhores do que os exatos. É o caso, por exemplo, quando o sistema linear é *de grande porte* e/ou quando a matriz dos coeficientes do sistema é uma *matriz esparsa*.

Relembre que um método numérico é *iterativo* quando fornece uma sequência de aproximações x_k para a solução $\overline{x}$, utilizando aproximações anteriores para calcular as novas aproximações. Em geral, o processo para obter cada nova aproximação é sempre o mesmo e, por esse motivo, dizemos que o método numérico iterativo é *estacionário*. É sempre desejável que, sob certas condições, a sequência construída convirja para a solução exata. Nesse caso, em um número finito de

repetições do procedimento, é possível obter uma aproximação que satisfaça uma precisão prefixada.

Como no caso dos métodos diretos, vamos considerar sistemas lineares de ordem n que tenham solução única, ou seja, sistemas lineares do tipo $Ax = b$, em que A é uma matriz quadrada de ordem n, x e b são vetores do $\mathbb{R}^n$ e tal que $\det(A) \neq 0$.

Seguindo a ideia dos métodos de ponto fixo para determinar aproximações para os zeros de funções, a fim de determinar uma aproximação para a solução de um sistema linear por métodos iterativos, transformamos o sistema linear original em outro sistema linear. Nesse novo sistema linear, definimos um processo iterativo. Será necessário que a solução obtida para o sistema transformado seja também a solução do sistema original, ou seja, que os sistemas lineares sejam equivalentes.

VOCÊ SABIA?

Um sistema de equações lineares é de **grande porte** se é constituído de um grande número de equações e/ou incógnitas, ou seja, tem ordem elevada. Uma matriz é dita **esparsa** quando tem a maioria de seus elementos iguais a zero, ou seja, quando possui relativamente poucos elementos não nulos. Muitos sistemas lineares que surgem de problemas reais são de ordem elevada e possuem matrizes esparsas.

Como vantagens dos métodos iterativos em relação aos métodos diretos, podemos dizer que eles

- → São mais eficientes para sistemas lineares de grande porte e/ou quando a matriz dos coeficientes do sistema é uma matriz esparsa.
- → Ocupam menos memória.
- → São mais simples de serem implementados no computador.
- → Estão menos sujeitos ao acúmulo de erros de arredondamento.
- → Podem se autocorrigir, caso um erro seja cometido.
- → Podem, sob certas condições, ser aplicados para resolver sistemas não lineares.

As restritivas condições de convergência aparecem como uma das principais desvantagens dos métodos iterativos. Eles não podem ser aplicados para a resolução de todo sistema linear.

Portanto, o sistema $Ax = b$ é transformado em um sistema equivalente do tipo

$$x = Cx + d,$$

GUARDE BEM ISSO!

Dois sistemas lineares são equivalentes se têm as mesmas soluções.

em que C é uma matriz quadrada de ordem n, x e d são vetores do $\mathbb{R}^n$. Um exemplo de sistema transformado seria aquele do tipo $x = Cx + d$, tal que $C = I - A$ e $d = b$. Verifique!

Podemos definir a função $\varphi : \mathbb{R}^n \to \mathbb{R}^n$, dada por $\varphi(x) = Cx + d$ que funciona como função de iteração na forma matricial. Desse modo, o problema de resolver o sistema linear $Ax = b$ é transformado no problema de encontrar um ponto fixo para φ.

Partindo de uma aproximação inicial x^0 para a solução $\overline{x}$ do sistema linear, podemos construir uma sequência de aproximações de $x^0, x^1, x^2, \ldots$, na qual a aproximação x^{k+1} depende da aproximação x^k pela relação

$x^{k+1} = \varphi(x^k)$, $k = 0, 1, 2, \ldots$,

ou seja, definimos uma sequência de aproximações para a solução da seguinte maneira:

$$x^{k+1} = Cx^k + d, \quad k = 0, 1, 2, \ldots,$$

em que x^0 é uma aproximação inicial dada.

Verifica-se que se a sequência $\{x^k\}$ converge para $\overline{x}$, isto é,

$$\lim_{k \to \infty} x^k = \overline{x},$$

então $\overline{x}$ é a solução do sistema $Ax = b$. De fato, passando-se ao limite (quando $k \to \infty$) ambos os membros da igualdade $x^{k+1} = Cx^k + d$, obtemos

$$\overline{x} = C\overline{x} + d.$$

Pela equivalência dos sistemas lineares, segue que $\overline{x}$ é também solução do sistema $Ax = b$.

ATENÇÃO!

No caso de métodos iterativos, é fundamental identificar se a sequência de aproximações que estamos obtendo está convergindo ou não para a solução desejada. Para tanto, é necessário ter em mente o significado de convergência de uma sequência de vetores (as aproximações são vetores). Veja este importante conceito abaixo. Você pode encontrá-lo também em livros de cálculo.

Definição 1: *Seja V um espaço vetorial. Dada uma sequência de vetores* $\{x^k\}$ *pertencentes a V e uma norma* $||.||$ *sobre V, dizemos que a sequência* $\{x^k\}$ *converge para* $\overline{x} \in V$ *se* $\lim_{k \to \infty} ||x^k - \overline{x}|| = 0$.

Talvez você ainda não conheça alguns termos nessa definição, como *espaço vetorial* e *norma*. Eles serão apresentados formalmente na disciplina de *Álgebra Linear* do próximo semestre. Uma vez que avaliaremos se uma dada aproximação é "boa" (ou seja, satisfaz uma

precisão prefixada) através da chamada norma do máximo, faremos uma breve introdução apresentando as normas mais usuais sobre o espaço vetorial $\mathbb{R}^n$.

É possível que você já tenha trabalhado com a chamada *norma euclideana padrão* sobre $\mathbb{R}^n$, que, a cada vetor $v=(v_1,v_2,\ldots,v_n)\in\mathbb{R}^n$, associa o número real

$$||v||_E=\sqrt{v_1^2+v_2^2+\ldots v_n^2}=\sqrt{\sum_{i=1}^{n}v_i^2}\ .$$

Além da norma euclideana padrão, outras normas sobre $\mathbb{R}^n$ bem conhecidas são a *norma da soma*, dada por

$$||v||_S=|v_1|+|v_2|+\ldots+|v_n|=\sum_{i=1}^{n}|v_i|,$$

e a norma do máximo, dada por

$$||v||_M=\max\{|v_1|,|v_2|,\ldots,|v_n|\}=\max\{|v_i|:i=1,2,\ldots,n\}.$$

Para fixar melhor, vejamos o exemplo a seguir.

ATENÇÃO!

Uma norma sobre o espaço vetorial $\mathbb{R}^n$ é uma função $||\ ||:\mathbb{R}^n\to\mathbb{R}$ que satisfaz as propriedades:

i. $||x||\geq 0$, $\forall x\in\mathbb{R}^n$ e $||x||=0\Leftrightarrow x=0$.

ii. $||x+y||\leq||x||+||y||$, $\forall x,y\in\mathbb{R}^n$.

iii. $||\alpha x||=|\alpha|\cdot||x||$, $\forall x\in\mathbb{R}^n$ e $\forall\alpha\in\mathbb{R}$.

Exemplo 1

Considerando o vetor $v=(2,-1,0,-5,3)\in\mathbb{R}^5$, teremos

$||v||_E=\sqrt{2^2+(-1)^2+0^2+(-5)^2+3^2}=\sqrt{39}$.

$||v||_S=|2|+|-1|+|0|+|-5|+|3|=11$.

$||v||_M=\max\{|2|,|-1|,|0|,|-5|,|3|\}=5$.

Um fato interessante é que toda norma $||.||$ sobre $\mathbb{R}^n$ induz uma distância d em $\mathbb{R}^n$ dada por

$$d(x,y)=||x-y||,\ \forall x,y\in\mathbb{R}^n.$$

Antes de passarmos aos métodos iterativos específicos que veremos, devemos deixar claro o critério de parada que adotaremos.

Supondo que $\overline{x}$ seja solução do sistema linear $Ax=b$ e que a sequência $\{x^k\}$ converge

ATENÇÃO!

Uma distância no espaço vetorial $\mathbb{R}^n$ é uma função $d:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}$ que satisfaz as propriedades:

i. $d(x,y)\geq 0$, $\forall x,y\in\mathbb{R}^n$ e $d(x,y)=0\Leftrightarrow x=y$.

ii. $d(x,y)=d(y,x)$, $\forall x,y\in\mathbb{R}^n$.

iii. $d(x,y)\leq d(x,z)+d(z,y)$, $\forall x,y,z\in\mathbb{R}^n$.

para $\overline{x}$ ($\lim_{k\to\infty} ||x^k - \overline{x}|| = 0$), é possível mostrar que

$$\lim_{k\to\infty} ||x^k - x^{k-1}|| = 0 ,$$

ou seja, a sequência dos termos consecutivos converge para 0.

Baseado nesse fato, dada uma precisão (tolerância) prefixada ε, paramos um processo iterativo para determinar uma aproximação para a solução $\overline{x}$ de um sistema linear determinado $Ax = b$ de ordem n se a aproximação x^k calculada na k-ésima iteração satisfaz

$$||x^k - x^{k-1}||_M < \varepsilon .$$

Isso corresponde à distância entre dois iterados (aproximação calculada em uma iteração) consecutivos ser menor que ε.

Portanto, interrompemos o processo iterativo quando o vetor x^k estiver suficientemente próximo do vetor x^{k-1} ou, mais precisamente, quando a distância entre os vetores x^k e x^{k-1}, dada por

$$d^k = d(x^k, x^{k-1}) = ||x^k - x^{k-1}||_M = \max\{|x_i^k - x_i^{k-1}| : i = 1, 2, ..., n\} ,$$

satisfaz $d^k < \varepsilon$.

Do mesmo modo que para os métodos iterativos para obter aproximações para zeros de funções, podemos efetuar o teste do erro relativo, em que fazemos

$$d_r^k = \frac{d^k}{\max\{|x_i^k| : i = 1, 2, ..., n\}} .$$

É interessante também exigir que o número de iterações não ultrapasse um limite máximo *N* de iterações preestabelecido, ou seja, paramos também se $k = N$.

Estamos agora em condições de conhecer alguns métodos numéricos iterativos específicos para o cálculo de uma aproximação para a solução de um sistema linear determinado de ordem *n*. Então, vamos ao próximo tópico, no qual veremos primeiro método.

TÓPICO 2

Método de Gauss-Jacobi

OBJETIVOS

- Compreender o funcionamento do método de Gauss-Jacobi
- Calcular aproximações para soluções de sistemas lineares
- Estabelecer o critério das linhas para convergência do método

O que caracteriza cada método iterativo para resolver sistemas lineares é a forma como o sistema $Ax = b$ é transformado no sistema equivalente $x = Cx + d$, ou seja, a forma como é definida a função de iteração matricial $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ dada por $\varphi(x) = Cx + d$. Neste tópico, analisaremos o modo particular que o método de Gauss-Jacobi faz tal transformação, ou seja, veremos como é feito o isolamento de x no método de Gauss-Jacobi.

Vamos considerar um sistema linear de ordem n nas incógnitas $x_1, x_2, \ldots, x_n$,

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \cdots & + & a_{nn}x_n & = & b_n \end{array},$$

que pode ser escrito na forma matricial $Ax = b$, em que a matriz A dos coeficientes do sistema é quadrada de ordem n e b é um vetor do $\mathbb{R}^n$. Suponhamos que $a_{ii} \neq 0$, $i = 1, 2, \ldots, n$ (todos os elementos da diagonal da matriz A são não nulos).

O método de Gauss-Jacobi faz o isolamento do vetor x pela diagonal do seguinte modo:

ATENÇÃO!

Muitas vezes, a condição $a_{ii} \neq 0$, $i = 1, 2, \ldots, n$ pode não ser cumprida pelo sistema original. Em alguns desses casos, uma reordenação das equações e/ou incógnitas pode tornar a condição satisfeita.

$$
\begin{aligned}
x_1 &= \frac{1}{a_{11}}(b_1 - a_{12}x_2 - a_{13}x_3 - \cdots - a_{1n}x_n)\\
x_2 &= \frac{1}{a_{22}}(b_2 - a_{21}x_1 - a_{23}x_3 - \cdots - a_{2n}x_n)\\
&\vdots \qquad\qquad \vdots \qquad\qquad ,\\
x_n &= \frac{1}{a_{nn}}(b_n - a_{n1}x_1 - a_{n2}x_2 - \cdots - a_{nn}x_n)
\end{aligned}
$$

ou seja, isolamos a incógnita x_1 pela primeira equação, a incógnita x_2 pela segunda equação e, sucessivamente, isolamos a incógnita x_n pela n-ésima equação. Note que isto só é possível porque estamos supondo $a_{ii} \neq 0$, $i = 1, 2, \ldots, n$.

Na forma matricial, temos $x = Cx + d$, com

$$
C = \begin{pmatrix}
0 & -\frac{a_{12}}{a_{11}} & -\frac{a_{13}}{a_{11}} & \cdots & -\frac{a_{1n}}{a_{11}} \\
-\frac{a_{21}}{a_{22}} & 0 & -\frac{a_{23}}{a_{22}} & \cdots & -\frac{a_{2n}}{a_{22}} \\
-\frac{a_{31}}{a_{33}} & -\frac{a_{32}}{a_{33}} & 0 & \cdots & -\frac{a_{3n}}{a_{33}} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
-\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & -\frac{a_{n3}}{a_{nn}} & \cdots & 0
\end{pmatrix} \text{ e } d = \begin{pmatrix}
\frac{b_1}{a_{11}} \\
\frac{b_2}{a_{22}} \\
\frac{b_3}{a_{33}} \\
\vdots \\
\frac{b_n}{a_{nn}}
\end{pmatrix}.
$$

GUARDE BEM ISSO!

Substituindo o vetor aproximação x^k (seus componentes) no lado direito das equações acima, obteremos uma nova aproximação x^{k+1}, sendo que, para o cálculo do i-ésima componente do vetor x^{k+1}, dado por

$$
x_i^{k+1} = \frac{1}{a_{ii}}(b_i - a_{i1}x_1^k - a_{i2}x_2^k - \cdots - a_{i,i-1}x_{i-1}^k - a_{i,i+1}x_{i+1}^k - \cdots - a_{in}x_n^k)
$$

$i = 1, 2, \ldots, n$,

utilizamos todos os componentes do vetor x^k, exceto o componente x_i^k.

Desse modo, fornecida uma aproximação inicial $x^0 = (x_1^0, x_2^0, \ldots, x_n^0)$, o método de Gauss-Jacobi consiste em construir uma sequência de aproximações $x^0, x^1, x^2, \ldots$, dada pela relação recursiva

$$x^{k+1} = Cx^k + d,$$

ou seja, por

$$
\begin{aligned}
x_1^{k+1} &= \frac{1}{a_{11}}(b_1 - a_{12}x_2^k - a_{13}x_3^k - \cdots - a_{1n}x_n^k)\\
x_2^{k+1} &= \frac{1}{a_{22}}(b_2 - a_{21}x_1^k - a_{23}x_3^k - \cdots - a_{2n}x_n^k)\\
&\vdots \qquad\qquad \vdots \qquad\qquad .\\
x_n^{k+1} &= \frac{1}{a_{nn}}(b_n - a_{n1}x_1^k - a_{n2}x_2^k - \cdots - a_{nn}x_n^k)
\end{aligned}
$$

Para uma melhor apropriação do processo iterativo de Gauss-Jacobi, vejamos um exemplo.

Exemplo 1

Considere o sistema linear

$$\begin{aligned} 2x_1 &+ x_2 = 1 \\ -x_1 &+ 4x_2 = -5 \end{aligned}$$

O processo iterativo de Gauss-Jacobi é dado por

$$\begin{aligned} x_1^{k+1} &= \frac{1}{2}(1 - x_2^k) \\ x_2^{k+1} &= \frac{1}{4}(-5 + x_1^k) \end{aligned}$$

Trabalhando com representação em ponto fixo com 5 casas decimais e fazendo arredondamentos, partindo da aproximação inicial $x^0 = (0,0)$, obteremos os seguintes resultados para as iterações:

k	x_1^k	x_2^k
0	0,00000	0,00000
1	0,50000	− 1,25000
2	1,25000	− 1,25000
3	1,06250	− 0,96875
4	0,98438	− 0,98438
5	0,99219	− 1,00391
6	1,00195	− 1,00195
7	1,00098	− 0,99951
8	0,99976	− 0,99976
9	0,99988	− 1,00006
10	1,00003	− 1,00003
11	1,00001	− 0,99999
12	1,00000	− 1,00000

Tabela 1: Iterações do exemplo 1

O sistema linear desse exemplo é bem simples e sua solução exata $\overline{x} = (1,-1)$ pode ser obtida por um método direto qualquer.

Nesse exemplo 1, não adotamos qualquer critério de parada. Entretanto, no caso geral, quando não se conhece a solução exata do sistema, precisaremos estipular quando o processo iterativo será interrompido, ou seja, precisamos de uma precisão prefixada e considerar o critério de parada apresentado no tópico 1 ou algum outro.

Observe que as iterações no exemplo 1 estão se aproximando da solução exata do sistema linear. Entretanto, não podemos esperar isso sempre. Para motivar a necessidade de estabelecer condições que garantam a convergência da sequência de aproximações gerada pelo método de Gauss-Jacobi, vejamos mais um exemplo.

Exemplo 2

Considere o sistema linear

$$\begin{aligned} x_1 + 3x_2 - x_3 &= 3 \\ 5x_1 - 2x_2 + 2x_3 &= 8 \\ 3x_2 + 4x_3 &= -4 \end{aligned}.$$

O processo iterativo de Gauss-Jacobi é dado por

$$\begin{aligned} x_1^{k+1} &= 3 - 3x_2^k + x_3^k \\ x_2^{k+1} &= -\frac{1}{2}(8 - 5x_1^k - 2x_3^k) \\ x_3^{k+1} &= \frac{1}{4}(-4 - 3x_2^k) \end{aligned}.$$

Usando novamente representação em ponto fixo com 5 casas decimais e fazendo arredondamentos, partindo da aproximação inicial $x^0 = (1,1,1)$, obteremos os seguintes resultados para as iterações:

k	x_1^k	x_2^k	x_3^k	d^k
0	1,00000	− 0,50000	− 1,75000	2,75000
1	2,75000	− 3,25000	− 0,62500	2,75000
2	12,12500	2,25000	1,43750	9,37500
3	− 2,31250	27,75000	− 2,68750	25,50000
4	− 82,93750	− 12,46880	− 21,81250	80,62500
5	18,59380	− 233,15600	8,35156	220,68800
6	710,82000	50,83590	173,86700	692,22700

Tabela 2: Iterações do exemplo 2

A solução exata deste sistema linear é $\overline{x} = (2,0,-1)$ e as iterações parecem estar divergindo de $\overline{x}$. Observe que a distância d^k ente os dois iterados consecutivos x^k e x^{k-1} está aumentando.

Portanto, será fundamental estabelecer critérios que assegurem a convergência da sequência de aproximações gerada pelo método de Gauss-Jacobi. *O critério das linhas*, apresentado no teorema seguinte, estabelece uma condição suficiente para tal garantia conhecida (RUGGIERO E LOPES, 1996).

Teorema 1: *Seja o sistema linear* $Ax = b$ *de ordem n e seja*

$$\alpha_k = \frac{1}{|a_{kk}|}\sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}|.$$

Se $\alpha = \max\{\alpha_k : k = 1, 2, ..., n\} < 1$, *então o método de Gauss-Jacobi gera uma sequência* $\{x_k\}$ *convergente para a solução do sistema dado, independente da escolha da aproximação inicial* x_0.

GUARDE BEM ISSO!

O número α_k associado à linha k é o quociente entre a soma dos valores absolutos (módulos) de todos os coeficientes da linha k da matriz A, exceto o coeficiente a_{kk} pelo valor absoluto do coeficiente a_{kk}.

Como exemplo de aplicação do critério das linhas, verifique que ele é satisfeito para o sistema linear do exemplo 1. Faremos a seguir a verificação para o sistema do exemplo 2.

Exemplo 3

Vamos verificar o critério das linhas para o sistema linear do exemplo 2. Temos

$$\alpha_1 = \frac{|a_{12}|+|a_{13}|}{|a_{11}|} = \frac{|3|+|-1|}{|1|} = 4,$$

$$\alpha_2 = \frac{|a_{21}|+|a_{23}|}{|a_{22}|} = \frac{|5|+|2|}{|-2|} = \frac{7}{2} \text{ e}$$

$$\alpha_3 = \frac{|a_{31}|+|a_{32}|}{|a_{33}|} = \frac{|0|+|3|}{|4|} = \frac{3}{4}.$$

$$\alpha = \max\{\alpha_k : k = 1, 2, 3\} = \max\{4, \frac{7}{2}, \frac{3}{4}\} = 4 > 1.$$

Portanto, o *critério das linhas* não é satisfeito e não podemos garantir (por este critério) que a sequência gerada pelo método de Gauss-Jacobi irá convergir. De fato, pelo que observamos da construção da sequência, ela parece divergir da solução exata.

GUARDE BEM ISSO!

O critério das linhas dá uma condição suficiente para garantir a convergência da sequência. Entretanto, ela pode não ser necessária, ou seja, a sequência pode convergir sem que o critério das linhas seja satisfeito.

Voltando ao exemplo 2, se reordenarmos o sistema permutando a primeira com a segunda equação, obtemos o sistema linear

$$\begin{aligned} 5x_1 - 2x_2 + 2x_3 &= 8 \\ x_1 + 3x_2 - x_3 &= 3 \\ 3x_2 + 4x_3 &= -4 \end{aligned}$$

Esse novo sistema linear é equivalente ao sistema original e satisfaz o critério das linhas. Verifique, como forma de exercício, este fato.

Desse modo, é mais adequado aplicarmos o método de Gauss-Jacobi a esta nova disposição, pois há garantia de que a sequência gerada irá convergir para a solução do novo sistema que, por sua vez, é a solução do sistema original. Isso motiva uma ideia interessante, exposta em Ruggiero e Lopes (1996, p. 161): *"... sempre que o critério das linhas não for satisfeito, devemos tentar uma permutação de linhas e/ou colunas de forma a obtermos uma disposição para a qual a matriz dos coeficientes satisfaça o critério das linhas"*. Mas atenção: nem sempre é possível obter tal disposição!

Neste tópico, vimos o método de Gauss-Jacobi, estabelecendo uma condição para garantia de sua convergência. A seguir, apresentaremos o método de Gauss-Seidel.

TÓPICO 3

Método de Gauss-Seidel

OBJETIVOS

- Compreender o funcionamento do método de Gauss-Seidel
- Calcular aproximações para soluções de sistemas lineares
- Estabelecer o critério de Sassenfeld para convergência do método

Neste tópico, apresentaremos o método iterativo de Gauss-Seidel para resolver sistemas lineares. Ele pode ser visto como uma variação do método de Gauss-Jacobi em que, para o cálculo de um componente da nova aproximação, são usados, além dos componentes da aproximação anterior, os já calculados da nova aproximação.

Essa é uma ideia bem interessante, uma vez que podemos esperar que, no caso de haver convergência para a solução exata do sistema, os componentes da nova aproximação sejam "melhores" que os componentes da aproximação anterior.

Mais precisamente, supondo que $a_{ii} \neq 0$, $i = 1, 2, ..., n$, o processo iterativo para o método de Gauss-Seidel consiste em, partindo de uma aproximação inicial $x^0 = (x_1^0, x_2^0, \ldots, x_n^0)$, construir uma sequência de aproximações $x^0, x^1, x^2, \ldots$, dada pelas relações recursivas.

$$
\begin{aligned}
x_1^{k+1} &= \frac{1}{a_{11}}(b_1 - a_{12}x_2^k - a_{13}x_3^k - a_{14}x_4^k - \cdots - a_{1n}x_n^k) \\
x_2^{k+1} &= \frac{1}{a_{22}}(b_2 - a_{21}x_1^{k+1} - a_{23}x_3^k - a_{34}x_4^k - \cdots - a_{2n}x_n^k) \\
x_3^{k+1} &= \frac{1}{a_{33}}(b_3 - a_{31}x_1^{k+1} - a_{32}x_2^{k+1} - a_{34}x_4^k - \cdots - a_{3n}x_n^k) \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
x_n^{k+1} &= \frac{1}{a_{nn}}(b_n - a_{n1}x_1^{k+1} - a_{n2}x_2^{k+1} - a_{34}x_4^{k+1} - \cdots - a_{nn}x_n^{k+1})
\end{aligned}
$$

Portanto, o i-ésima componente do vetor x^{k+1}, dado por

$$x_i^{k+1} = \frac{1}{a_{ii}}(b_i - a_{i1}x_1^{k+1} - a_{i2}x_2^{k+1} - \cdots - a_{i,i-1}x_{i-1}^{k+1} - a_{i,i+1}x_{i+1}^{k} - \cdots - a_{in}x_n^{k}),$$

$i = 1, 2, \ldots, n$,

é calculado utilizando todos os componentes do vetor x^{k+1} já calculados (componentes do vetor x^{k+1} com índices menores que i) e os componentes do vetor x^k com índices maiores que i, ou seja, usando os componentes $x_1^{k+1}, x_2^{k+1}, \cdots, x_{i-1}^{k+1}$ do vetor x^{k+1} e os componentes $x_{i+1}^{k}, x_{i+2}^{k}, \cdots, x_n^{k}$ do vetor x^k.

Como vantagens do *método de Gauss-Seidel* em relação ao método de Gauss-Jacobi, podemos esperar que

→ a convergência seja acelerada

→ os critérios de convergência sejam menos restritivos.

SAIBA MAIS!

O método da Gauss-Seidel é conhecido também por *Método dos Deslocamentos Sucessivos*, uma vez que, para o cálculo de uma componente de x^{k+1}, utilizam-se os valores mais recente das demais componentes.

Para exemplificar, vamos repetir o que foi feito no exemplo 1, desta vez usando o processo iterativo de Gauss-Seidel.

Exemplo 4

Considere o sistema linear

$$\begin{aligned} 2x_1 &+ x_2 &= 1 \\ -x_1 &+ 4x_2 &= -5 \end{aligned}.$$

O processo iterativo de Gauss-Jacobi é dado por

$$\begin{aligned} x_1^{k+1} &= \frac{1}{2}(1 - x_2^k) \\ x_2^{k+1} &= \frac{1}{4}(-5 + x_1^{k+1}) \end{aligned}.$$

Trabalhando com representação em ponto fixo com 5 casas decimais e fazendo arredondamentos, partindo da aproximação inicial $x^0 = (0,0)$, obtemos os seguintes resultados para as iterações:

k	x_1^k	x_2^k
0	0,00000	0,00000
1	0,50000	− 1,12500

2	1,06250	− 0,98438
3	0,99219	− 1,00195
4	1,00098	− 0,99976
5	0,99988	− 1,00003
6	1,00001	− 1,00000
7	1,00000	− 1,00000

Tabela 3: Iterações do exemplo 1

Observe que pelo método de Gauss-Seidel, com o sistema de numeração escolhido, foram necessárias apenas 7 iterações para obter a solução $\overline{x} = (1,00000; -1,00000)$, enquanto que pelo método de Gauss-Jacobi precisamos de 12 iterações.

Do mesmo modo que no método de Gauss-Jacobi, o método de Gauss-Seidel transforma o sistema original $Ax = b$ de ordem n em um sistema equivalente do tipo $x = Cx + d$, ou seja, a função de iteração matricial é dada por $\varphi(x) = Cx + d$.

Assim, apesar de utilizarmos componentes do vetor x^{k+1}, nas relações recursivas para o processo de Gauss-Seidel apresentadas acima, o processo iterativo para o método pode ser escrito como

$$x^{k+1} = Cx^k + d,$$

ou seja, com os componentes da nova aproximação sendo dados em termos apenas dos componentes da aproximação anterior. Para isso, devemos fazer

$$C = -(I + L_1)^{-1} R_1^{-1} \text{ e } d = (I + L_1)^{-1} D^{-1} b.$$

em que

$$L_1 = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ \frac{a_{21}}{a_{22}} & 0 & 0 & \cdots & 0 \\ \frac{a_{31}}{a_{33}} & \frac{a_{32}}{a_{33}} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{a_{n1}}{a_{nn}} & \frac{a_{n2}}{a_{nn}} & \frac{a_{n3}}{a_{nn}} & \cdots & 0 \end{pmatrix}, \quad R_1 = \begin{pmatrix} 0 & \frac{a_{12}}{a_{11}} & -\frac{a_{13}}{a_{11}} & \cdots & \frac{a_{1n}}{a_{11}} \\ 0 & 0 & \frac{a_{23}}{a_{22}} & \cdots & \frac{a_{2n}}{a_{22}} \\ 0 & 0 & 0 & \cdots & \frac{a_{3n}}{a_{33}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix},$$

$$D = \begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ 0 & 0 & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{pmatrix} \text{ e } I = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Portanto, o processo iterativo do método de Gauss-Seidel é dado pela relação recursiva

$$x^{k+1} = -(I + L_1)^{-1} R_1^{-1} x^k + (I + L_1)^{-1} D^{-1} b .$$

Você pode encontrar uma demonstração desse fato em Ruggiero e Lopes (1996) ou em outras referências da área.

Passaremos agora a estabelecer critérios que garantam a convergência da sequência de aproximações gerada pelo método de Gauss-Seidel.

O critério das linhas, usado para avaliar a convergência do método de Gauss-Jacobi, pode ser aplicado também para estabelecer uma condição suficiente para a convergência do método de Gauss-Seidel (RUGGIERO E LOPES, 1996). Então, temos o teorema seguinte.

Teorema 2: *Seja o sistema linear* $Ax = b$ *de ordem n e seja*

$$\alpha_k = \frac{1}{|a_{kk}|} \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{kj}| .$$

Se $\alpha = \max\{\alpha_k : k = 1, 2, ..., n\} < 1$*, então o método de Gauss-Seidel gera uma sequência* $\{x_k\}$ *convergente para a solução do sistema dado, independente da escolha da aproximação inicial* x_0*.*

Outro critério que estabele uma condição suficiente para garantir a convergência da sequência de aproximações gerada pelo método de Gauss-Seidel é o critério de Sassenfeld, apresentado no teorema abaixo. Você pode encontrar este resultado demonstrado em Ruggiero e Lopes (1996) ou em outras referências da área.

Teorema 3: *Seja o sistema linear* $Ax = b$ *de ordem n e seja*

$$\beta_k = \frac{1}{|a_{kk}|} \left(\sum_{j=1}^{k-1} |a_{kj}| \beta_j + \sum_{j=k+1}^{n} |a_{kj}| \right).$$

Se $\beta = \max\{\beta_k : k = 1, 2, ..., n\} < 1$*, então o método de Gauss-Seidel gera uma sequência* $\{x_k\}$ *convergente para a solução do sistema dado, independente da escolha da aproximação inicial* x_0*.*

O número β_k associado à linha *k* é o quociente entre a soma dos valores absolutos (módulos) de todos os coeficientes da linha *k* da matriz *A*, exceto o coeficiente a_{kk} pelo valor absoluto do coeficiente a_{kk}, sendo que os valores

absolutos dos coeficientes com índice j menor que k são multiplicados por β_k, ou seja, os números β_k são dados por

$$\beta_1 = \frac{|a_{12}|+|a_{13}|+\cdots+|a_{1n}|}{|a_{11}|} \text{ e}$$

$$\beta_k = \frac{|a_{k1}|\beta_1+|a_{k2}|\beta_2+\cdots+|a_{k,k-1}|\beta_{k-1}+|a_{k,k+1}|+\cdots+|a_{kn}|}{|a_{kk}|}.$$

Note que o número β_1 é igual ao número α_1 do *critério das linhas*.

O número β está associado à ordem de convergência da sequência gerada pelo método, entretanto a convergência será tanto mais rápida quanto menor for o valor de β.

O critério de Sassenfeld apresenta uma condição menos restritiva que o critério das linhas. É possível mostrar que o critério de Sassenfeld é satisfeito sempre que o critério das linhas for satisfeito. Entretanto, a recíproca desse resultado não é verdadeira, ou seja, é possível que o critério de Sassenfeld seja satisfeito sem que o critério das linhas seja satisfeito. O exemplo seguinte é uma ilustração desse fato.

ATENÇÃO!

Os critérios das linhas ou de Sassenfeld não dependem das constantes (dos termos independentes) do sistema. Assim, se um sistema linear $Ax = b$ cumpre a condição de um desses critérios, dizemos também que a matriz A dos coeficientes do sistema satisfaz essa condição.

Exemplo 5

Considere o sistema linear

$$\begin{aligned} 5x_1 &- x_2 + x_3 = 3 \\ 3x_1 &+ 4x_2 + 2x_3 = 5. \\ -3x_1 &+ 3x_2 + 6x_3 = -6 \end{aligned}$$

Vamos verificar o critério das linhas para o sistema. Temos

$$\alpha_1 = \frac{|a_{12}|+|a_{13}|}{|a_{11}|} = \frac{|-1|+|1|}{|5|} = \frac{2}{5},$$

$$\alpha_2 = \frac{|a_{21}|+|a_{23}|}{|a_{22}|} = \frac{|3|+|2|}{|4|} = \frac{5}{4} \text{ e}$$

$$\alpha_3 = \frac{|a_{31}|+|a_{32}|}{|a_{33}|} = \frac{|-3|+|3|}{|6|} = 1.$$

$$\alpha = \max\{\alpha_k : k = 1, 2, 3\} = \max\{\frac{2}{5}, \frac{5}{4}, 1\} = \frac{5}{4} > 1.$$

Logo, o critério das linhas não é satisfeito e não podemos garantir (por este critério) que a sequência gerada pelo método de Gauss-Seidel irá convergir. Note que não precisaríamos sequer calcular α_3, pois do fato que $\alpha_2 = \frac{5}{4} > 1$ já poderíamos afirmar que $\alpha = \max\{\alpha_k : k = 1, 2, 3\} > 1$.

GUARDE BEM ISSO!

Como o critério das linhas, o critério de Sassenfeld dá uma condição suficiente para garantir a convergência da sequência. Entretanto, ela pode não ser necessária, ou seja, a sequência pode convergir sem que o critério de Sassenfeld seja satisfeito.

Vamos agora verificar o *critério de Sassenfeld* para o sistema. Temos

$$\beta_1 = \frac{|a_{12}| + |a_{13}|}{|a_{11}|} = \frac{|-1| + |1|}{|5|} = \frac{2}{5},$$

$$\beta_2 = \frac{|a_{21}|\beta_1 + |a_{23}|}{|a_{22}|} = \frac{|3| \times \frac{2}{5} + |2|}{|4|} = \frac{4}{5} \text{ e}$$

$$\beta_3 = \frac{|a_{21}|\beta_1 + |a_{23}|\beta_2}{|a_{22}|} = \frac{|-3| \times \frac{2}{5} + |3| \times \frac{4}{5}}{|4|} = \frac{9}{10}.$$

$$\beta = \max\{\beta_k : k = 1, 2, 3\} = \max\{\frac{2}{5}, \frac{4}{5}, \frac{9}{10}\} = \frac{9}{10} < 1.$$

Portanto, o critério de Sassenfeld é satisfeito e podemos garantir que a sequência gerada pelo método de Gauss-Seidel irá convergir. Que tal determinar uma aproximação para a solução desse sistema linear pelo método de Gauss-Seidel com erro inferior a $\varepsilon = 10^{-2}$? Faça isso como exercício!

Do mesmo modo que observamos para aplicação do critério das linhas no método de Gauss-Jacobi, caso o critério de Sassenfeld não seja satisfeito para um sistema dado, você pode tentar uma nova disposição (um sistema equivalente), permutando linhas e/ou colunas para examinar o critério. Obviamente, caso haja tal disposição para a qual o critério seja satisfeito, devemos aplicar o método de Gauss-Seidel a ela por termos a garantia de convergência. Mas lembre: nem sempre é possível obter tal disposição!

Neste tópico, vimos o método de Gauss-Seidel, estabelecendo condições para garantia de sua convergência. Com isso, completamos nossos estudos sobre técnicas numéricas para resolver sistemas lineares. Agora, você já tem bastantes ferramentas para tratar esta importante classe de problemas: os métodos diretos, discutidos na aula 4; e os métodos iterativos, vistos nesta aula. Nas próximas aulas, você conhecerá outros tipos de problemas que podem ser tratados por métodos numéricos e terá a oportunidade de aplicar os conhecimentos adquiridos até aqui.

SAIBA MAIS!

Aprofunde seus conhecimentos consultando as referências que citamos ou outras da área e/ou acessando páginas da internet relacionadas ao tema. Abaixo, listamos algumas páginas que poderão ajudá-lo. Bons estudos!

http://www.profwillian.com/_diversos/download/livro_metodos.pdf

http://www.das.ufsc.br/~camponog/Disciplinas/DAS-5103/LN.pdf

www.ime.usp.br/~asano/LivroNumerico/LivroNumerico.pdf

http://www.dma.uem.br/kit/arquivos/arquivos_pdf/sassenfeld.pdf

AULA 6

Interpolação Polinomial

Olá a todos! Vamos continuar nosso estudo de Cálculo Numérico e das ferramentas de aproximação de resultados. Em muitas situações, obtemos dados pontuais para o estudo de determinado fenômeno. Se tivermos condições de, a partir dos dados obtidos, conseguir uma função que represente (ou aproxime) o processo, poderemos fazer simulações para resultados intermediários ou próximos, diminuindo a necessidade de repetição para os experimentos ou obtendo valores em intervalos fora da precisão da máquina.

Por exemplo, um responsável por um laboratório pode fazer medições regulares da pressão de um determinado gás e obter como dados $\{(t_1, P_1), (t_2, P_2), (t_3, P_3), (t_4, P_4)\}$. Uma função $f(t)$ tal que, para cada um dos tempos dados, satisfaça $f(t_i) = P_i$ (ou sejam bem próximos) permitirá uma boa avaliação da pressão no gás em outros tempos, sem que seja necessária a medição.

Nesta aula, estudaremos especificamente a aproximação por polinômios dos dados apresentados.

Objetivos

- Analisar aproximações de dados por funções
- Apresentar métodos de obtenção dos polinômios interpoladores

TÓPICO 1

Definições Iniciais

Objetivos

- Formular o problema de interpolação polinomial
- Resolver problemas de interpolação pelo método direto

Imaginemos, inicialmente, a seguinte situação da Física: um móvel se desloca em uma trajetória orientada passando sucessivamente pelos pontos s =20m, s =30m e s =50m para tempos iguais a 3s, 5s e 7s, respectivamente. Colocando esses dados em uma tabela, obtemos

t (em segundos)	s (em metros)
3	20
5	30
7	50

Tabela 1: Representação dos dados do problema

A partir desses dados, podemos nos perguntar qual a posição do móvel para t=4s. Como responder satisfatoriamente a essa pergunta se não foi feita a observação do espaço do móvel no tempo dado? Se a velocidade dele fosse constante, poderíamos simplesmente fazer a média aritmética entre os valores para t=3s e para t=7s. Entretanto, pelos dados do problema, verifica-se imediatamente que o movimento não é uniforme (pois, de 3 a 5 segundos, ele percorreu $10\ m$, enquanto nos dois segundos seguintes foram percorrido 20m). Outra informação da qual não dispomos é se a aceleração é constante ou não.

Se tivéssemos uma função $s(t)$ que descrevesse esse movimento, bastaria substituir t=4s para encontrar o espaço desejado. Com apenas os pontos dados, algo que podemos fazer para ter uma boa noção da posição do móvel para t=4s, de modo a não perder as informações, seria admitir um comportamento para $s(t)$, que poderia ser o de uma função exponencial, trigonométrica ou polinomial, sendo

essa última alternativa mais simples para fins de cálculo. Então, supondo que *s(t)* é uma função polinomial de *t*, tal que $s(3) = 20,\ s(5) = 30$ e $s(7) = 50$, podemos ter uma boa aproximação para o valor de $s(4)$.

Exemplo 1

Encontre um polinômio $s(t)$, de segundo grau, tal que $s(3) = 20,\ s(5) = 30$ e $s(7) = 50$.

Solução:

Um polinômio do segundo grau é da forma $s(t) = at^2 + bt + c$. Devemos encontrar, então, números reais a, b e c para que $s(3) = 20,\ s(5) = 30$ e $s(7) = 50$, ou seja, $a.3^2 + b.3 + c = 20;\ a.5^2 + b.5 + c = 30$ e $a.7^2 + b.7 + c = 50$ que equivale a

$$\begin{aligned} 9a + 3b + c &= 30 \\ 25a + 5b + c &= 30 \\ 49a + 7b + c &= 50 \end{aligned}.$$

Usando algum dos métodos que conhecemos para resolver sistemas lineares, encontraremos a solução (exata) $a = 1{,}25,\ b = -5$ e $c = 23{,}75$. Assim, o polinômio desejado será $s(t) = 1{,}25t^2 - 5t + 23{,}75$.

Empregando a solução encontrada no exemplo 1, podemos obter uma aproximação para $s(4) = 1{,}25.4^2 - 5.4 + 23{,}75 = 23{,}75$. Com isso, conseguimos, sem desprezar os dados apresentados, aproximar a posição do móvel para $t = 4\ s$ por $s = 23{,}75\ \text{m}$.

Vista essa situação inicial, podemos formular o problema da interpolação polinomial.

> **Problema 1:** Para o conjunto de dados $\{(x_0, y_0), (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, encontre um polinômio $p(x)$, de grau menor ou igual a *n*, para o qual $p(x_i) = y_i$, para $i = 0, 1, 2, \ldots, n$.

Em outras palavras, interpolar polinomialmente alguns dados consiste em encontrar uma função polinomial cujo gráfico passe pelos pontos dados.

Aqui surgem dois questionamentos:

→ o problema tem solução?

→ a solução é única?

Para responder às duas perguntas ao mesmo tempo, deve-se observar que todo polinômio de grau menor ou igual a *n* pode ser escrito da forma

$p(x) = a_n x^n + ... + a_1 x + a_0$ .Substituindo os pontos dados, devemos ter, necessariamente: $p(x_i) = y_i$, para todo i = 0, 1, ..., n. Ou seja:

$$p(x_0) = a_n x_0^n + ... + a_1 x_0 + a_0 = y_0$$
$$p(x_1) = a_n x_1^n + ... + a_1 x_1 + a_0 = y_1$$

...

$p(x_n) = a_n x_n^n + ... + a_1 x_n + a_0 = y_n$, que gera um sistema nas incógnitas a_n, ..., a_1, a_0 dado por

$$\begin{cases} a_n x_0^n + ... + a_1 x_0 + a_0 = y_0 \\ a_n x_1^n + ... + a_1 x_1 + a_0 = y_1 \\ \qquad\qquad ... \\ a_n x_n^n + ... + a_1 x_n + a_0 = y_n \end{cases}$$, matricialmente equivalente a

$$\begin{bmatrix} x_0^n & ... & x_0 & 1 \\ x_1^n & ... & x_1 & 1 \\ ... & \ddots & ... & ... \\ x_n^n & ... & x_n & 1 \end{bmatrix} . \begin{bmatrix} a_n \\ a_{n-1} \\ ... \\ a_0 \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ ... \\ y_n \end{bmatrix} .$$

Uma vez que a matriz dos coeficientes é de Vandermonde (ou de potências), seu determinante será diferente de zero sempre que os valores de x_i forem todos distintos. Desse modo, teremos um sistema possível e determinado, de onde podemos concluir que a solução do problema existe e é única.

A partir de agora, sabendo que o problema de interpolação polinomial sempre terá solução (o que nos tranquiliza um bocado), nossa preocupação será em COMO resolvê-lo de forma eficiente.

Observação 1: Ao polinômio solução para o problema 1, damos o nome de *polinômio interpolador*.

Observação 2: Para um conjunto de $n + 1$ dados, devemos encontrar um polinômio de grau menor ou igual a n, ou seja, o grau máximo do polinômio interpolador será um a menos que a quantidade de pontos.

Exemplo 2

Encontre o polinômio interpolador para o conjunto de dados $\{(-1,0),\ (1,2),\ (2,7),\ (3,26)\}$.

Solução:

Uma vez que o conjunto de dados possui pontos com abscissas todas

distintas, o problema terá solução. Assim, buscaremos um polinômio de grau menor ou igual a 3 (pois há quatro pontos). Um polinômio de grau menor ou igual a 3 é da forma $p(x) = ax^3 + bx^2 + cx + d$. Com as condições do problema, devemos ter $p(-1) = 0$, $p(1) = 2$, $p(2) = 7$ e $p(3) = 26$. Por isso, devemos resolver o sistema

$$\begin{cases} -a + b - c + d = 0 \\ a + b + c + d = 2 \\ 8a + 4b + 2d + c = 7 \\ 27a + 9b + 3c + d = 26 \end{cases}$$

. Para tanto, devemos fazer uso de algum método para resolução de sistemas lineares, como visto nas últimas aulas ou pelos conhecimentos adquiridos em outras disciplinas. A solução para o sistema é $a = 1$, $b = c = 0$ e $d = -1$. Assim, o polinômio procurado é $p(x) = x^3 - 1$.

Exemplo 3

Em um laboratório, um físico fez medições regulares na pressão de um gás e organizou os resultados na seguinte tabela:

tempo(s)	pressão(atm)
5	2,5
8	6,8
13	11,9

Usando interpolação polinomial, estime a pressão do gás para t = 10 s.

Solução:

Temos o conjunto de dados $\{(5;2,5),\ (8;6,8),\ (13;11,9)\}$. Aqui usamos ponto e vírgula para separar as coordenadas de modo a evitar confusão com a vírgula que separa a parte decimal. O polinômio procurado será de grau menor ou igual a 2, sendo, portanto, da forma $p(t) = at^2 + bt + c$. De maneira análoga ao exemplo anterior, devemos resolver o sistema $\begin{cases} 25a + 5b + c = 2,5 \\ 64a + 8b + c = 6,8 \\ 169a + 13b + c = 11,9 \end{cases}$. Obviamente, aqui temos um trabalho maior que no exemplo anterior por causa dos dados "quebrados". Realizando um processo qualquer da aula passada, podemos encontrar aproximações até a segunda casa decimal para $a = -0,05$; $b = 2,1$ e $c = -6,73$. Assim, um polinômio que aproxima a pressão a qualquer instante é $p(t) = -0,05t^2 + 2,1t - 6,73$. Desse modo, uma estimativa para a pressão do gás em t=10s pode ser obtida por $p(10) = -0,05.10^2 + 2,1.10 - 6,73 = 9,27$ atm.

Pelo que vimos neste tópico, podemos sempre aproximar um conjunto de dados por um polinômio. Entretanto, dependendo da quantidade de dados, esse processo pode ser muito trabalhoso de ser realizado diretamente pela solução de um sistema linear. Nos próximos tópicos, veremos métodos para encontrar o polinômio interpolador.

TÓPICO 2

O método de Lagrange

Objetivos

- Apresentar o método de Lagrange para obtenção do polinômio interpolador
- Comparar o método de Lagrange com o método direto

Como vimos no tópico anterior, $p(x)$ é o polinômio interpolador para um conjunto de dados $\{(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)\}$ se $p(x_i) = y_i$, para $i = 0, 1, 2, \ldots, n$. Tal polinômio sempre existe e, de modo a torná-lo único, pedimos que o seu grau fosse menor ou igual a n.

Neste tópico, descreveremos um método atribuído ao matemático nascido da Itália e naturalizado francês Joseph Louis Lagrange (1736 - 1812), a quem são devidos muitos importantes teoremas, como o Teorema do Valor Médio, do Cálculo Diferencial.

A ideia consiste basicamente em escrever o polinômio como soma de polinômios, ditos elementares, que se anulem em todos os valores do conjunto de dados, menos em um.

Exemplo 1

Encontre um polinômio $p(x)$, tal que $p(3) = 1$ e que tenha 2, 4 e 6 como raízes.

Solução:

Do estudo de polinômios, sabemos que, se ξ é uma raiz do polinômio $p(x)$, então $p(x)$ é divisível por $x - \xi$ (ver Teorema de D´Alembert). Assim, para que 2, 4 e 6 sejam raízes de um polinômio, ele deve ser divisível por $(x-2)(x-4)(x-6)$. Por simplicidade, poderíamos colocar $p(x) = (x-2)(x-4)(x-6)$. Entretanto, dessa forma, $p(3) = (3-2)(3-4)(3-6) = 3$. Para atingir o nosso objetivo, basta,

então, que dividamos $(x-2)(x-4)(x-6)$ por 3. Ou seja, o polinômio com as características procuradas é $p(x)=\frac{(x-2)(x-4)(x-6)}{(3-2)(3-4)(3-6)}=\frac{(x-2)(x-4)(x-6)}{3}$.

De modo geral, é facilmente verificável que o polinômio $L_0(x)=\frac{(x-x_1)(x-x_2)...(x-x_n)}{(x_0-x_1)(x_0-x_2)...(x_0-x_n)}$, o qual se anula para todos os elementos de $\{x_1, x_2, ..., x_n\}$ e satisfaz $L_0(x_0)=1$. Da mesma forma, podemos encontrar polinômios $L_1(x), L_2(x), ..., L_n(x)$ tais que $L_i(x_i)=1$ e $L_i(x_j)=0$, se $i\neq j$, cada um dos quais com grau *n*. Definimos, então, o polinômio

$$p(x)=y_0.L_0(x)+y_1.L_1(x)+...+y_n.L_n(x),$$

que é um polinômio de grau menor ou igual a *n* tal que:

$$p(x_0)=y_0.L_0(x_0)+y_1.L_1(x_0)+...+y_n.L_n(x_0)=y_0.1+y_1.0+...+y_n.0=y_0;$$
$$p(x_1)=y_0.L_0(x_1)+y_1.L_1(x_1)+...+y_n.L_n(x_1)=y_0.0+y_1.1+...+y_n.0=y_1$$
...
$$p(x_n)=y_0.L_0(x_n)+y_1.L_1(x_n)+...+y_n.L_n(x_n)=y_0.0+y_1.0+...+y_n.1=y_n,$$

ou seja, é o polinômio interpolador para o conjunto de dados $\{(x_0, y_0), (x_1, y_1), ..., (x_n, y_n)\}$.

Vejamos, a seguir, como aplicar o método de Lagrange.

Exemplo 2

Usando o método de Lagrange, encontre o polinômio interpolador para o conjunto de dados {(1, 3), (4, 18)}.

Solução:

O conjunto de dados contém dois pontos, logo o polinômio interpolador terá grau 1 e será da forma $p(x)=y_0.L_0(x)+y_1.L_1(x)$, sendo $(x_0, y_0)=(1,3)$ e $(x_1, y_1)=(4,18)$. Comecemos encontrando os polinômios elementares $L_0(x)$ e $L_1(x)$. Temos

$$L_0(x)=\frac{(x-x_1)}{(x_0-x_1)}=\frac{(x-4)}{(1-4)}=\frac{(x-4)}{-3} \text{ e}$$
$$L_1(x)=\frac{(x-x_0)}{(x_1-x_1)}=\frac{(x-1)}{(4-1)}=\frac{(x-1)}{3}.$$

Dessa maneira, encontraremos $p(x)=y_0.L_0(x)+y_1.L_1(x) = y_0.\frac{(x-4)}{-3}+y_1.\frac{(x-1)}{3} = 3.\frac{(x-4)}{-3}+18.\frac{(x-1)}{3} = -(x-4)+6(x-1) = 5x-2$.

Podemos escrever a definição dos polinômios elementares usando o símbolo de produtório (a letra grega Π) da seguinte forma:

$L_i(x) = \dfrac{\prod_{\substack{k=0 \\ k \neq i}}^{n} (x - x_k)}{\prod_{\substack{k=0 \\ k \neq i}}^{n} (x_i - x_k)}$ e o polinômio interpolador fica $p(x) = \sum_{i=0}^{n} y_i . L_i(x)$.

As expressões acima são apenas formas mais compactas de escrever o que já obtemos antes do exemplo. Na prática, ao procurar pelo polinômio interpolador, usa-se a forma extensa, pois precisaremos colocar os dados do conjunto.

Exemplo 3

(situação inicial da aula) Um móvel desloca-se em uma trajetória orientada de acordo com os seguintes dados:

t (em segundos)	s (em metros)
3	20
5	30
7	50

Usando interpolação polinomial, através do método de Lagrange, encontre uma estimativa para a posição do móvel para t = 4 s.

Solução:

Para o conjunto de dados $\{(3,20),\ (5,30),\ (7,50)\}$, o polinômio interpolador terá grau 2 (no máximo) da forma $p(x) = y_0 . L_0(x) + y_1 . L_1(x) + y_2 L_2(x)$, sendo $(x_0, y_0) = (3,20);\ (x_1, y_1) = (5,30)$ e $(x_2, y_2) = (7,50)$. Comecemos encontrando os polinômios elementares $L_0(x), L_1(x)$ e $L_2(x)$. Temos

$$L_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} = \frac{(x-5)(x-7)}{(3-5)(3-7)} = \frac{(x-5)(x-7)}{8};$$

$$L_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} = \frac{(x-3)(x-7)}{(5-3)(5-7)} = \frac{(x-3)(x-7)}{-4}$$

$$L_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} = \frac{(x-3)(x-5)}{(7-3)(7-5)} = \frac{(x-3)(x-5)}{8}.$$

Assim, o polinômio interpolador será da forma

$$p(x) = y_0 . L_0(x) + y_1 . L_1(x) + y_2 . L_2(x) = 20 . L_0(x) + 30 . L_1(x) + 50 . L_2(x)$$

$$= 20.\frac{(x-5)(x-7)}{8} + 30.\frac{(x-3)(x-7)}{-4} + 50.\frac{(x-3)(x-5)}{8}.$$

Como o objetivo não é encontrar o polinômio em si, não precisamos desenvolver os produtos. Podemos, apenas, substituir x = 4 para obter $p(4) = 20.\frac{(4-5)(4-7)}{8} + 30.\frac{(4-3)(4-7)}{-4} + 50.\frac{(4-3)(4-5)}{8} = 20.\frac{3}{8} + 30.\frac{(-3)}{-4} + 50.\frac{(-1)}{8} = 23{,}75$. Obviamente, encontramos o mesmo resultado do método direto.

Como sugestão para encerrar o tópico, recomendamos que você refaça os exemplos do tópico 1, usando o método de Lagrange, para que fique claro o uso da fórmula, com a comodidade de já sabermos as respostas.

TÓPICO 3

O método de Newton

Objetivos

- Apresentar o método de Newton para obtenção do polinômio interpolador
- Calcular diferenças divididas em um conjunto de dados

Neste tópico, ainda em relação ao problema de encontrar o polinômio interpolador, descreveremos um método atribuído ao famoso matemático inglês Isaac Newton (1643 - 1727).

Inicialmente, definamos *diferença dividida* para um conjunto de dados da seguinte forma:

Definição 1: Para o conjunto de dados $\{(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)\}$, a diferença dividida de ordem 0 em relação a x_i será dada por $\nabla_i^0 = y_i$.

Definição 2: Para o conjunto de dados $\{(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)\}$, a diferença dividida de ordem 1 em relação a x_i será dada por $\nabla_i^1 = \dfrac{\nabla_{i+1}^0 - \nabla_i^0}{x_{i+1} - x_i}$.

Observe que, nesta definição, podemos calcular os valores ∇_i^1 apenas para $i = 0, 1, \ldots, n-1$.

Exemplo 1

Para o conjunto de dados $\{(1,2),(3,7),(5,19)\}$, podemos calcular $\nabla_0^0 = y_0 = 2$; $\nabla_1^0 = y_1 = 7$ e $\nabla_2^0 = y_2 = 19$. Também podemos determinar $\nabla_0^1 = \dfrac{\nabla_1^0 - \nabla_0^0}{x_1 - x_0} = \dfrac{7-2}{3-1} = \dfrac{5}{2}$ e $\nabla_1^1 = \dfrac{\nabla_2^0 - \nabla_1^0}{x_2 - x_1} = \dfrac{19-7}{5-3} = \dfrac{12}{2} = 6$, mas não podemos calcular ∇_2^1, pois não há x_3.

Definição geral (por recorrência): Para o conjunto de dados $\{(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)\}$, a diferença dividida de ordem k em relação a x_i, com $1 \leq k \leq n$, será dada por $\nabla_i^k = \dfrac{\nabla_{i+1}^{k-1} - \nabla_i^{k-1}}{x_{i+k} - x_i}$.

Exemplo 1

(continuação): Para o conjunto de dados $\{(1,2), (3,7), (5,19)\}$, podemos calcular $\nabla_0^2 = \dfrac{\nabla_1^1 - \nabla_0^1}{x_2 - x_0} = \dfrac{6-(5/2)}{5-1} = \dfrac{7/2}{4} = \dfrac{7}{8}$ e organizar os resultados em uma tabela:

ATENÇÃO!

Na Definição geral, podemos determinar os valores ∇_i^k apenas para $i = 0, 1, \ldots, n-k$.

x_i	$\nabla_i^0 = y_i$	∇_i^1	∇_i^2
1	2	5/2	7/8
3	7	6	---
5	19	---	---

Assim, por exemplo, para encontrar a diferença dividida de ordem 4 de um determinado valor, precisamos das diferenças divididas de ordem 3 e, por isso, de todas as diferenças divididas de ordem menor que 4.

Exemplo 2

Para o conjunto de dados $\{(2,3), (3,5), (4,14), (5,27), (6,42)\}$, encontre o valor de ∇_0^4.

Solução:

Para que determinemos uma diferença dividida de ordem 4, devemos encontrar as diferenças divididas de todas as ordem menores que 4. Comecemos pelas de ordem 0:

$$\nabla_0^0 = y_0 = 3,\ \nabla_1^0 = y_1 = 5,\ \nabla_2^0 = y_2 = 14,\ \nabla_3^0 = y_3 = 27,\ \nabla_4^0 = y_4 = 42.$$

Seguimos para determinar as diferenças divididas de ordem 1:

$$\nabla_0^1 = \frac{\nabla_1^0 - \nabla_0^0}{x_1 - x_0} = \frac{5-3}{3-2} = 2,\ \nabla_1^1 = \frac{\nabla_2^0 - \nabla_1^0}{x_2 - x_1} = \frac{14-5}{4-3} = 9,$$

$$\nabla_2^1 = \frac{\nabla_3^0 - \nabla_2^0}{x_3 - x_2} = \frac{27-14}{5-4} = 13,\ \nabla_3^1 = \frac{\nabla_4^0 - \nabla_3^0}{x_4 - x_3} = \frac{42-27}{6-5} = 15.$$ Veja que

não há ∇_4^1, pois não existe x_5 para o conjunto de dados. Podemos guardar estes dados para referência na seguinte tabela:

x_i	$\nabla_i^0 = y_i$	∇_i^1	∇_i^2	∇_i^3	∇_i^4
2	3	2			
3	5	9			---
4	14	13		---	---
5	27	15	---	---	---
6	42	---	---	---	---

Encontremos, agora, as diferenças divididas de ordem 2:

$\nabla_0^2 = \dfrac{\nabla_1^1 - \nabla_0^1}{x_2 - x_0} = \dfrac{9-2}{4-2} = \dfrac{7}{2}$, $\nabla_1^2 = \dfrac{\nabla_2^1 - \nabla_1^1}{x_3 - x_1} = \dfrac{13-9}{5-3} = 2$ e

$\nabla_2^2 = \dfrac{\nabla_3^1 - \nabla_2^1}{x_4 - x_2} = \dfrac{15-13}{6-4} = 2$. Aqui não calculamos ∇_3^2, pois não existe x_5 para o conjunto de dados. Analisando o cálculo para essas diferenças divididas, observe que, no numerador, subtraímos as diferenças divididas consecutivas de ordem 1, mas, no denominador, não subtraímos x_i consecutivos, há um "salteamento".

Agora as diferenças divididas de ordem 3:

$\nabla_0^3 = \dfrac{\nabla_1^2 - \nabla_0^2}{x_3 - x_0} = \dfrac{2 - 7/2}{5-2} = \dfrac{-3/2}{3} = -\dfrac{1}{2}$ e $\nabla_1^3 = \dfrac{\nabla_2^2 - \nabla_1^2}{x_4 - x_1} = \dfrac{2-2}{6-3} = 0$.

Aqui não calculamos ∇_2^3, pois não existe x_5 para o conjunto de dados. Analisando o cálculo para essas diferenças divididas, observe que, no numerador, subtraímos as diferenças divididas consecutivas de ordem 2, mas, no denominador, não subtraímos xi consecutivos, há um "salteamento duplo".

Por último, com ordem 4:

$\nabla_0^4 = \dfrac{\nabla_1^3 - \nabla_0^3}{x_4 - x_0} = \dfrac{0 - (-1/2)}{6-2} = \dfrac{1/2}{4} = \dfrac{1}{8}$, que completa a tabela:

x_i	$\nabla_i^0 = y_i$	∇_i^1	∇_i^2	∇_i^3	∇_i^4
2	3	2	7/2	-1/2	1/8
3	5	9	2	0	---
4	14	13	2	---	---
5	27	15	---	---	---
6	42	---	---	---	---

Exemplo 3

Para o conjunto de dados $\{(1,3),\ (2,5),\ (3,9),\ (4,17),\ (5,33),\ (6,65)\}$, podemos construir a tabela (verifique):

x_i	$\nabla_i^0 = y_i$	∇_i^1	∇_i^2	∇_i^3	∇_i^4	∇_i^5
1	3	2	1	1/3	1/12	1/60
2	5	4	2	2/3	2/12	---
3	9	8	4	4/3	---	---
4	17	16	8	---	---	---
5	33	32	---	---	---	---
6	65	---	---	---	---	---

As diferenças divididas podem ser usadas para determinar o polinômio interpolador para um conjunto de dados de acordo com o que segue.

Proposição: Para o conjunto de dados $\{(x_0, y_0),\ (x_1, y_1),\ \ldots,\ (x_n, y_n)\}$, o polinômio interpolador pode ser obtido pela expressão:
$p(x) = y_0 + \nabla_0^1.(x - x_0) + \nabla_0^2.(x - x_0)(x - x_1) + \ldots + \nabla_0^n.(x - x_0)(x - x_1)\ldots(x - x_{n-1})$.

Vejamos como pode ser encontrado o polinômio interpolador pelo uso da proposição acima.

Exemplo 4

Usando o método de Newton, encontre o polinômio interpolador para os dados $\{(1,4),\ (3,8),\ (6,29)\}$.

Solução:

Fazendo $(x_0, y_0) = (1,4)$, $(x_1, y_1) = (3,8)$ e $(x_2, y_2) = (6,29)$, podemos encontrar as diferenças divididas

De ordem 0:

$\nabla_0^0 = y_0 = 4$, $\nabla_1^0 = y_1 = 8$, $\nabla_2^0 = y_2 = 29$.

De ordem 1:

$$\nabla_0^1 = \frac{\nabla_1^0 - \nabla_0^0}{x_1 - x_0} = \frac{8-4}{3-1} = 2 \text{ e } \nabla_1^1 = \frac{\nabla_2^0 - \nabla_1^0}{x_2 - x_1} = \frac{29-8}{6-3} = 7.$$

E de ordem 2:

$\nabla_0^2 = \dfrac{\nabla_1^1 - \nabla_0^1}{x_2 - x_0} = \dfrac{7-2}{6-1} = 1$, dados que podem ser tabelados:

x_i	$\nabla_i^0 = y_i$	∇_i^1	∇_i^2
1	**4**	**2**	**1**
3	**8**	**7**	**---**
6	**29**	**---**	**---**

Assim, o polinômio interpolador será

$$p(x) = y_0 + \nabla_0^1.(x - x_0) + \nabla_0^2.(x - x_0)(x - x_1) =$$

$$4 + 2.(x-1) + 1.(x-1)(x-3) = 4 + 2x - 2 + x^2 - 4x + 3 = x^2 - 2x + 5.$$

Exemplo 5

O volume de água em um reservatório foi medido em tempos regulares. Os resultados das medições aparecem na tabela abaixo. Usando interpolação polinomial, estime o volume de água no reservatório para t=2,5h.

t (em h)	0	1	2	3	4
V (em m³)	0	3	7	15	30

Solução:

Agrupando os dados $\{(0,0), (1,3), (2,7), (3,15), (4,30)\}$ na "tabela" do método de Newton, temos

x_i	$\nabla_i^0 = y_i$	∇_i^1	∇_i^2	∇_i^3	∇_i^4
0	0	3	1/2	1/2	0
1	3	4	2	1/2	---
2	7	8	7/2	---	---
3	15	15	---	---	---
4	30	---	---	---	---

Assim, o polinômio interpolador pode ser obtido por

$$p(x) = y_0 + \nabla_0^1.(x - x_0) + \nabla_0^2.(x - x_0)(x - x_1) + \nabla_0^3.(x - x_0)(x - x_1)(x - x_2)$$
$$+ \nabla_0^4.(x - x_0)(x - x_1)(x - x_2)(x - x_3)$$

$$p(x) = 0 + 3.(x - 0) + \frac{1}{2}.(x - 0)(x - 1) + \frac{1}{2}.(x - 0)(x - 1)(x - 2) + 0.(x - 0)$$
$$(x - 1)(x - 1)(x - 3)$$

$$p(x) = 3x + \frac{1}{2}x.(x-1) + \frac{1}{2}.x.(x-1)(x-2)$$

Para obter uma estimativa do volume do tanque para t=2,5h, calculamos $p(2,5) = 3.2,5 + \frac{1}{2}2,5.(2,5-1) + \frac{1}{2}.2,5.(2,5-1)(2,5-2) = 10,3125$, de onde podemos afirmar que o volume do tanque para t=2,5h é de, aproximadamente, 10,31m³.

Para encerrar a aula, acompanhe como a interpolação polinomial pode ser usada para aproximar raízes de funções.

Exemplo 6

Considere $f(x) = x^3 + 2x - 1$. Não há um método analítico simples para determinar as raízes de f, mas, como $f(0) = -1$ e $f(1) = 2$, temos a certeza de que a função *f* possui uma raiz entre 0 e 1 (ver Teorema de Bolzano). Uma aproximação para essa raiz pode ser obtida por algum dos métodos descritos nas primeiras aulas.

Algo diferente que podemos fazer é escolher um terceiro valor, de preferência perto de 0 e 1, substituir na função e obter três pontos, usar os três pontos para encontrar um polinômio *p*, de grau 2, que aproxime *f*, e aplicar a fórmula de Bhaskara para determinar a raiz de p que fica no intervalo considerado e usá-la como aproximação para a raiz de *f*.

Escolhendo, por exemplo, o número 0,5, temos $f(0,5) = 0,5^3 + 2.0,5 - 1 = 0,125$. Usemos então o conjunto de dados $\{(0;-1),\ (0,5;0,125),\ (1;2)\}$ e o método de Lagrange, começando pelos polinômios elementares $L_0(x), L_1(x)$ e $L_2(x)$. Temos

$$L_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{(x-0,5)(x-1)}{(0-0,5)(0-1)} = \frac{x^2-1,5x+0,5}{0,5};$$

$$L_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(x-0)(x-1)}{(0,5-0)(0,5-1)} = \frac{x^2-x}{-0,25};$$

$$L_2(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{(x-0)(x-0,5)}{(1-0)(1-0,5)} = \frac{x^2-0,5x}{0,5}.$$

Assim, o polinômio interpolador será da forma

$$p(x) = y_0.L_0(x) + y_1.L_1(x) + y_2.L_2(x) = (-1).L_0(x) + 0,125.L_1(x) + 2.L_2(x) =$$
$$= (-1).\frac{x^2-1,5x+0,5}{0,5} + 0,125.\frac{x^2-x}{-0,25} + 2.\frac{x^2-0,5x}{0,5} =$$
$$= -2(x^2-1,5x+0,5) - 0,5(x^2-x) + 4(x^2-0,5x) =$$

$$= -2x^2 + 3x - 1 - 0,5x^2 + 0,5x + 4x^2 - 2x =$$

$$= 1,5x^2 + 1,5x - 1 .$$

Dessa forma, podemos usar a fórmula de Bhaskara para o polinômio $p(x) = 1,5x^2 + 1,5x - 1$, resultando na raiz positiva

$$x = \frac{-1,5 + \sqrt{1,5^2 - 4.1,5.(-1)}}{2.1,5} \cong 0,457 .$$

Assim, como no final do tópico 2, sugerimos que os exemplos dos tópicos anteriores sejam refeitos através do método de Newton e que se analise as vantagens e desvantagens dos métodos descritos.

AULA 7

Integração Numérica

Olá alunos! Sejam bem-vindos.

Nesta nova aula, aproximaremos os valores das integrais definidas, como visto no Cálculo I. Recomendamos que você revise os conceitos aprendidos naquela disciplina, especialmente o de integral de Riemann, para que possamos tirar o maior proveito possível do estudo que se inicia agora.

Objetivos

- Descrever métodos de integração numérica
- Comparar métodos e aplicar processos de aproximação de funções

TÓPICO 1

Revisão de conceitos e definições iniciais

Objetivos

- Revisar os conceitos necessários para a formulação do problema
- Resolver problemas iniciais

Um problema central com o qual lidamos no Cálculo Diferencial e Integral é o que segue:

Problema: Encontre a área da região do plano cartesiano limitada pelo gráfico da função contínua $f:[a,b]\to\mathbb{R}_+$, pelo eixo x e pelas retas $x=a$ e $x=b$ (ver figura 1).

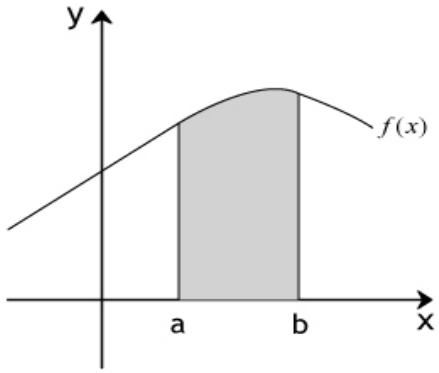


Figura 1: Área da região limitada pelas retas $x=a$ e $x=b$

No curso de Cálculo I, vimos que o problema pode ser resolvido a partir da determinação da integral definida $\int_a^b f(x)dx$ e que uma regra prática para se encontrar esse valor é dada pelo seguinte resultado crucial:

Teorema Fundamental do Cálculo: *Se* $f:[a,b]\to\mathbb{R}$ *é uma função contínua, e* F *é uma primitiva de* f *em* (a,b)*, ou seja, vale* $\frac{dF}{dx}(x)=f(x), \forall x\in(a,b)$*, então* $\int_a^b f(x)dx=F(b)-F(a)$.

Exemplo 1

Calcule a área da região do plano cartesiano limitada pelo gráfico de $f(x)=2x-1$, pelo eixo x e pela retas $x=1$ e $x=2$.

Solução:

Um esboço da região considerada pode ser visto na figura 2. Como a função não assume valores negativos no intervalo $[1,2]$, podemos calcular a área por $\int_1^2 (2x-1)dx$. Para tanto, encontramos uma primitiva para a função. É imediato verificar que $F(x)=x^2-x$ é uma primitiva para a função dada. Assim, usando o Teorema Fundamental do Cálculo, obtemos $\int_1^2 (2x-1)dx = F(2)-F(1)=2$.

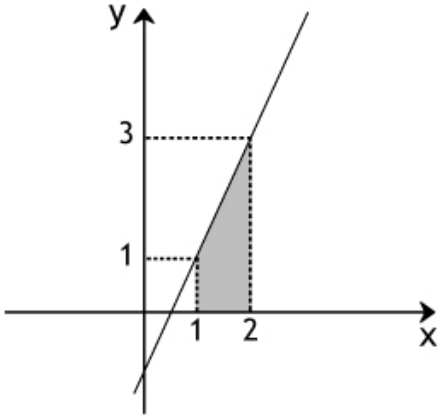


Figura 2: Gráfico da função $f(x)=2x-1$

Exemplo 2

Uma vez que $F(x)=x^3$ é uma primitiva para $f(x)=3x^2$, podemos, usando o Teorema Fundamental do Cálculo, escrever $\int_2^5 3x^2dx = \left[x^3\right]_{x=2}^{x=5} = 5^3-2^3 = 125-8 = 117$.

Embora a motivação inicial para o cálculo de integrais venha da Geometria Plana, na qual não interessam medidas negativas, podemos encontrar, via TFC, o valor de integrais definidas mesmo que as funções assumam valores negativos.

Exemplo 3

Visto que $F(x)=\frac{x^4}{4}$ é uma primitiva para $f(x)=x^3$, podemos, usando o Teorema Fundamental do Cálculo, escrever $\int_{-1}^{0} x^3dx = \left[\frac{x^4}{4}\right]_{x=-1}^{x=0} = \frac{0^4}{4}-\frac{(-1)^4}{4} = -\frac{1}{4}$.

As integrais definidas têm aplicação em várias áreas, com interpretações diversas (áreas, espaço percorrido, volume, trabalho, etc.); entretanto há duas

situações nas quais a determinação de seu valor pela aplicação do Teorema Fundamental do Cálculo é impraticável. Vejamos quais:

Situação 1

Para se encontrar o valor de $\int_a^b f(x)dx$, precisamos de uma primitiva para a função $f(x)$, o que pode ser bem difícil ou mesmo impossível de se obter por funções simples. Por exemplo, as funções $\frac{e^x}{x}$, e^{x^2}, $\sqrt{1+x^3}$ e $\frac{1}{\ln x}$ não possuem primitivas elementares, ou seja, não podemos determinar exatamente o valor de $\int_0^1 e^{x^2}dx$, $\int_2^e \frac{1}{\ln x}dx$ ou $\int_0^1 \sqrt{1+x^3}dx$ através das funções que estudamos nos cursos iniciais de Cálculo.

Situação 2

Outra impossibilidade de determinação do valor exato da integral é quando a função é obtida a partir de um experimento (por instrumentos de medida ou por dados coletados), caso no qual podemos não ter uma fórmula para expressá-la ou, por conhecê-la apenas em pontos isolados, não temos a confirmação do seu comportamento em intervalos.

A integração numérica estabelece métodos de aproximação para essas integrais, mas que, obviamente, também podem ser usados nos casos nos quais conhecemos a primitiva para a função, mas saber o valor exato da integral não é o objetivo ou não é algo simples de ser feito sem o uso de calculadoras, como é o caso da função $f(x)=\frac{1}{x}$, da qual conhecemos uma primitiva $F(x)=\ln x$. Assim, $\int_2^3 \frac{1}{x}dx = \ln 3 - \ln 2 = \ln\frac{3}{2}$, entretanto, pode ser que trabalhar com a função gere uma complexidade menor que fazer uma aproximação para o logaritmo.

TÓPICO 2

Soma de Riemann

OBJETIVOS

- Apresentar o método de integração por somas de Riemann
- Analisar geometricamente o método

Comecemos aqui recordando a definição de Integral de Riemann

Fonte: http://pt.wikipedia.org/

Figura 3: Georg Riemann

Dada a função contínua $f:[a,b]\to\mathbb{R}_+$, dividimos o intervalo considerado em n subintervalos de igual comprimento $\Delta x=\frac{b-a}{n}$ (ou seja, fazemos uma partição *uniforme* de $[a,b]$) e escolhemos em cada subintervalo $[x_{i-1},x_i]$ um valor qualquer x_i^*. Dessa forma, temos, por definição:

$$\int_a^b f(x)dx=\lim_{n\to\infty}\sum_{i=1}^{n} f\left(x_i^*\right)\Delta x\,.$$

Na definição acima x_i^*, cada pode ser escolhido como o final do intervalo, o começo, o ponto de máximo, o ponto de mínimo, o ponto médio ou qualquer outro ponto já que o resultado permaneceria o mesmo ao realizar o processo de limite. Um método que podemos usar para aproximar o valor da integral é considerar apenas a soma de Riemann para uma quantidade fixa de subintervalos, pois assim aproximaremos

$$\int_a^b f(x)dx\approx\sum_{i=1}^{n} f\left(x_i^*\right)\Delta x\,.$$

Na figura 4 a seguir, temos a interpretação geométrica desta aproximação, considerando x_i^* como o mínimo em cada subintervalo.

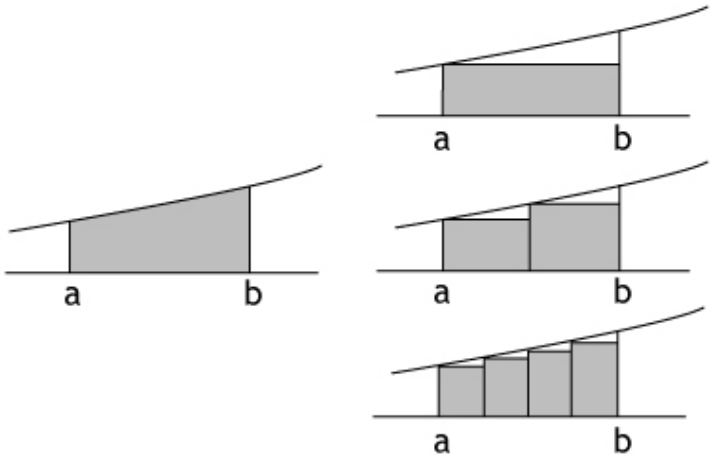


Figura 4: Área desejada (à esquerda) e suas aproximações (à direita) por somas inferiores de Riemann com 1, 2 e 4 subintervalos

Exemplo 1

Usando soma de Riemann, quatro subintervalos e escolhendo x_i^* como o final de cada subintervalo, aproxime $\int_0^2 e^{x^2} dx$.

Solução:

Inicialmente, dividimos o intervalo $[0;2]$ em quatro subintervalos, cada um deles com comprimento $\Delta x = \frac{2-1}{4} = 0,5$.

→ no primeiro subintervalo $[0;0,5]$, obtemos $x_1^* = 0,5$ e assim $f\left(x_1^*\right)\Delta x = e^{0,5^2}.0,5 = e^{0,25}.0,5$.

→ no segundo subintervalo $[0,5;1]$, temos $x_2^* = 1$, portanto encontraremos $f\left(x_2^*\right)\Delta x = e^{1^2}.0,5 = e.0,5$.

→ no terceiro subintervalo $[1;1,5]$, encontramos $x_3^* = 1,5$ e, por conseguinte, $f\left(x_3^*\right)\Delta x = e^{1,5^2}.0,5 = e^{2,25}.0,5$.

→ no quarto subintervalo $[1,5;2]$, temos $x_4^* = 2$ e assim $f\left(x_4^*\right)\Delta x = e^{2^2}.0,5 = e^4.0,5$

Desse modo, podemos aproximar

$$\int_0^2 e^{x^2} dx \approx \sum_{i=1}^{4} f\left(x_i^*\right)\Delta x = f\left(x_1^*\right)\Delta x + f\left(x_2^*\right)\Delta x + f\left(x_3^*\right)\Delta x + f\left(x_4^*\right)\Delta x =$$

$$= e^{0,25}.0,5 + e^1.0,5 + e^{2,25}.0,5 + e^4.0,5 =$$

$$= 0,5.(e^{0,25} + e + e^{2,25} + e^4) \cong 0,5.68,08 = 34,04$$

Como a função $f(x)=e^{x^2}$ é crescente em $[0;2]$, escolher o ponto final de cada subintervalo equivale a escolher o ponto de máximo, assim a aproximação feita no exemplo é por excesso, de onde podemos concluir que o valor exato da integral é menor que 34,04.

Exemplo 2

Usando soma de Riemann, cinco subintervalos e escolhendo x_i^* como o ponto médio de cada subintervalo, aproxime $\int_1^2 \frac{1}{x}dx$.

Solução:

Inicialmente, dividimos o intervalo $[1;2]$ em cinco subintervalos, cada um deles com comprimento $\Delta x=\frac{1}{5}$. Para a função $f(x)=\frac{1}{x}$:

→ no primeiro subintervalo $\left[1,\frac{6}{5}\right]$, temos $x_1^*=\frac{11}{10}$ e, assim, $f\left(x_1^*\right)\Delta x=\frac{1}{11/10}.\frac{1}{5}=\frac{2}{11}$;

→ no segundo subintervalo $\left[\frac{6}{5},\frac{7}{5}\right]$, temos $x_2^*=\frac{13}{10}$ e, assim, $f\left(x_2^*\right)\Delta x=\frac{1}{13/10}.\frac{1}{5}=\frac{2}{13}$;

→ no terceiro subintervalo $\left[\frac{7}{5},\frac{8}{5}\right]$, temos $x_3^*=\frac{15}{10}$ e, assim, $f\left(x_3^*\right)\Delta x=\frac{1}{15/10}.\frac{1}{5}=\frac{2}{15}$;

→ no quarto subintervalo $\left[\frac{8}{5},\frac{9}{5}\right]$, temos $x_4^*=\frac{17}{10}$ e, assim, $f\left(x_4^*\right)\Delta x=\frac{1}{17/10}.\frac{1}{5}=\frac{2}{17}$;

→ no quinto subintervalo $\left[\frac{9}{5},2\right]$, temos $x_5^*=\frac{19}{10}$ e, assim, $f\left(x_5^*\right)\Delta x=\frac{1}{19/10}.\frac{1}{5}=\frac{2}{19}$.

Desse modo, podemos aproximar

$$\int_1^2 \frac{1}{x}dx \approx \sum_{i=1}^{5} f\left(x_i^*\right)\Delta x = f\left(x_1^*\right)\Delta x+f\left(x_2^*\right)\Delta x+f\left(x_3^*\right)\Delta x+f\left(x_4^*\right)\Delta x+f\left(x_5^*\right)\Delta x=$$

$$=\frac{2}{11}+\frac{2}{13}+\frac{2}{15}+\frac{2}{17}+\frac{2}{19}=$$

$$=2.\left(\frac{1}{11}+\frac{1}{13}+\frac{1}{15}+\frac{1}{17}+\frac{1}{19}\right)\cong 0,692.$$

O exemplo 2 pode ser usado para se obter uma aproximação de $\ln 2$, pois,

pelo Teorema Fundamental do Cálculo, temos $\int_1^2 \frac{1}{x} dx = [\ln x]_{x=1}^{x=2} = \ln 2 - \ln 1 = \ln 2$. Assim, obtemos $\ln 2 \cong 0,692$.

Observação 1: Quanto maior for a quantidade de subintervalos, melhor será a aproximação, independente da escolha do x_i^*.

Observação 2: Escolhendo x_i^* como sendo o máximo em cada subintervalo, teremos uma aproximação por excesso e, escolhendo x_i^* como sendo o mínimo em cada subintervalo, teremos uma aproximação por falta. Em geral, a melhor aproximação da integral por soma de Riemann será feita pela escolha do ponto médio.

TÓPICO 3

A regra dos trapézios

OBJETIVOS

- Apresentar e justificar a regra dos trapézios para integração numérica
- Analisar geometricamente o método

No tópico anterior, analisamos aproximações de integral por somas de Riemann, que consistem em somas de áreas de retângulos. No presente tópico, faremos uma aproximação por trapézios, como o nome da regra sugere. Acompanhe a situação na figura 5.

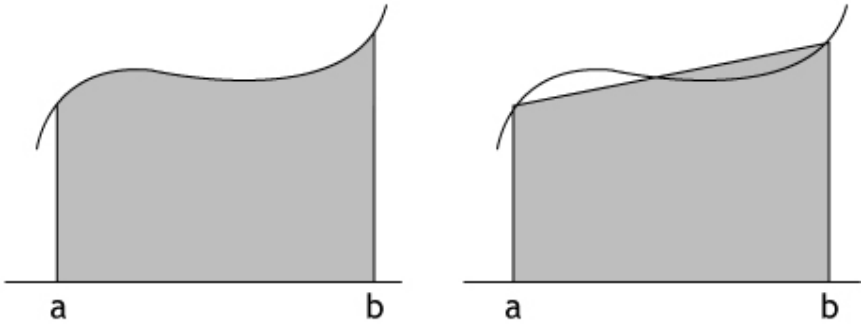


Figura 5: Área pretendida (à esquerda) e aproximação por um trapézio (à direita)

Relembrando que, se um trapézio tem bases de medidas B e b, e altura h, então sua área vale $\frac{h}{2}.(B+b)$. Na situação do gráfico de $f(x)$, a altura do trapézio é o comprimento do intervalo e as bases medem $f(b)$ e $f(a)$. Assim, podemos aproximar

$$\int_a^b f(x)dx \approx \frac{b-a}{2}(f(b)+f(a)).$$

EXEMPLO 1

Use a regra do trapézio para estimar o valor da integral $\int_0^2 \sqrt{1+x^3}dx$.

Solução:

Para a função $f(x)=\sqrt{1+x^3}$, podemos fazer

$$\int_0^2 \sqrt{1+x^3}\,dx \approx \frac{2-0}{2}.(f(2)+f(0)) = 1.\left(\sqrt{1+2^3}+\sqrt{1+0^3}\right) = \sqrt{9}+\sqrt{1} = 4.$$

Podemos também dividir o intervalo considerado e aplicar a regra do trapézio em cada um dos subintervalos, de acordo com o esquema abaixo, no qual $h=\Delta x=\frac{b-a}{n}$:

$$\begin{aligned}\int_a^b f(x)dx &= \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \ldots + \int_{x_{n-1}}^{x_n} f(x)dx \approx \\ &\approx \frac{h}{2}\big(f(x_0)+f(x_1)\big) + \frac{h}{2}\big(f(x_1)+f(x_2)\big) + \ldots + \frac{h}{2}\big(f(x_{n-1})+f(x_n)\big) = \\ &= \frac{h}{2}\big(f(x_0)+2.f(x_1)+\ldots+2.f(x_{n-1})+f(x_n)\big).\end{aligned}$$

Observe a figura a seguir na qual a regra do trapézio foi usada para quatro subintervalos.

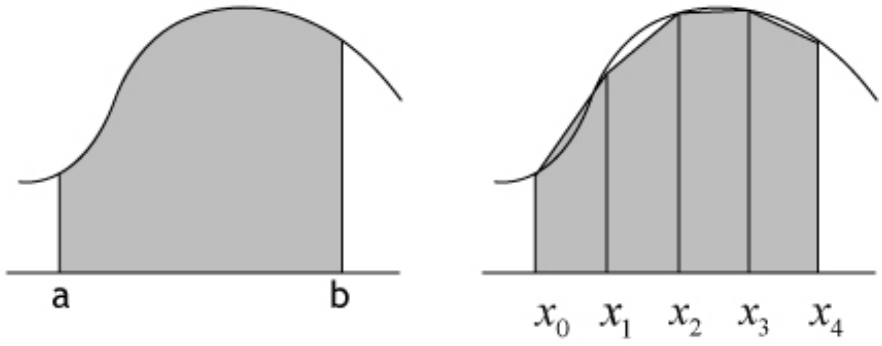


Figura 6: Aproximação pela regra do trapézio com quatro subintervalos

Exemplo 2

Se $p(x)$ é o polinômio interpolador para o conjunto de dados $\{(1,3),(2,7),(3,15),(4,31),(5,59)\}$, encontre uma aproximação para o valor de $\int_1^5 p(x)dx$.

Solução:

Aqui temos o caso no qual a função que vamos integrar é desconhecida, mas sabemos quanto ela vale em alguns pontos específicos. Se considerarmos os subintervalos $[1,2]$, $[2,3]$, $[3,4]$ e $[4,5]$, podemos aproximar o valor de $\int_1^5 p(x)dx$ pela regra dos trapézios, pois sabemos que $p(1)=3$, $p(2)=7$, $p(3)=15$, $p(4)=31$ e $p(5)=59$. Assim

$$\int_1^5 p(x)dx \approx \frac{h}{2}\left(p(x_0)+2.p(x_1)+2.p(x_2)+2.p(x_3)+p(x_4)\right)=$$

$$=\frac{1}{2}(3+2.7+2.15+2.31+59)=\frac{1}{2}.168=84.$$

Exemplo 3

Usando as técnicas de integração vistas no Cálculo, podemos obter $\int_0^1 \frac{1}{1+x^2}dx = \text{arctg } 1 - \text{arctg } 0 = \frac{\pi}{4}-0=\frac{\pi}{4}$. Assim, se fizermos uma aproximação para o valor de $\int_0^1 \frac{1}{1+x^2}dx$, teremos uma aproximação para $\frac{\pi}{4}$ e, multiplicando por 4, uma aproximação para π.

Usemos aqui a regra dos trapézios para cinco subintervalos (de comprimento 0,2), os pontos considerados são

$x_0=0$; $x_1=0,2$; $x_2=0,4$; $x_3=0,6$; $x_4=0,8$ e $x_5=1$.

Assim, para a função $f(x)=\frac{1}{1+x^2}$, encontramos

$$f(x_0)=\frac{1}{1+0^2}=1;\ f(x_1)=\frac{1}{1+0,2^2}=\frac{1}{1,04};\ f(x_2)=\frac{1}{1+0,4^2}=\frac{1}{1,16};$$

$$f(x_3)=\frac{1}{1+0,6^2}=\frac{1}{1,36};\ f(x_4)=\frac{1}{1+0,8^2}=\frac{1}{1,64} \text{ e } f(x_5)=\frac{1}{1+1^2}=\frac{1}{2}.$$

A partir daí, a aproximação ficará

$$\int_0^1 f(x)dx \approx \frac{0,2}{2}\left(f(x_0)+2.f(x_1)+2.f(x_2)+2.f(x_3)+2.f(x_4)+f(x_5)\right)=$$

$$=0,1\left(1+2.\frac{1}{1,04}+2.\frac{1}{1,16}+2.\frac{1}{1,36}+2.\frac{1}{1,64}+\frac{1}{2}\right)\approx 0,1.7,837=0,7837.$$

Logo, encontramos uma aproximação para $\pi \cong 4.0,7837=3,1348$.

Em geral, a regra dos trapézios oferece uma aproximação equivalente àquela obtida pela soma de Riemann com ponto médio, mas tem vantagem sobre as outras escolhas de pontos, especialmente em funções de crescimento acentuado.

Usando soma de Riemann, aproximamos a função em cada subintervalo por uma função constante, ou seja, de grau 0. A regra dos trapézios aproxima o gráfico da função em cada subintervalo por um segmento de reta, isto é, o gráfico de uma função de primeiro grau. O próximo passo será aproximar as funções em cada subintervalo por uma parábola, ou seja, por uma função de segundo grau, e, para tanto, podemos fazer uso de interpolação polinomial. Por ora, sugerimos que você refaça os exemplos do tópico 1, usando a regra dos trapézios, e compare os resultados obtidos.

TÓPICO 4

A regra de Simpson

Objetivos

- Apresentar e justificar a regra de Simpson para integração numérica
- Analisar geometricamente o método

Nos tópicos iniciais, vimos como aproximar o gráfico de uma função por segmentos de reta, horizontais (soma de Riemann) ou não (regra dos trapézios), com o objetivo de encontrar o valor aproximado da integral da função. Neste tópico, aproximaremos as funções por arcos de parábola, ou seja, por funções de segundo grau. Vimos, na aula passada, que um polinômio de segundo grau fica bem determinado por três pontos. Assim, precisaremos de três pontos do intervalo e não apenas dos extremos, como nos métodos anteriores. Observe a figura 7:

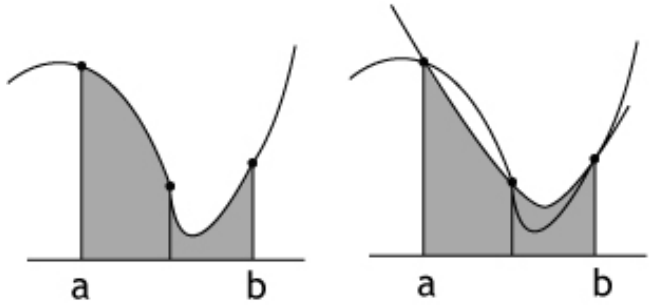


Figura 7: Área pretendida (à esquerda) e sua aproximação por um arco de parábola (à direita)

Por simplicidade, consideraremos os pontos igualmente espaçados, sendo a origem o ponto médio. Serão, portanto, os pontos $x_0 = -h$, $x_1 = 0$ e $x_2 = h$ com imagens y_0, y_1 e y_2, respectivamente. Escrevendo o polinômio interpolador para estes dados como $p(x) = ax^2 + bx + c$, teremos

$$\int_{x_0}^{x_2} p(x)dx = \int_{-h}^{h} \left(ax^2+bx+c\right)dx =$$

$$= \left[\frac{ax^3}{3}+\frac{bx^2}{2}+cx\right]_{x=-h}^{x=h} =$$

$$= \left(\frac{ah^3}{3}+\frac{bh^2}{2}+ch\right)-\left(\frac{-ah^3}{3}+\frac{bh^2}{2}-ch\right) =$$

$$= \frac{2ah^3}{3}+2ch^2 = \frac{h}{3}\left(2ah^2+6c\right).$$

Porém, como a parábola passa pelos pontos $(-h, y_0)$, $(0, y_1)$ e (h, y_2), devemos ter

$$y_0 = a(-h)^2+b(-h)+c = ah^2-bh+c$$

$$y_1 = a.0^2+b.0+c = c \text{ e}$$

$$y_2 = ah^2+bh+c\,.$$

Assim, obtemos $y_0+4y_1+y_2 = 2ah^2+6c$, de modo que podemos escrever a integral de $p(x)$ na forma $\int_{x_0}^{x_2} p(x)dx = \frac{h}{3}\left(2ah^2+6c\right) = \frac{h}{3}\left(y_0+4y_1+y_2\right)$.

Agora, fazendo a parábola mover-se horizontalmente para outros pontos $x_0, x_1 = x_0+h$ e $x_2 = x_1+h$, com imagens y_0, y_1 e y_2, a área sob a parábola não se altera. Desse modo, podemos enunciar que

ATENÇÃO!

O resultado enunciado na proposição (Regra de Simpson) já era conhecido por matemáticos do século XVII, mas foi popularizado nos textos do britânico Thomas Simpson (1710 – 1761), reconhecido por muitos como um dos melhores matemáticos ingleses do século XVIII. Em sua homenagem, damos ao método o nome de *Regra de Simpson*.

Proposição (Regra de Simpson)

Se $f(x)$ é uma função contínua, e os pontos (x_0, y_0), (x_1, y_1) e (x_2, y_2) do gráfico de $f(x)$ estão igualmente espaçados horizontalmente, ou seja, se $x_2-x_1 = x_1-x_0 = h$, então podemos aproximar:

$$\int_{x_0}^{x_2} f(x)dx \cong \frac{h}{3}\left(y_0+4y_1+y_2\right).$$

Exemplo 1

Usando a regra de Simpson, faça uma aproximação para $\int_0^1 e^{-x^2}dx$.

Solução:

Para usar a fórmula acima, precisamos de três pontos igualmente espaçados.

Tomemos, então, o ponto médio do intervalo e calculemos

$x_0 = 0$, logo $y_0 = e^{-0^2} = 1$;

$x_1 = 0{,}5$, logo $y_1 = e^{-0{,}5^2} \cong 0{,}7788$ e

$x_2 = 1$, logo $y_2 = e^{-1^2} \cong 0{,}3679$.

Como o intervalo tem comprimento 1, vale $h = \frac{1}{2} = 0{,}5$. Assim, podemos aproximar

$$\int_0^1 e^{-x^2} dx \approx \frac{0{,}5}{3}(y_0 + 4y_1 + y_2)$$

$$\cong \frac{0{,}5}{3}(1 + 4.0{,}7788 + 0{,}3679) \cong 0{,}7472.$$

ATENÇÃO!

1. Por causa da expressão obtida, a regra de aproximação acima também recebe o nome de regra 1/3 de Simpson.
2. Os valores de y_i aparecem na expressão abaixo obedecendo à seguinte regra: o primeiro e o último serão multiplicados por 1, e os demais, alternadamente, multiplicados por 4 e 2, sempre começando por 4.

Exemplo 2

Use a regra de Simpson para estimar o valor da integral $\int_2^3 \sqrt{1+x^3}\,dx$.

Solução:

De maneira análoga ao exemplo 1, precisamos de três pontos igualmente espaçados. Tomemos, então, o ponto médio do intervalo $[2,3]$ e calculemos

$x_0 = 2$, logo $y_0 = \sqrt{1+2^3} = 3$;

$x_1 = 2{,}5$, logo $y_0 = \sqrt{1+2{,}5^3} \cong 4{,}0774$ e

$x_2 = 3$, logo $y_0 = \sqrt{1+3^3} = 5{,}2915$.

Assim, podemos aproximar, para $h = \frac{3-2}{2} = 0{,}5$

$$\int_2^3 \sqrt{1+x^3}\,dx = \frac{0{,}5}{3}(y_0 + 4y_1 + y_2) \cong \frac{0{,}5}{3}(3 + 4.4{,}0744 + 5{,}2915) \cong 4{,}0982.$$

Por fim, podemos refinar a regra de Simpson, usando-a repetidamente. Se dividirmos o intervalo $[a,b]$ em n subintervalos, essa quantidade deve ser par, a fim de que possamos aplicar a regra "de dois em dois". Acompanhe o esquema, no qual $h = \frac{b-a}{n}$:

$$\int_a^b f(x)dx = \int_{x_0}^{x_2} f(x)dx + \int_{x_2}^{x_4} f(x)dx + \ldots + \int_{x_{n-2}}^{x_n} f(x)dx \approx$$

$$\approx \frac{h}{3}\left(y_0+4y_1+y_2\right)+\frac{h}{3}\left(y_2+4y_3+y_4\right)+\ldots+\frac{h}{3}\left(y_{n-2}+4y_{n-1}+y_n\right)=$$
$$=\frac{h}{3}\left(y_0+4y_1+2y_2+4y_3+2y_4+\ldots+2y_{n-2}+4y_{n-1}+y_n\right).$$

Exemplo 3

Usando a regra de Simpson para oito subintervalos, aproxime $\int_1^3 \frac{1}{x}dx$.

Solução:

Aqui o intervalo $[1,3]$ deve ser dividido em oito partes iguais, cada uma de comprimento $h=\frac{3-1}{8}=\frac{2}{8}=0,25$. Assim, os valores a serem empregados são:

$$x_0=1\Rightarrow y_0=\frac{1}{1};\ x_1=1,25\Rightarrow y_1=\frac{1}{1,25};\ x_2=1,5\Rightarrow y_2=\frac{1}{1,5};$$

$$x_3=1,75\Rightarrow y_3=\frac{1}{1,75};\ x_4=2\Rightarrow y_0=\frac{1}{2};\ x_5=2,25\Rightarrow y_5=\frac{1}{2,25};$$

$$x_6=2,5\Rightarrow y_6=\frac{1}{2,5};\ x_7=2,75\Rightarrow y_7=\frac{1}{2,75}\ \text{e}\ x_8=3\Rightarrow y_8=\frac{1}{3}.$$

Logo, podemos fazer a aproximação:

$$\int_1^3 \frac{1}{x}dx \approx \frac{h}{3}\left(y_0+4y_1+2y_2+4y_3+2y_4+4y_5+2y_6+4y_7+y_8\right)=$$
$$=\frac{0,25}{3}\left(1+4.\frac{1}{1,25}+2.\frac{1}{1,5}+4.\frac{1}{1,75}+2.\frac{1}{2}+4.\frac{1}{2,25}+2.\frac{1}{2,5}+4.\frac{1}{2,75}+\frac{1}{3}\right)\cong$$
$$\cong 0,0833\left(1+3,2+1,3333+2,2857+1+1,7778+0,8+1,4545+0,3333\right)\cong$$
$$\cong 1,098277$$

Podemos usar o valor acima como aproximação para $\ln 3\cong 1,098277$.

Exemplo 4

Foram feitas medições regulares na largura de uma piscina, de dois em dois metros, com resultados apresentados na figura abaixo, na qual as unidades estão em metros. Sabendo que a piscina tem profundidade constante de 1,5 m, use a regra de Simpson para estimar a sua capacidade.

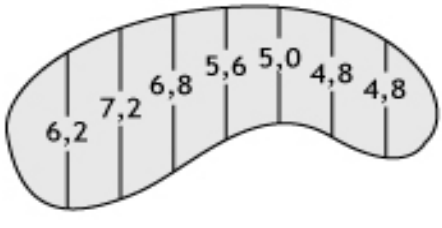


Figura 9: Planta de uma piscina

Solução:

Inicialmente, relembremos que o volume de um sólido de altura constante pode ser encontrado multiplicando-se a altura pela área da "base". Devemos, para começar, aproximar a área da piscina. Como as medições foram feitas de 2 em 2 metros, podemos considerar para cada x o valor da largura correspondente, de acordo com a seguinte tabela:

x	0	2	4	6	8	10	12	14	16
Largura (L)	0	6,2	7,2	6,8	5,6	5,0	4,8	4,8	0

Desse modo, podemos aproximar a área da piscina pela integral $\int_0^{16} L(x)dx$ e usar a divisão feita pelas medições, ou seja, $h = 2$. Fazendo as contas, obtemos

$$\begin{aligned}\int_0^{16} L(x)dx &\approx \frac{h}{3}\left(y_0 + 4y_1 + 2y_2 + 4y_3 + 2y_4 + 4y_5 + 2y_6 + 4y_7 + y_8\right) = \\ &= \frac{2}{3}(0 + 4.6{,}2 + 2.7{,}2 + 4.6{,}8 + 2.5{,}6 + 4.5{,}0 + 2.4{,}8 + 4.4{,}8 + 0) = \\ &= \frac{2}{3}.126{,}4 = \frac{252{,}8}{3}.\end{aligned}$$

Multiplicando o resultado acima pela profundidade da piscina, obteremos uma aproximação para o seu volume. O resultado é $\frac{252{,}8}{3}.1{,}5 = 126{,}4$ metros cúbicos. Uma estimativa para a capacidade da piscina é, portanto, de 126 400 litros.

Depois desse exemplo, chegamos ao fim da aula. Sugerimos que você refaça alguns exemplos usando um método diferente daquele empregado no texto. Compare os resultados e decida quais são mais precisos. Em geral, a regra de Simpson oferece uma aproximação melhor que os outros métodos e/ou com uma quantidade diferente de subintervalos. Se dispuser de um sistema computacional que calcule integrais, compare os resultados obtidos.

AULA 8

O método dos mínimos quadrados

Olá a todos!

Dando prosseguimento ao nosso estudo de aproximação de dados por funções conhecidas, trataremos nesta aula do problema dos mínimos quadrados. Um caso simples é o de encontrar a reta que melhor se ajusta a três ou mais pontos não alinhados. Há algumas maneiras de medir o quanto a função de aproximação difere dos dados do problema. Aqui levaremos em consideração a distância entre os pontos dados e os pontos aproximados ou, equivalentemente, o quadrado dessa distância.

Precisaremos de conceitos iniciais do trato de funções e análise de gráficos. Não hesite em recorrer a outras fontes, como o material de disciplinas anteriores, para revisar esses assuntos. Vamos ao trabalho, então?!

Objetivos

- Aproximar dados por funções conhecidas minimizando as distâncias
- Apresentar e discutir métodos e casos do problema de mínimos quadrados

TÓPICO 1

O caso linear discreto

OBJETIVOS

- Descrever aproximação de dados por funções
- Definir desvios quadrados
- Formular o problema dos mínimos quadrados para o caso linear

Em nossos estudos de Interpolação Polinomial, vimos como obter um polinômio que sirva de modelo para descrever certos fenômenos, visando à coincidência de pontos dados com os pontos gerados. Sabemos que há uma única parábola que passa por três pontos não colineares.

EXEMPLO 1

Encontre a equação da parábola que passa pelos pontos (1, 6), (2, 13) e (4, 45).

Solução:

Uma parábola tem equação do tipo $y = ax^2 + bx + c$ e como queremos que passe pelos pontos dados, devemos ter:

$$\begin{aligned} x = 1, y = 6 \quad &\Rightarrow a.1^2 + b.1 + c = 6 \quad \Rightarrow a + b + c = 6 \\ x = 2, y = 13 \quad &\Rightarrow a.2^2 + b.2 + c = 13 \quad \Rightarrow 4a + 2b + c = 13 \\ x = 4, y = 45 \quad &\Rightarrow a.4^2 + b.4 + c = 45 \quad \Rightarrow 16a + 4b + c = 45 \end{aligned}$$

Resolvendo, então, o sistema $\begin{cases} a + b + c = 6 \\ 4a + 2b + c = 13 \\ 16a + 4b + c = 45 \end{cases}$, obtemos $a = 3, b = -2$ e $c = 5$,

de onde podemos escrever a equação da parábola $y = 3x^2 - 2x + 5$.

Note que, no exemplo anterior, o sistema linear obtido é possível e determinado, ou seja, a solução é única. Se quiséssemos encontrar uma função de terceiro grau para os mesmos pontos, teríamos várias soluções, o que nos daria

mais alternativas. Um problema surge quando temos que aproximar um conjunto de $n+1$ dados por um polinômio de grau menor que n.

Exemplo 2

Encontre a equação da reta (função do primeiro grau) que passa pelos pontos (1, 6), (2, 13) e (4, 45).

Solução:

Uma reta tem equação do tipo $y = ax + b$ e como queremos que passe pelos pontos dados, devemos ter:

$$\begin{aligned} x=1, y=6 \quad &\Rightarrow a.1+b=6 \quad &&\Rightarrow a+b=6 \\ x=2, y=13 \quad &\Rightarrow a.2+b=13 \quad &&\Rightarrow 2a+b=13 \\ x=4, y=45 \quad &\Rightarrow a.4+b=45 \quad &&\Rightarrow 4a+b=45 \end{aligned}$$

Como o sistema $\begin{cases} a+b=6 \\ 2a+b=13 \\ 4a+b=45 \end{cases}$ possui três equações e duas incógnitas, uma maneira de saber as suas soluções é trabalhar com as duas primeiras e verificar se a solução obtida também é a mesma da terceira, mas $\begin{cases} a+b=6 \\ 2a+b=13 \end{cases} \Rightarrow a=7, b=-1$, e $4.7-1=27 \neq 45$, ou seja, o sistema é impossível.

No exemplo que acabamos de estudar, o problema não tem solução. Uma interpretação geométrica para esse fato é que os pontos (1, 6), (2, 13) e (4, 45) não estão alinhados, como pode ser facilmente verificado por algum método de Geometria Analítica.

Como nenhuma reta passa pelos três pontos dados, poderíamos escolher dois dos pontos e encontrar a reta que passa por eles, usando-a como função de aproximação. Mas quais dos pontos devem ser escolhidos? Como dizer se uma aproximação é "melhor" que outra sem termos a função? Uma reta que não passa pelos pontos pode ser uma melhor aproximação?

Uma maneira de medir o quanto uma reta $y = ax + b$ se distancia de um conjunto de dados $\{(x_0, y_0), \ldots, (x_n, y_n)\}$ é o cálculo da distância vertical entre (x_i, y_i) e seu correspondente pela reta $(x_i, ax_i + b)$, a saber $|y_i - (ax_i + b)|$, como sugere a figura 1.

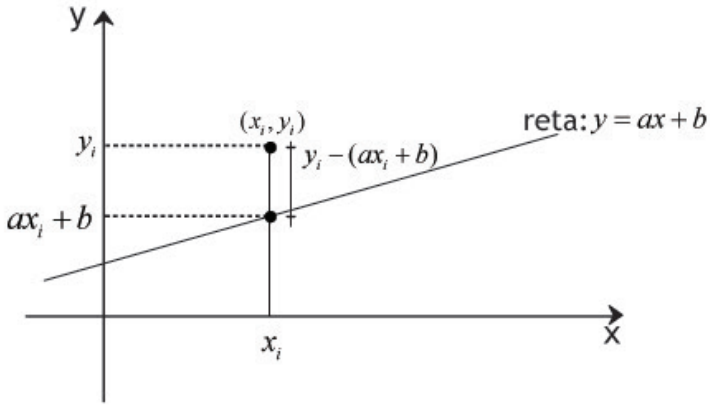


Figura 1: Distância vertical entre (x_i, y_i) e $(x_i, ax_i + b)$

Calcular o módulo diretamente dividiria os casos em que os pontos estão acima ou abaixo da reta. Para simplificar o processo, calculamos diretamente o quadrado desse valor. Definimos, então, o desvio quadrado por:

$$dq_i = \left(y_i - (ax_i + b)\right)^2$$

Exemplo 3

Para o conjunto de dados $\{(1,6),(2,13),(4,45)\}$ e para a reta $y = 8x + 2$, calcule todos os desvios quadrados.

Solução:

Substituindo *x* por 1, 2 e 4 na equação da reta, obtemos 10, 18 e 34, respectivamente. Assim, os desvios quadrados serão:

$dq_0 = (y_0 - (8x_0 + 2))^2 = (6 - 10)^2 = 16$;

$dq_1 = (y_1 - (8x_1 + 2))^2 = (13 - 18)^2 = 25$;

$dq_2 = (y_2 - (8x_2 + 2))^2 = (34 - 45)^2 = 121$.

Exemplo 4

Para o conjunto de dados $\{(1,2),(3,9),(5,16),(7,20)\}$ e para a reta $y = 3x - 1$, calcule todos os desvios quadrados.

Solução:

Substituindo *x* por 1, 3, 5 e 7 na equação da reta, obtemos 2, 8, 14 e 20, respectivamente. Assim, os desvios quadrados serão:

$dq_0 = (y_0 - (3x_0 - 1))^2 = (2 - 2)^2 = 0$

$dq_1 = (y_1 - (3x_1 - 1))^2 = (9 - 8)^2 = 1$

$dq_2 = (y_2 - (3x_2 - 1))^2 = (16 - 14)^2 = 4$

$$dq_3 = (y_3 - (3x_3 - 1))^2 = (20 - 20)^2 = 1$$

Procuraremos, assim, minimizar a soma dos desvios quadrados.

Problema - Para o conjunto de dados $\{(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)\}$, encontrar a reta $y = ax + b$ que minimiza a soma dos desvios quadrados, ou seja, tal que o valor de

$Q = \sum_{i=0}^{n} dq_i = \sum_{i=0}^{n} (y_i - (ax_i + b))^2$ seja o menor possível.

Aqui temos uma justificativa para o nome método dos mínimos quadrados. Observe que o problema consiste em encontrar os valores de a e b que minimizem a expressão Q. Do cálculo de duas variáveis, sabemos que os pontos de mínimo possuem derivadas nulas em relação às variáveis a e b. A derivada de Q em relação a a é representada por $\frac{\partial Q}{\partial a}$ e por ser igual a zero, devemos ter:

$$\begin{aligned}
\frac{\partial Q}{\partial a} = 0 \quad & \Leftrightarrow -2\sum_{i=0}^{n} x_i (y_i - (ax_i + b)) = 0 \Leftrightarrow \\
& \Leftrightarrow \sum_{i=0}^{n} x_i (y_i - ax_i - b) = 0 \Leftrightarrow \\
& \Leftrightarrow \sum_{i=0}^{n} x_i y_i - \sum_{i=0}^{n} ax_i^2 - \sum_{i=0}^{n} bx_i = 0 \Leftrightarrow \\
& \Leftrightarrow \sum_{i=0}^{n} ax_i^2 + \sum_{i=0}^{n} bx_i = \sum_{i=0}^{n} x_i y_i \Leftrightarrow \\
& \Leftrightarrow a\sum_{i=0}^{n} x_i^2 + b\sum_{i=0}^{n} x_i = \sum_{i=0}^{n} x_i y_i.
\end{aligned}$$

Analogamente, a derivada de Q em relação a b é representada por $\frac{\partial Q}{\partial b}$ e para que seja igual a zero, devemos ter:

$$\begin{aligned}
\frac{\partial Q}{\partial b} = 0 \quad & \Leftrightarrow -2\sum_{i=0}^{n} (y_i - (ax_i + b)) = 0 \Leftrightarrow \\
& \Leftrightarrow \sum_{i=0}^{n} (y_i - ax_i - b) = 0 \Leftrightarrow \\
& \Leftrightarrow \sum_{i=0}^{n} y_i - \sum_{i=0}^{n} ax_i - \sum_{i=0}^{n} b = 0 \Leftrightarrow \\
& \Leftrightarrow \sum_{i=0}^{n} ax_i + \sum_{i=0}^{n} b = \sum_{i=0}^{n} y_i \Leftrightarrow \\
& \Leftrightarrow a\sum_{i=0}^{n} x_i + b(n+1) = \sum_{i=1}^{n} y_i.
\end{aligned}$$

Juntando as equações resultantes, as quais chamamos de equações normais do problema, obtemos o sistema nas incógnitas a e b:

$$\begin{cases} a\sum_{i=0}^{n} x_i^2 + b\sum_{i=0}^{n} x_i = \sum_{i=0}^{n} x_i y_i \\ a\sum_{i=0}^{n} x_i + b(n+1) = \sum_{i=0}^{n} y_i \end{cases}.$$

Observe, no exemplo a seguir, como determinar cada um dos elementos envolvidos nas equações normais e como resolver o problema.

Exemplo 5

Usando o método dos mínimos quadrados, encontre a reta que melhor se ajusta ao conjunto de dados $\{(1,6),(2,13),(4,45)\}$.

Solução:

Para o conjunto de dados, temos $x_0 = 1; x_1 = 2; x_2 = 4$ e $y_0 = 6; y_1 = 13; y_2 = 45$. Assim, podemos encontrar:

$$\sum_{i=0}^{2} x_i^2 = x_0^2 + x_1^2 + x_2^2 = 1^2 + 2^2 + 4^2 = 1 + 4 + 16 = 21.$$

$$\sum_{i=0}^{2} x_i = x_0 + x_1 + x_2 = 1 + 2 + 4 = 7.$$

$$\sum_{i=0}^{2} x_i y_i = x_0 y_0 + x_1 y_1 + x_2 y_2 = 1.6 + 2.13 + 4.45 = 6 + 26 + 180 = 212.$$

$$\sum_{i=0}^{2} y_i = y_0 + y_1 + y_2 = 6 + 13 + 45 = 64.$$

Assim, o sistema de equações normais $\begin{cases} a\sum_{i=0}^{n} x_i^2 + b\sum_{i=0}^{n} x_i = \sum_{i=0}^{n} x_i y_i \\ a\sum_{i=0}^{n} x_i + b(n+1) = \sum_{i=0}^{n} y_i \end{cases}$ fica $\begin{cases} 21a + 7b = 212 \\ 7a + 3b = 64 \end{cases}$, que tem solução $a = \frac{94}{7}$ e $b = -10$. Dessa forma, a reta procurada tem equação $y = \frac{94}{7}x - 10$.

Observe que estamos querendo uma reta que minimize os desvios quadrados. No exemplo que acabamos de resolver, a reta não passa por nenhum dos pontos. Ao processo descrito acima, damos também o nome de regressão linear dos dados e os coeficientes procurados podem ser encontrados diretamente em algumas calculadoras científicas. Acompanhe o próximo exemplo do tópico, conferindo as contas feitas.

Exemplo 6

Usando o método dos mínimos quadrados, encontre a equação da reta que melhor se ajusta ao conjunto de dados $\{(1,2),(3,9),(5,16),(7,20)\}$.

Solução:

Temos $x_0 = 1; x_1 = 3; x_2 = 5; x_3 = 7$ e $y_0 = 2; y_1 = 9; y_2 = 16; y_3 = 20$. Daí calculamos:

$$\sum_{i=0}^{3} x_i^2 = 1^2 + 3^2 + 5^2 + 7^2 = 84.$$

$$\sum_{i=0}^{3} x_i = 1 + 3 + 5 + 7 = 16.$$

$$\sum_{i=0}^{3} x_i y_i = 1.2 + 3.9 + 5.16 + 7.20 = 249.$$

$$\sum_{i=0}^{3} y_i = 2 + 9 + 16 + 20 = 47.$$

Assim, o sistema de equações normais $\begin{cases} a\sum_{i=0}^{n} x_i^2 + b\sum_{i=0}^{n} x_i = \sum_{i=0}^{n} x_i y_i \\ a\sum_{i=0}^{n} x_i + b(n+1) = \sum_{i=0}^{n} y_i \end{cases}$ fica $\begin{cases} 84a + 16b = 249 \\ 16a + 4b = 47 \end{cases}$, que tem solução $a = \frac{61}{20}$ e $b = -\frac{9}{20}$. Dessa forma, a reta procurada tem equação $y = \frac{61}{20}x - \frac{9}{20}$.

Antes de encerrar o tópico, acompanhe mais um exemplo, com o qual ganhamos mais um método para aproximar integrais.

Exemplo 7

Usando a função do primeiro grau obtida pelos métodos dos mínimos quadrados, podemos obter um valor aproximado para $\int_1^2 \frac{1}{x} dx$ com quatro subintervalos. Os pontos dessa divisão são $x_0 = 1; x_1 = \frac{5}{4}; x_2 = \frac{3}{2}; x_3 = \frac{7}{4}; x_4 = 2$, com imagens correspondentes pela função $f(x) = \frac{1}{x}$ iguais a $y_0 = 1; y_1 = \frac{4}{5}; y_2 = \frac{2}{3}; y_3 = \frac{4}{7}; y_4 = \frac{1}{2}$. Para esse conjunto de dados, podemos encontrar:

$$\sum_{i=0}^{4} x_i^2 = 1^2 + \left(\frac{5}{4}\right)^2 + \left(\frac{3}{2}\right)^2 + \left(\frac{7}{4}\right)^2 + 2^2 = \frac{95}{8}.$$

$$\sum_{i=0}^{4} x_i = 1 + \frac{5}{4} + \frac{3}{2} + \frac{7}{4} + 2 = \frac{15}{2}.$$

$$\sum_{i=0}^{4} x_i y_i = 1.1 + \frac{5}{4}.\frac{4}{5} + \frac{3}{2}.\frac{2}{3} + \frac{7}{4}.\frac{4}{7} + 2.\frac{1}{2} = 5.$$

$$\sum_{i=0}^{4} y_i = 1+\frac{4}{5}+\frac{2}{3}+\frac{4}{7}+\frac{1}{2}=\frac{743}{210}.$$

Assim, o sistema de equações normais $\begin{cases} a\sum_{i=0}^{n} x_i^2 + b\sum_{i=0}^{n} x_i = \sum_{i=0}^{n} x_i y_i \\ a\sum_{i=0}^{n} x_i + b(n+1) = \sum_{i=0}^{n} y_i \end{cases}$ fica $\begin{cases} \frac{95}{8}a+\frac{15}{2}b=5 \\ \frac{15}{2}a+5b=\frac{743}{210} \end{cases}$, que tem solução $a \cong -0,49143$ e $b \cong 1,44476$. Dessa forma, a parte do gráfico da função $f(x)=\frac{1}{x}$ para valores de $x \in [1,2]$ pode ser aproximada pela reta $y=-0,49143x+1,44476$. Assim,

$$\int_1^2 \frac{1}{x}dx \approx \int_1^2 (ax+b)dx = \left[\frac{ax^2}{2}+bx\right]_{x=1}^{x=2} = (2a+2b)-\left(\frac{a}{2}+b\right) =$$
$$= \frac{3a}{2}+b \cong \frac{3.(-0,49143)}{2}+1,44476 = 0,707615.$$

Com o que temos neste exemplo, aliado ao exposto na aula 7, podemos também aproximar o valor $\ln 2 \cong 0,707615$. Sugerimos que se use o método acima para obter outras aproximações para as integrais discutidas naquela aula.

Por fim, observe que, se escrevermos

$F=\sum_{i=0}^{n} x_i^2; G\sum_{i=0}^{n} x_i; H=\sum_{i=0}^{n} x_i y_i; I=n+1$ e $J=\sum_{i=0}^{n} y_i$, o sistema de equações normais de que tanto falamos reduz-se a $\begin{cases} Fa+Gb=H \\ Ga+Ib=J \end{cases}$, que é matricialmente equivalente a $\begin{bmatrix} F & G \\ G & I \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix}=\begin{bmatrix} H \\ J \end{bmatrix}$. Uma vez que a matriz dos coeficientes desse sistema é simétrica, podemos usar o método de Cholesky para resolvê-lo (ou aproximar a solução).

Neste tópico, tratamos um conjunto de dados isolados (caso discreto) que foi aproximado por uma função do primeiro grau (caso linear). Há várias outras possibilidades também para dados contínuos e outros tipos de funções (exponenciais, logarítmicas, trigonométricas, polinomiais etc). Algumas dessas aproximações serão discutidas nos próximos tópicos, sempre tendo em vista a melhor relação entre aproximação dos dados e complexidade da função de ajuste.

TÓPICO 3

Caso discreto geral

Objetivos

- Formular o método dos mínimos quadrados no caso geral
- Analisar o caso de funções do segundo grau

No tópico anterior, vimos como aproximar um conjunto de dados por uma função do primeiro grau, resolvendo as suas equações normais e obtendo os coeficientes da equação da reta. Em alguns problemas, pode ficar evidente, pela quantidade de pontos e pelo seu comportamento, o uso de outros tipos de funções.

Com mais rigor, dado o conjunto de pontos $\{(x_0, y_0),(x_1, y_1),...,(x_n, y_n)\}$, os desvios da função $\varphi(x)$ são definidos por $d_i = |\varphi(x_i) - y_i|$ e os desvios quadrados por $dq_i = (\varphi(x_i) - y_i)^2$. O método dos mínimos quadrados consiste em encontrar a função, dentro de um modelo pré-estabelecido, que minimize a soma dos desvios quadrados. Para a soma $Q = \sum_{i=0}^{n} (\varphi(x_i) - y_i)^2$, vale sempre $Q \geq 0$, de onde temos que ela deve assumir um mínimo, que é o objetivo do nosso problema. Note que, ao considerar os desvios quadrados, a ordem da subtração não influencia o resultado, ou seja, poderíamos igualmente definir $Q = \sum_{i=0}^{n} (y_i - \varphi(x_i))^2$.

A escolha do tipo da função $\varphi(x)$ depende do fenômeno descrito pelos dados ou da análise gráfica dos pontos. Por exemplo, se a marcação dos pontos sugerir uma parábola, procuraremos uma função do segundo grau, e a determinação dos coeficientes será feita de modo semelhante ao desenvolvido no tópico 1.

Exemplo 1

Marque os pontos do conjunto $\{(-2;14,5),(-1;7,5),(0;4,5),(1;2,5),(2;2),(3;4,5)\}$ no plano cartesiano.

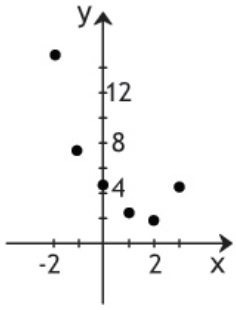


Figura 2: Plano Cartesiano

Solução:

Um esboço da marcação dos pontos pode ser visto na figura 2. Pelo que vimos na aula 6, para um conjunto com seis pontos, o polinômio interpolador terá grau 5, mas o diagrama sugere uma parábola.

Se fizermos o processo para encontrar uma parábola que passa pelos seis pontos dados no exemplo, encontraremos um sistema impossível, mas podemos encontrar uma função do segundo grau cujo gráfico aproxime bem esses pontos, ou seja, que passe o mais perto possível dos pontos dados. Uma parábola tem equação do tipo $\varphi(x) = ax^2 + bx + c$. Para cada ponto (x_i, y_i) do conjunto de dados, podemos definir o desvio quadrado por $dq_i = \left(\varphi(x_i) - y_i\right)^2 = \left(ax_i^2 + bx_i + c - y_i\right)^2$. Dessa forma, a expressão da soma dos desvios quadrados fica:

$$Q = \sum_{i=0}^{n} \left(\varphi(x_i) - y_i\right)^2 = \sum_{i=0}^{n} \left(ax_i^2 + bx_i + c - y_i\right)^2 .$$

Para este problema, devemos encontrar *a*, *b* e *c* que minimizem o valor de Q. Assim como o desenvolvido no caso linear, aqui faremos $\frac{\partial Q}{\partial a} = \frac{\partial Q}{\partial b} = \frac{\partial Q}{\partial c} = 0$, o que irá gerar três equações normais. Acompanhe com atenção os cálculos abaixo, pois eles poderão ser usados para qualquer outro caso no qual o conjunto de dados sugerir uma parábola.

$Q = \sum_{i=0}^{n} \left(ax_i^2 + bx_i + c - y_i\right)^2 \Rightarrow \frac{\partial Q}{\partial a} = \sum_{i=0}^{n} 2x_i^2 \left(ax_i^2 + bx_i + c - y_i\right)$. Daí temos:

$$\frac{\partial Q}{\partial a} = 0 \Leftrightarrow \sum_{i=0}^{n} x_i^2 \left(ax_i^2 + bx_i + c - y_i\right) = 0 \Leftrightarrow$$

$$\Leftrightarrow \sum_{i=0}^{n} ax_i^4 + bx_i^3 + cx_i^2 - x_i^2 y_i = 0 \Leftrightarrow$$

$$\Leftrightarrow \sum_{i=0}^{n} ax_i^4 + \sum_{i=0}^{n} bx_i^3 + \sum_{i=0}^{n} cx_i^2 - \sum_{i=0}^{n} x_i^2 y_i = 0 \Leftrightarrow$$

$$\Leftrightarrow a\sum_{i=0}^{n}x_i^4+b\sum_{i=0}^{n}x_i^3+c\sum_{i=0}^{n}x_i^2=\sum_{i=0}^{n}x_i^2y_i.$$

Agora em relação a b:

$Q=\sum_{i=0}^{n}\left(ax_i^2+bx_i+c-y_i\right)^2\Rightarrow\frac{\partial Q}{\partial b}=\sum_{i=0}^{n}2x_i\left(ax_i^2+bx_i+c-y_i\right)$. Daí temos:

$$\frac{\partial Q}{\partial b}=0\Leftrightarrow\sum_{i=0}^{n}x_i\left(ax_i^2+bx_i+c-y_i\right)=0\Leftrightarrow$$

$$\Leftrightarrow\sum_{i=0}^{n}ax_i^3+bx_i^2+cx_i-x_iy_i=0\Leftrightarrow$$

$$\Leftrightarrow\sum_{i=0}^{n}ax_i^3+\sum_{i=0}^{n}bx_i^2+\sum_{i=0}^{n}cx_i-\sum_{i=0}^{n}x_iy_i=0\Leftrightarrow$$

$$\Leftrightarrow a\sum_{i=0}^{n}x_i^3+b\sum_{i=0}^{n}x_i^2+c\sum_{i=0}^{n}x_i=\sum_{i=0}^{n}x_iy_i.$$

E, por fim, em relação a c:

$Q=\sum_{i=0}^{n}\left(ax_i^2+bx_i+c-y_i\right)^2\Rightarrow\frac{\partial Q}{\partial c}=\sum_{i=0}^{n}2\left(ax_i^2+bx_i+c-y_i\right)$. Então, temos:

$$\frac{\partial Q}{\partial c}=0\Leftrightarrow\sum_{i=0}^{n}\left(ax_i^2+bx_i+c-y_i\right)=0\Leftrightarrow$$

$$\Leftrightarrow\sum_{i=0}^{n}ax_i^2+\sum_{i=0}^{n}bx_i+\sum_{i=0}^{n}c-\sum_{i=0}^{n}y_i=0\Leftrightarrow$$

$$\Leftrightarrow a\sum_{i=0}^{n}x_i^2+b\sum_{i=0}^{n}x_i+c(n+1)=\sum_{i=0}^{n}y_i.$$

Juntando os três resultados, obtemos o sistema de equações normais:

$$\begin{cases}a\sum_{i=0}^{n}x_i^4+b\sum_{i=0}^{n}x_i^3+c\sum_{i=0}^{n}x_i^2=\sum_{i=0}^{n}x_i^2y_i\\ a\sum_{i=0}^{n}x_i^3+b\sum_{i=0}^{n}x_i^2+c\sum_{i=0}^{n}x_i=\sum_{i=0}^{n}x_iy_i\\ a\sum_{i=0}^{n}x_i^2+b\sum_{i=0}^{n}x_i+c(n+1)=\sum_{i=0}^{n}y_i\end{cases}.$$

Para cada conjunto de dados, os valores $\sum_{i=0}^{n}x_i^4,\sum_{i=0}^{n}x_i^3,\sum_{i=0}^{n}x_i^2,\sum_{i=0}^{n}x_i,\sum_{i=0}^{n}x_i^2y_i,\sum_{i=0}^{n}x_iy_i$ e $\sum_{i=0}^{n}y_i$ são facilmente determinados, embora seja um processo demorado de ser realizado manualmente para uma grande quantidade de pontos. Uma vez determinados os valores citados, passa-se a resolver o sistema de equações normais para a determinação dos coeficientes da função $\varphi(x)=ax^2+bx+c$.

Exemplo 2

Usando o método dos mínimos quadrados, encontre a equação da parábola que melhor se ajusta ao conjunto de dados $\{(-2;14,5),(-1;7,5),(0;4,5),(1;2,5),(2;2),(3;4,5)\}$.

Solução:

Para determinar os coeficientes da equação $\varphi(x)=ax^2+bx+c$, devemos resolver o sistema de equações normais e, para tanto, devemos encontrar os valores de:

$$\sum_{i=0}^{5} x_i^4 = x_0^4+x_1^4+x_2^4+x_3^4+x_4^4+x_5^4+x_6^4 =$$
$$=(-2)^4+(-1)^4+0^4+1^4+2^4+3^4=16+1+0+1+16+81=115;$$

$$\sum_{i=0}^{5} x_i^3 = x_0^3+x_1^3+x_2^3+x_3^3+x_4^3+x_5^3+x_6^3 =$$
$$=(-2)^3+(-1)^3+0^3+1^3+2^3+3^3=(-8)+(-1)+0+1+8+27=27;$$

$$\sum_{i=0}^{5} x_i^2 = x_0^2+x_1^2+x_2^2+x_3^2+x_4^2+x_5^2+x_6^2 =$$
$$=(-2)^2+(-1)^2+0^2+1^2+2^2+3^2=4+1+0+1+4+9=19;$$

$$\sum_{i=0}^{5} x_i = x_0+x_1+x_2+x_3+x_4+x_5+x_6 =$$
$$=(-2)+(-1)+0+1+2+3=3;$$

$$\sum_{i=0}^{5} x_i^2 y_i = x_0^2y_0+x_1^2y_1+x_2^2y_2+x_3^2y_3+x_4^2y_4+x_5^2y_5 =$$
$$=(-2)^2.14,5+(-1)^2.7,5+0^2.4,5+1^2.2,5+2^2.2+3^2.4,5=$$
$$=58+7,5+0+2,5+8+40,5=116,5;$$

$$\sum_{i=0}^{5} x_i y_i = x_0y_0+x_1y_1+x_2y_2+x_3y_3+x_4y_4+x_5y_5 =$$
$$=(-2).14,5+(-1).7,5+0.4,5+1.2,5+2.2+3.4,5=$$
$$=-29-7,5+0+2,5+4+13,5=-16,5;$$

$$\sum_{i=0}^{5} y_i = y_0+y_1+y_2+y_3+y_4+y_5 =$$
$$=14,5+7,5+4,5+2,5+2+4,5=35,5;$$

Assim, o sistema de equações normais descrito acima fica $\begin{cases} 115a+27b+19c=116,5 \\ 27a+19b+3c=-16,5 \\ 19a+3b+6c=35,5 \end{cases}$, cuja solução pode ser encontrada (ou aproximada) por algum dos métodos vistos nas aulas 4 e 5 (inclusive o de Cholesky, pois a matriz dos

coeficientes é simétrica). Temos $a \cong 1,0269$, $b \cong -2,9839$ e $c \cong 4,1571$. Assim, a parábola procurada tem equação $y = 1,0269x^2 - 2,9839x + 4,1571$.

O método empregado no exemplo anterior pode ser estendido para encontrar polinômios de qualquer grau cujo gráfico aproxime um conjunto de pontos. Entretanto, o processo ganha complexidade à medida que o grau do polinômio aumenta, como pode ser visto já no caso de aumentar o grau de 1 pra 2. Problemas semelhantes podem ser resolvidos quando os pontos sugerirem uma função trigonométrica, logarítmica ou exponencial. No próximo tópico, estudaremos o método dos mínimos quadrados para dados contínuos, ou seja, para um intervalo em vez de dados isolados.

TÓPICO 3

O caso contínuo

OBJETIVOS

- Descrever o método dos mínimos quadrados para variável contínua
- Analisar expressões obtidas por derivação parcial

Em vez de um conjunto de dados, no caso contínuo do método dos mínimos quadrados, teremos uma função $f:[a,b]\to\mathbb{R}$, a qual aproximaremos por outra $\varphi:[a,b]\to\mathbb{R}$. Como o conjunto base não é mais formado por pontos isolados, não podemos definir o desvio total pela soma dos desvios em cada ponto. Esse problema é contornado pela definição a seguir:

Definição - Dada a função $f:[a,b]\to\mathbb{R}$, o desvio quadrado total de $\varphi:[a,b]\to\mathbb{R}$ em relação a f é dado por $Q=\int_a^b \left(f(x)-\varphi(x)\right)^2 dx$.

O objetivo aqui, então, será minimizar o valor de Q dentro de determinado modelo para $\varphi(x)$. Por exemplo, poderemos aproximar um polinômio de grau elevado por um de grau 2, ou uma função trigonométrica por uma polinomial. A dificuldade nesse caso será o cálculo das integrais, portanto recomendamos uma revisão sobre integrais definidas.

EXEMPLO 1

Encontre a função do primeiro grau que minimiza o desvio quadrado total em relação à função $f(x)=x^3+6$ no intervalo $[0,1]$.

Solução:

Uma função do primeiro grau é do tipo $\varphi(x)=ax+b$. Assim, o desvio quadrado total no intervalo dado é calculado por $Q=\int_0^1 \left((x^3+6)-(ax+b)\right)^2 dx$.

Simplifiquemos, então:

$$
\begin{aligned}
Q &= \int_0^1 \left((x^3+6)-ax+b)\right)^2 dx = \\
&= \int_0^1 \left((x^3+6)^2 - 2(x^3+6)(ax+b)+(ax+b)^2\right)dx = \\
&= \int_0^1 \left(x^6+12x^3+36-2(ax^4+bx^3+6ax+6b)+a^2x^2+2abx+b^2\right)dx = \\
&= \int_0^1 \left(x^6+12x^3+36-2ax^4-2bx^3-12ax-12b+a^2x^2+2abx+b^2\right)dx = \\
&= \left[\frac{x^7}{7}+3x^4+36x-2a\frac{x^5}{5}-b\frac{x^4}{2}-6ax^2-12bx+a^2\frac{x^3}{3}+abx^2+b^2x\right]_{x=0}^{x=1} = \\
&= \frac{1}{7}+3+36-\frac{2a}{5}-\frac{b}{2}-6a-12b+\frac{a^2}{3}+ab+b^2 = \\
&= \frac{274}{7}-\frac{32a}{5}-\frac{25b}{2}+\frac{a^2}{3}+ab+b^2.
\end{aligned}
$$

Com o objetivo de minimizar o valor de $Q=\frac{274}{7}-\frac{32a}{5}-\frac{25b}{2}+\frac{a^2}{3}+ab+b^2$, devemos anular suas derivadas parciais em relação a a e a b. Assim, calculamos:

$\frac{\partial Q}{\partial a}=-\frac{32}{5}+\frac{2a}{3}+b$ e $\frac{\partial Q}{\partial b}=-\frac{25}{2}+a+2b$. Igualando as duas expressões a zero, obtemos as equações $\frac{2a}{3}+b=\frac{32}{5}$ e $a+2b=\frac{25}{2}$. Multiplicando a primeira equação por 15 e a segunda por 2, obtemos o sistema $\begin{cases}10a+15b=96\\ 2a+4b=25\end{cases}$, que tem solução $a=\frac{9}{10}=0,9$ e $b=\frac{29}{5}=5,8$. Assim, a função procurada é a de equação $\varphi(x)=0,9x+5,8$.

Como se percebe, ajustar curvas pelo método dos mínimos quadrados pode ser um processo bem trabalhoso (imagine fazer o exemplo anterior ajustando por uma função só de segundo grau). Além disso, é necessário entender os passos, deve ficar claro que, assim como no caso de interpolação polinomial, estamos encontrando um modelo (ou simplificando um modelo pré-existente) de uma função dada por uma expressão ou conjunto de dados. A diferença central entre os dois métodos é que, na interpolação, a função dada e o ajuste que fazemos coincidem nos pontos; enquanto no método dos mínimos quadrados, como o nome sugere, ajustamos por uma curva que passe o mais perto possível dos pontos dados.

O ajuste pelos mínimos quadrados permite, também, obter aproximações para valores fora do intervalo considerado com certa segurança. Se os dados vierem de experimentos sujeitos a erros de medição, é possível que tenhamos mais de um valor para determinado ponto, de acordo com que escolhamos modelos diferentes para o ajuste. Na prática, algo razoável para contornar essa provável ambiguidade é a média aritmética entre os valores possíveis dentre os modelos aceitáveis.

REFERÊNCIAS

ANTON, Howard e BUSBY, Robert C. **Álgebra linear contemporânea**.Tradução Claus Ivo Doering. Porto Alegre: Bookman, 2006.

ASANO, Claudio Hirofume e COLLI, Eduardo. **Cálculo numérico**: fundamentos e aplicações. 2007. Disponível em: <*www.ime.usp.br/~asano/LivroNumerico/LivroNumerico.pdf*>. Acesso em: 20 jul. 2009.

BERLEZE, Caren Saccol e BISOGNIN, Eleni. Interdisciplinaridade: equações quadráticas associadas à ionização de ácidos. In: ENCONTRO GAÚCHO DE EDUCAÇÃO MATEMÁTICA - EGEM, 9., 2006, Caxias do Sul, RS. Anais do IX EGEM. Caxias do Sul: UCS, 2006. Disponível em: <*www.ccet.ucs.br/eventos/outros/egem/cientificos/cc46.pdf*>. Acesso em: 20 jul. 2009.

BIEMBENGUT, Maria S. e HEIN, Nelson. **Modelagem matemática no ensino**. São Paulo: Contexto, 2000.

BRASIL. Ministério da Educação. Secretaria de Educação Básica. **Orientações curriculares para o ensino médio**: Ciências da Natureza Matemática e suas Tecnologias, v. 2. Brasília: MEC/SEB, 2006.

BUFFONI, Salete Souza de Oliveira. **Apostila de introdução aos métodos numéricos** (Parte 1). Volta Redonda, Rio de Janeiro: Universidade Federal Fluminense, 2002. Disponível em: <*www.professores.uff.br/salete/imn/calnuml.pdf*>. Acesso em: 20 jul. 2009.

CAMPONOGARA, Eduardo ; CASTELAN NETO, Eugênio de Bona. **Cálculo numérico para controle e automação** (Versão preliminar). Florianópolis: Universidade Federal de Santa Catarina / Departamento de Automação e Sistemas, 2008 Disponível em: <*http://www.das.ufsc.br/~camponog/Disciplinas/DAS-5103/LN.pdf*>. Acesso em: 20 jul. 2009.

CARNEIRO, José P. Q. Raiz quadrada utilizando médias. **Revista do Professor de Matemática**, n. 45, p. 21-28, 2001. Disponível em: <*http://www.bibvirt.futuro.usp.br/textos/periodicos/revista_do_professor_de_matematica/vol_0_no_45*>. Acesso em: 20 jul. 2009.

FREITAS, Sérgio Roberto. **Métodos Numéricos**. Campo Grande: Universidade Federal de Mato Grosso do Sul, 2000. Disponível em: <*www.profwillian.com/_diversos/download/livro_metodos.pdf.*>. Acesso em: 20 jul. 2009.

HUMES, Ana F. P. C. L. et. al. **Noções de Cálculo Numérico**. São Paulo: McGraw-Hill do Brasil, 1984.

HUMES, Ana Flora Pereira de Castro Lages. **Noções de Cálculo Numérico**. São Paulo: McGraw-Hill do Brasil, 1984.

LIMA, Elon L. et. al. **A matemática do ensino médio**. Vol. 1. Rio de Janeiro: Sociedade Brasileira de Matemática, 2003.

_____. **Exame de textos**: análise de livros de matemática para o ensino médio. Rio de Janeiro: Sociedade Brasileira de Matemática, 2001.

_____. **Curso de análise**. Vol. 1, 11. ed. Projeto Euclides. Rio de Janeiro: Instituto de Matemática Pura e Aplicada, 2004.

LINHARES, O.D., **Cálculo Numérico B**. Departamento de Ciências de Computação e Estatística do ICMSC, 1969.

LIPSCHUTZ, Seymour. **Álgebra linear**: teoria e problemas. 3. ed. Tradução Alfredo Alves de Farias. São Paulo: Pearson Makron Books, 1994.

OHSE, Marcos L.A matemática como modelo (ferramenta). Pedagobrasil, **Revista Eletrônica de Educação**, v. 1, n. 1, p. 1-2, 2005. Disponível em: *<http://www.pedagobrasil.com.br/pedagogia/amatematica.htm>.* Acesso em: 20 jul. 2009.

RUGGIERO, Marcia Aparecida Gomes e LOPES, Vera Lúcia da Rocha. **Cálculo numérico**: aspectos teóricos e computacionais. 2. ed. São Paulo: Makron Books, 1996.

STEWART, James. **Cálculo**. Vol 1. 5. ed. São Paulo: Cengage Learning.

CURRÍCULO

FRANCISCO GÊVANE MUNIZ CUNHA

Francisco Gêvane Muniz Cunha é professor efetivo do Instituto Federal do Ceará – IFCE desde 1993. Nascido em São João do Jaguaribe – CE em 1970, é técnico em informática industrial pela Escola Técnica Federal do Ceará (1993). Licenciado (1993) e bacharel (1994) em matemática pela Universidade Federal do Ceará – UFC. Possui mestrado em matemática (1997) e mestrado em ciência da computação (2002), ambos pela UFC. É doutor em engenharia de sistemas e computação (2007) pela Universidade Federal do Rio de Janeiro com tese na linha de otimização. Tem experiência na área de matemática aplicada, no ensino de matemática, na formação de professores, no uso de tecnologias e no ensino na modalidade a distância. Atualmente é professor de disciplinas de matemática dos cursos de licenciatura em matemática, engenharias e outros do IFCE. Na modalidade semi-presencial é professor conteudista e formador de disciplinas de matemática do curso licenciatura em matemática do IFCE, tendo produzido diversos livros didáticos. Orienta alunos em nível de graduação e pós-graduação em matemática, ensino de Matemática ou educação Matemática. Tem interesse no uso de ambientes informatizados e, em especial, no uso de softwares educativos como apoio para o ensino de matemática. Dentre outras atividades, gosta de ler a bíblia, ajudar as pessoas, ensinar, estudar matemática e computação e assistir corridas de fórmula 1.

JÂNIO KLÉO SOUSA CASTRO

Jânio Kléo começou seus estudos de Matemática em 2000, quando ingressou no bacharelado da Universidade Federal do Ceará, colando grau em julho de 2004. A partir de 2001 e por três anos, foi monitor de Cálculo Diferencial e Integral na

UFC, desempenhando atividade de acompanhamento e tira-dúvidas para alunos de graduação. Durante os anos de 2006, 2007 e 2008, foi professor da UFC, com turmas de diversos cursos, ministrando aulas de Álgebra Linear, Equações Diferenciais, Variáveis Complexas e Geometria Hiperbólica, entre outras. Desde o começo de 2009 é professor do Instituto Federal de Educação, Ciência e Tecnologia do Ceará, atuando nos campus de Fortaleza e Maracanaú, nos cursos presenciais e semipresenciais.