
# CÁLCULO NUMÉRICO

Elaborado por
**C. Bombardelli**
cb_kxt@hotmail.com - fone: 41 30228546

Junho / 2002

REFERÊNCIA BIBLIOGRAFICA:

Albrecht, P., **Análise Numérica. Um Curso Moderno**, LTC, Rio de Janeiro, 1973;
Barroso, L. C. et al., **Cálculo Numérico com Aplicações**, 2 ed. Ed. HARBRA, São Paulo, 1987;
Brassard, G., Bratley, P., **Algoritmics. Theory & Practice**, Prentice-Hall, N. Jersey, 1988;
Cutlip, M. B., Shacham, M., **Problem Solving in Chemical Engineering with Numerical Methods**, Prentice-Hall, Upper Saddle River, 1999;
Dorn, W. D. et al., **Cálculo Numérico com Estudos de Casos em FORTRAN IV**, Ed Univ. de São Paulo, 1981.
Jenson, V. G., Jeffreys, G. V., **Mathematical Methods in Chemical Engineering**, 2ª ed., Academic Press, London, 1994;
Royo dos Santos, J. A., **Mini-Calculadoras Eletrônicas**, Ed. Edgard Blücher, 1977;
Ruggiero, M. A. G.; Lopes, V. L. R., **Cálculo Numérico: Aspectos Teóricos e Computacionais**, 2 ed. Ed Makron Books do Brasil, 1997.

Se o que faz lhe parecer difícil, é provável que não esteja usando o método adequado.

## Números aproximados

No estudo da matemática, admite-se usualmente que os números que intervêm nas operações são perfeitamente exatos, ou seja, quando possuem infinitas casas se forem escritos nos sistemas de base, como por exemplo, o decimal usual. Assim, mesmo que não esteja expressamente indicado, entende-se que:

2 = 2,00000000000000000000.....
pi = 3,14159265358979238462643.....
raiz quadrada de dois = 1,41421356237309504880168724209698078569 6.....
1/3 = 0,33333333333333333333.....

Números exatos são os que podem ser perfeitamente representados por um símbolo.
Exemplos: são números exatos 1
7
275
0,027
1/2
3/7

As frações, entretanto, podem não ter representação exata no sistema decimal ou em qualquer outro sistema. Por exemplo, não é possível escrever 2/3 exatamente no sistema decimal. O mesmo seria escrito 0,3333333333333..., consistindo numa dízima infinita e somente se pode escrever um valor aproximado como....0,33333, mas o valor exato é impossível. O mesmo ocorre com os números irracionais, PI, *e*, raiz quadrada de 3, etc.

***Um número é dito aproximado quando ele representa um valor que não é exatamente o seu.***

Admite-se também que todos os cálculos são feitos usando as infinitas casas de todos os números. Ora, enquanto isto é possível em operações com números inteiros, tais como:

```
      2,000000000000000000.....
+     2,000000000000000000.....
      ------------------------------------
      4,000000000000000000.....
```

Claramente não é possível na simples operação **2 x pi**

Os erros resultantes do fato de operações serem efetuadas na realidade com representações aproximadas dos números, são usualmente desprezados para não complicar a análise. Estes erros, entretanto, devem ser levados em consideração em muitos casos e seu efeito é estudado na parte da matemática denominada: ***Cálculo Numérico***. Esta disciplina estuda também vários métodos de solução de problemas que não resultam em soluções perfeitamente exatas, mas em *soluções tão exatas quanto seja necessário*. Tais métodos, além de permitirem a solução de muitos problemas não solúveis por outros processos, são freqüentemente mais vantajosos na solução de muitos outros problemas que possuem outros métodos de solução. Quase todos os métodos que serão propostos para a solução de problemas de engenharia podem ser convertidos em programas para serem processados em computador ou serem simplesmente resolvidos manualmente com a ajuda de calculadora.

Os números aproximados surgem nos cálculos por muitas razões, das quais as principais talvez sejam:

- Não é possível efetuar cálculos usando as infinitas casas de um número (você já usou pi com todas as casas indicadas anteriormente?). Assim, é necessário interromper o número após um número conveniente de casas;

- O número é obtido através de medida. Medidas nunca são executadas dando resultados perfeitamente exatos. (Será que um quilo de batatas tem exatamente 1.000,00000000000000000..... gramas?);

- O número é obtido estatisticamente (medias, etc.). Como exemplo, o caso de um Boeing 747 que normalmente pode levar cerca de 400 pessoas, mas se o clube dos gastrônomos, cujos sócios tem um peso médio de 150 kg, fretar um, ele nem sairá do chão;

*Ao se levar em consideração os erros nos números, ou seja, ao se entrar nos domínios do cálculo numérico, desaparecem várias propriedades comuns na matemática dos números exatos.*

Por exemplo:
A + B + C pode ser diferente de C + B + A;
A – B pode ser uma operação impossível, dependendo dos valores de A e de B;
O sinal = passa a significar "iguais dentro de uma determinada precisão".

Mais ainda, no cálculo numérico desaparecem os processos infinitos, tais como limites, derivadas, integrais, etc., que são substituídos por processos finitos aproximados. Os teoremas de existência usualmente não são suficientes no cálculo numérico, ou seja, não interessa apenas saber que uma solução existe, mas sim, saber se é possível obter seu valor (aproximado). Por outro lado, aparecem métodos de obtenção de soluções sem justificativas perfeitamente fundamentais (sem deduções rigorosas), mas baseados em simples plausibilidade ou em raciocínios heurísticos.

***A melhor prova de que um método é certo é o fato de que ele funciona.***

EXATIDÃO E PRECISÃO
Dois conceitos muito confundidos, ambos ligados a números aproximados. A *precisão* está relacionada com o número de casas com que o número é conhecido; a e*xatidão* está relacionada com quão bem ele representa uma grandeza.

Exemplo1:
Na representação de PI, 3,1416 é mais preciso que 3,14, pois tem maior número de casas decimais. 3,1416 é também mais exato que 3,14, pois se aproxima melhor do valor real de PI.

Exemplo 2:
Na representação de um comprimento cujo valor real é 4,273500000..., o número 4,277452 é mais preciso que 4,27 mas é menos exato, pois se aproxima menos do valor real.

ERRO ABSOLUTO, ERRO RELATIVO E ERRO PERCENTUAL.
A diferença numérica entre uma grandeza e o número usado para representá-lo é chamada *erro absoluto* (normalmente em valor absoluto)
Denomina-se *erro relativo* ao erro absoluto dividido pelo valor exato. *Erro percentual* é o erro relativo multiplicado por 100. Se Q for o valor real e $Q_1$ o número aproximado, fica:

Erro absoluto = $| Q_1 – Q |$

Erro relativo = $(Q_1 – Q) / Q | = |Q/Q|$

Erro percentual = $100 . | (Q_1 – Q) / Q |$

Usa-se um número aproximado ou porque não é possível escrever o número exato em forma decimal ou porque não se conhece o valor exato, o que é mais comum. Então, o valor $Q_1 – Q$ usualmente não pode ser calculado, pois Q não pode ser conhecido. Usa-se, normalmente, ao invés do erro absoluto, uma cota superior do erro absoluto, tal que:

$$Q \text{ maior ou igual a } | Q_1 – Q |$$

Neste caso, o erro relativo também pode ser escrito como um limite superior

$$dQ = | Q/ Q| \text{ maior ou igual a } Q/ (|Q_1| - Q)$$

Ou se $Q <<<$ ( muito menor que ) $Q_1$

$$\mathbf{dQ = | Q/ Q| \text{ menor ou igual a } Q/ |Q_1|}$$

ALGARISMOS SIGNIFICATIVOS

Algarismos significativos de um número decimal são todos os seus algarismos de 1 a 9. O 0 (zero) é significativo se não estiver substituindo algarismos desconhecidos ou somente fixando a posição da vírgula.

Exemplo: 0,0158 - são significativos: 1, 5 e 8
4096 - são significativos: 4, 0, 9 e 6
38500 - são significativos: ?

NOTAÇÃO CIENTÍFICA

A notação científica tem por finalidade eliminar a ambigüidade dos zeros do último exemplo. Consiste em escrever 0,..., seguido pelos algarismos significativos multiplicados pela base do sistema numérico elevado a uma potência conveniente, como mostram os exemplos abaixo:

0,0158 = $0{,}158 \times 10^{-1}$
4096 = $0{,}4096 \times 10^{4}$
38500 = $0{,}385 \times 10^{5}$
= $0{,}3850 \times 10^{5}$
= $0{,}38500 \times 10^{5}$ **

** O valor a ser escolhido dependerá de quantos zeros são significativos

ARREDONDAMENTO DE NÚMEROS

É necessário na representação de um valor com muitas (ou infinitas) casas decimais por um número com poucos algarismos.

Exemplo: 4/7 = 0,57014285...

Com três algarismos significativos:
= 0,570

Com cinco algarismos significativos:
= 0,57014

etc.

A eliminação das casas em excesso pode ser feita de três modos distintos:

- por cancelamento: as casas decimais são simplesmente abandonadas. Este tipo de eliminação recebe às vezes o nome de truncamento;
- Arredondamento: O último algarismo conservado é aumentado de uma unidade quando o algarismo eliminado for maior ou igual a cinco.
  Este é o método mais utilizado;
- Arredondamento especial: Feito da mesma forma que a anterior, mas se os algarismos forem 500..., o último conservado será aumentado de uma unidade somente se for ímpar;

Exemplos:

Eliminar as casas a partir da segunda após a vírgula:

| Valor | cancelamento | arredondamento | arredondamento especial |
|---|---|---|---|
| 46,12416 | 46,12 | 46,12 | 46,12 |
| 31,34792 | 31,34 | 34,35 | 31,35 |
| 52,27500 | 52,27 | 52,28 | 52,28 |
| 2,38500 | 2,38 | 2,39 | 2,38 |
| 2,38501 | 2.38 | 2,39 | 2,39 |

Representar os valores seguintes com quatro algarismos significativos:

| Valor | cancelamento | arredondamento | arredondamento especial |
|---|---|---|---|
| 316,8972 | 316,8 | 316,9 | 316,9 |
| 4,168500 | 4,168 | 4,169 | 4,168 |
| 0,0011118 | 0,001111 | 0,001112 | 0,001112 |
| 19,265001 | 19,26 | 19,27 | 19,27 |

ALGARISMOS SIGNIFICATIVOS CORRETOS

São os algarismos de um número que tem significado na representação do valor que o número pretende representar. Como exemplo: Para representar o valor 2/3, o número 0,66667 tem menos algarismos que o número 0,66699842693, mas este último tem menos algarismos que contribuem para a representação de 2/3. Portanto, um algarismo significativo é correto se o arredondamento do número até a sua posição resulta em um erro absoluto menor do que meia unidade na sua posição, ou seja, menor que 5 na primeira casa abandonada.

PROPAGAÇÃO DE ERROS NA ARITMÉTICA DOS NÚMEROS APROXIMADOS

Quando se executam operações aritméticas com números aproximados, os erros originais dos dados se propagam pelos resultados intermediários podendo inclusive eliminar todo o significado do resultado final. É portanto, importante conhecer como os erros se propagam e controlar a exatidão do resultado final.

ADIÇÃO: O erro absoluto do resultado é a soma dos erros absolutos das parcelas. O erro relativo é intermediário entre o das parcelas. Como cota superior pode-se tomar o maior deles. Cuidado: A soma de números aproximados não é comutativa. Isto significa dizer que (a + b) é diferente de (b + a).

SUBTRAÇÃO:

É a operação proibida no cálculo numérico. O erro no resultado de uma subtração pode ser muito maior do que qualquer de suas parcelas.

*Se não for possível evitar subtração em uma solução, devem ser tomadas providências para assegurar que o cálculo não perderá todo o seu significado.*

MULTIPLICAÇÃO E DIVISÃO:

O erro relativo do produto é aproximadamente igual à soma dos erros relativos das parcelas.

## Matrizes

Introdução:
A notação e os métodos matriciais têm sido cada vez mais usados em problemas de matemática aplicada à engenharia e à ciência e constituem um assunto muito ligado a sistemas de equações lineares.

Definição e notação:
Matriz é um arranjo de elementos afins. é denotada por dois colchetes ou parêntesis grandes.
É comum denotar os elementos de uma matriz por letras minúsculas com duplo índice e representar a matriz pela letra maiúscula correspondente, ou pela minúscula com duplo índice literal.

Exemplo:
Elementos de matriz A: $a_{ij}$

Tipo de matrizes:
Os tipos mais usuais de matrizes e sua nomenclatura são:

- Matriz retangular – quando a quantidade de linhas difere da quantidade de colunas;
- Matriz quadrada – quando a quantidade de linhas é igual à quantidade de colunas;
- Matriz fila – quando a matriz possui uma única linha (ou fila);
- Matriz coluna – quando a matriz possui uma única coluna;
- Matriz triangular – quando uma matriz retangular possuir o caso específico de ter todos os seus elementos abaixo da diagonal principal nulos, esta levará a designação de matriz triangular superior. Se os elementos nulos estiverem acima da diagonal principal, tem-se então uma matriz triangular inferior;
- Matriz diagonal – quando a matriz só possuir elementos não nulos na diagonal principal;
- Matriz simétrica – quando os valores são simétricos com relação a diagonal, ou seja, $a_{ij}=a_{ji}$;
- Matriz zero – quando todos elementos da matriz são nulos ou iguais a zero;
- Matriz unidade – também chamada de matriz identidade. É a matriz cujos elementos da diagonal são todos iguais à UNIDADE (iguais a 1);
- Matriz transposta – Diz que a matriz B é transposta de A quando $b_{ij} = a_{ji}$ e neste caso a matriz é denotada $A^t$.

Operações sobre matrizes:

Igualdade: duas matrizes A e B, ambas de dimensão mxn, são ditas iguais somente se $a_{ij} = b_{ij}$ para todo ***i*** e ***j***.

Adição: a matriz C é dita a soma das matrizes A e B, denotada C = A + B quando $c_{ij} = a_{ij} + b_{ij}$.
Vê-se que duas matrizes devem ser de mesma dimensão para poderem ser somadas.

Multiplicação por um número: A matriz B é dita o produto da matriz A pelo número k quando $b_{ij} = k \cdot a_{ij}$ para todo ***i*** e ***j***.

Subtração: a subtração consiste em somar uma matriz com outra multiplicada por –1.

Produto de duas matrizes: para que duas matrizes sejam multiplicadas é necessário que o número de colunas da primeira seja igual ao número de filas da segunda. Tais matrizes são ditas conformáveis. O produto de duas matrizes A e B é outra matriz, C, formada da seguinte maneira: o elemento na i-ésima fila e j-ésima coluna são a soma dos produtos dos elementos da i-ésima fila de A pelos elementos de mesma posição da j-ésima coluna de B. A multiplicação matricial não é comutativa, ou seja, AxB é diferente de BxA. *(É comum ocorrer inclusive que as matrizes A e B sejam conformáveis neste sentido AxB e não o seja no sentido contrário. Mesmo que duas matrizes sejam conformáveis nos dois sentidos, seus produtos em sentidos diferentes normalmente são diferentes).*

Propriedades (resumo)

- Adição:

$A + B = B + A$
$A + (B+C) = (A+B) + C$
$A + 0 = A$ (0 = matriz zero)

- Produto de número por matriz:

$p . (q . A) = (p . q) . A$
$p . (A+B) = p . A + p . B$
$(p+q) . A = p . A + q . A$
$1 . A = A$ (1 = matriz unidade)

- Produto de matrizes:

A . B não é necessariamente igual a B . A
$(A . B) . C = A . (B . C)$
$(A + B) . C = A . C + B . C$
$A . (B + C) = A . B + A . C$
$(k . A) . B = A . (k . B) = k . (A . B)$

- Relações entre as matrizes transpostas e as operações sobre matrizes:

$(A + B)^t = A^t + B^t$
$(p . A)^t = p(A^t)$
$(A . B)^t = B^t . A^t$

Determinante de uma matriz quadrada:
O determinante formado com os mesmos elementos de uma matriz A é denominado determinante da matriz A e é denotado por

$$\det A \text{ ou } |A|$$

Um método de cálculo do valor de um determinante consiste no seu desenvolvimento segundo uma fila ou uma coluna. Dado um determinante de ordem m, denomina-se ***determinante menor*** correspondente ao elemento $a_{ij}$ ao determinante formado eliminando-se a fila i e a coluna j do determinante original. Ao determinante menor precedido por $(-1)^{(i+j)}$, sendo i e j as ordens da fila e da coluna correspondente, denomina-se **determinante cofator** correspondente ao elemento.

***Desenvolver um determinante segundo uma fila (ou uma coluna), consiste em multiplicar cada elemento da fila (ou coluna) pelo seu determinante cofator e somar tudo.***

Matriz inversa:
Chama-se matriz inversa de uma matriz quadrada A. a matriz denotada $A^{-1}$ tal que $A^{-1} . A = A . A^{-1} = I$.
I representa no caso, a matriz identidade. Neste caso especial o produto é igual nos dois sentidos. Um método de obter a matriz inversa $A^{-1}$, de uma dada matriz A, consta de:

1. Substituir todos os elementos da matriz pelos valores dos determinantes cofatores correspondentes;
2. Transpor a matriz resultante;
3. Multiplicar pelo inverso do valor do determinante da matriz original;

*Observação: Se o det A = 0, então a matriz inversa não é definida. (Não existe).*

Matriz adjunta:
À matriz obtida pela multiplicação (det A) . $A^{-1}$ denomina-se matriz adjunta da matriz A, denotando-se por adj A.

$$AdjA = A_{cof}^{T} = \det A \times A^{-1}$$

**Solução de sistema de equações lineares por inversão de matriz:**

Em engenharia é muito comum ao equacionar um problema se chegar a equações do tipo:

$ax + by + cz = r$

as quais podem ser escritas também na forma:

$ax_1 + bx_2 + cx_3 + dx_4 + ex_5 + .... + mx_n = r$

Tais equações são composições de segmentos de retas, onde cada x representa uma incógnita a ser determinada. Tais equações têm solução possível somente quando se consegue montar um sistema onde o número de equações independentes seja igual ao número de incógnitas.

$$
\begin{array}{ccccccccccc}
a_{11}\,x_1 & + & a_{12}\,x_2 & + & a_{13}\,x_3 & + & ..... & + & a_{1n}\,x_n & = & b_1 \\
a_{21}\,x_1 & + & a_{22}\,x_2 & + & a_{23}\,x_3 & + & ..... & + & a_{2n}\,x_n & = & b_2 \\
a_{31}\,x_1 & + & a_{32}\,x_2 & + & a_{33}\,x_3 & + & ..... & + & a_{3n}\,x_n & = & b_3 \\
..... & ... & ..... & ... & ..... & ... & ..... & ... & ..... & ... & .... \\
..... & ... & ..... & ... & ..... & ... & ..... & ... & ..... & ... & .... \\
..... & ... & ..... & ... & ..... & ... & ..... & ... & ..... & ... & .... \\
 & & & & & & & & & & \\
a_{n1}\,x_1 & + & a_{n2}\,x_2 & + & a_{n3}\,x_3 & + & ..... & + & a_{nn}\,x_n & = & b_n
\end{array}
$$

Uma vez montado o sistema, este pode ser transformado em matriz usando-se os coeficientes de cada equação, e tais sistemas, contendo n equações e n incógnitas possuem solução única se o determinante dos coeficientes $a_{ij}$ for **não nulo**, ou seja, deve ser diferente de zero. A unicidade da solução exige que todos os valores envolvidos nas equações sejam perfeitamente exatos e que todos os cálculos para a obtenção da solução sejam feitos usando infinitas casas. Nos casos práticos, a solução é determinada aproximadamente dentro da faixa de precisão desejada.

Um sistema de equações lineares pode ser escrito em notação matricial como no exemplo abaixo:

$$
\begin{bmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & a_{nn} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ ... \\ b_n \end{bmatrix}
$$

Chamando de A, a matriz de coeficientes e de X e B os vetores de incógnitas e segundo membros, respectivamente, fica:

$$A \cdot X = B$$

Se a inversa $A^{-1}$ for calculada, pode-se pré-multiplicar ambos os lados da equação por ela, ficando:

$$A^{-1} \cdot A \cdot X = A^{-1} \cdot B$$

Mas como:

$$A^{-1} \cdot A = I \qquad \text{e} \qquad I \cdot X = X$$

Fica:

$$X = A^{-1} \cdot B$$

E o sistema fica resolvido.

Exemplo 1:
Calcular o determinante e a matriz inversa da matriz abaixo.

- Pelo método do cofator

$$A=\begin{bmatrix} 3 & 1 & 0 \\ -2 & 0 & 2 \\ 1 & 3 & 4 \end{bmatrix} \qquad \det A=\begin{vmatrix} 3 & 1 & 0 \\ -2 & 0 & 2 \\ 1 & 3 & 4 \end{vmatrix}$$

Determinante menor correspondente ao elemento $a_{11}$

$$\det {}_mA=\begin{vmatrix} 0 & 2 \\ 3 & 4 \end{vmatrix}=0\times 4-2\times 3=-6$$

Determinante menor correspondente ao elemento $a_{12}$

$$\det {}_mA=\begin{vmatrix} -2 & 2 \\ 1 & 4 \end{vmatrix}=-2\times 4-2\times 1=-10$$

Determinante menor correspondente ao elemento $a_{13}$

$$\det {}_mA=\begin{vmatrix} -2 & 0 \\ 1 & 3 \end{vmatrix}=-2\times 3-0\times 1=-6$$

Determinante cofator correspondente ao elemento $a_{11}$

$$\det {}_{cf}A=(-1)^{(1+1)}\begin{vmatrix} 0 & 2 \\ 3 & 4 \end{vmatrix}=2\times 3-0\times 1=-6$$

Determinante cofator correspondente ao elemento $a_{12}$

$$\det {}_{cf}A=(-1)^{(1+2)}\begin{vmatrix} -2 & 2 \\ 1 & 4 \end{vmatrix}=(-1)(-2\times 4-2\times 1)=10$$

Determinante cofator correspondente ao elemento $a_{13}$

$$\det {}_{cf}A=(-1)^{(1+3)}\begin{vmatrix} -2 & 0 \\ 1 & 3 \end{vmatrix}=-2\times 3-0\times 1=-6$$

Cálculo do valor do determinante de A, pelo método do determinante cofator:

$$\det A=3\times\begin{vmatrix} 0 & 2 \\ 3 & 4 \end{vmatrix}+(-1)\times 1\begin{vmatrix} -2 & 2 \\ 1 & 4 \end{vmatrix}+0\times\begin{vmatrix} -2 & 0 \\ 1 & 3 \end{vmatrix}=$$

$$\det A=3\times(-6)+(-1)\times 1\times(-10)+0\times(-6)=-18+10=-8$$

*No cálculo de determinante pelo método do determinante cofator, é sempre recomendável eliminar a linha ou coluna que tenha o maior número de elementos nulos.*

- Cálculo do valor do determinante de A, pelo método de chió:

$$\det A = \frac{\begin{vmatrix} 3\times 0-1\times(-2) & 3\times 2-0\times(-2) \\ 3\times 3-1\times 1 & 3\times 4-0\times 1 \end{vmatrix}}{3} = \frac{\begin{vmatrix} 2 & 6 \\ 8 & 12 \end{vmatrix}}{3} = \frac{2\times 12-6\times 8}{3} = \frac{-24}{3} = -8$$

*Para evitar a divisão do determinante pelo elemento pivô, deve-se escolher a linha e a coluna que tenha como elemento comum (pivô) igual a 1.*

- Cálculo do valor do determinante de A, pelo método de Sarrus

$$\det A = 3\times 0\times 4+1\times 2\times 1+0\times(-2)\times 3-0\times 0\times 1-3\times 3\times 2-1\times(-2)\times 4 =$$
$$\det A = 0+2+0-0-18+8 = -8$$

Exemplo 2:
Cálculo da matriz inversa pelo método da matriz adjunta:

1° passo: substituir todos os elementos da matriz pelos valores dos determinantes cofatores correspondentes

$$\begin{bmatrix} +\begin{vmatrix} 0 & 2 \\ 3 & 4 \end{vmatrix} & -\begin{vmatrix} -2 & 2 \\ 1 & 4 \end{vmatrix} & +\begin{vmatrix} -2 & 0 \\ 1 & 3 \end{vmatrix} \\ -\begin{vmatrix} 1 & 0 \\ 3 & 4 \end{vmatrix} & +\begin{vmatrix} 3 & 0 \\ 1 & 4 \end{vmatrix} & -\begin{vmatrix} 3 & 1 \\ 1 & 3 \end{vmatrix} \\ +\begin{vmatrix} 1 & 0 \\ 0 & 2 \end{vmatrix} & -\begin{vmatrix} 3 & 0 \\ -2 & 2 \end{vmatrix} & +\begin{vmatrix} 3 & 1 \\ -2 & 0 \end{vmatrix} \end{bmatrix} = \begin{bmatrix} -6 & +10 & -6 \\ -4 & +12 & -8 \\ +2 & -6 & +2 \end{bmatrix}$$

2° passo: transpor a matriz obtida. Ao transpor a matriz dos cofatores teremos a matriz adjunta de A

$$\begin{bmatrix} -6 & -4 & +2 \\ +10 & +12 & -6 \\ -6 & -8 & +2 \end{bmatrix}$$

3° passo: multiplicar a matriz transposta pelo recíproco do determinante de A. O valor do determinante foi previamente calculado e é igual a –8, portanto:

$$A^{-1} = -\frac{1}{8}\begin{bmatrix} -6 & -4 & +2 \\ +10 & +12 & -6 \\ -6 & -8 & +2 \end{bmatrix} = \begin{bmatrix} +0{,}75 & +0{,}50 & -0{,}25 \\ -1{,}25 & -1{,}50 & +0{,}75 \\ +0{,}75 & +1{,}00 & -0{,}25 \end{bmatrix}$$

Verificação:

$$\begin{bmatrix} +0{,}75 & +0{,}50 & -0{,}25 \\ -1{,}25 & -1{,}50 & +0{,}75 \\ +0{,}75 & +1{,}00 & -0{,}25 \end{bmatrix} \times \begin{bmatrix} 3 & 1 & 0 \\ -2 & 0 & 2 \\ 1 & 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Exemplo 3:
Resolver o sistema de equações abaixo:

Seja o sistema:

$$\begin{array}{rcrcrcl} 3x_1 & + & x_2 & + & & = & -2 \\ -2x_1 & & & + & 2x_3 & = & 1 \\ x1 & + & 3x_2 & + & 4x_3 & = & 0 \end{array}$$

Montando as matrizes correspondentes fica:

$$A = \begin{bmatrix} 3 & 1 & 0 \\ -2 & 0 & 2 \\ 1 & 3 & 4 \end{bmatrix} \qquad X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad B = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

A matriz inversa foi calculada no exercício anterior, portanto:

$$X = A^{-1} \cdot B$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0{,}75 & 0{,}5 & -0{,}25 \\ -1{,}25 & -1{,}5 & 0{,}75 \\ 0{,}75 & 1 & 0{,}25 \end{bmatrix} \times \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ -0{,}5 \end{bmatrix}$$

$$x_1 = -1$$
$$x_2 = 1$$
$$x_3 = -0{,}5$$

Verificação:

$$3 \times (-1) + 1 \times 1 = -3 + 1 = -2$$
$$-2 \times (-1) + 2 \times (-0{,}5) = 2 - 1 = 1$$
$$1 \times (-1) + 3 \times 1 + 4 \times (-0{,}5) = -1 + 3 - 2 = 0$$

*Aplicando os valores x obtidos respectivamente, nas equações acima, verifica-se a igualdade em todas equações, o que confirma a validade do cálculo.*

Exemplo 4:
Calcular o valor do determinante de matriz de 4ª. ordem:

$$\begin{vmatrix} 2 & 0 & -2 & 3 \\ -1 & 3 & -1 & -2 \\ 0 & 0 & 4 & 0 \\ 4 & 5 & 0 & 4 \end{vmatrix} = 4 \times \begin{vmatrix} 2 & 0 & 3 \\ -1 & 3 & -2 \\ 4 & 5 & 4 \end{vmatrix} = 4 \times \left( 3 \times \begin{vmatrix} 2 & 3 \\ 4 & 4 \end{vmatrix} - 5 \times \begin{vmatrix} 2 & 3 \\ -1 & -2 \end{vmatrix} \right) =$$

$$4 \times (3 \times (-4) - 5 \times (-1) = 4 \times (-12 + 5) = -28$$

*Observações: No cálculo de determinante pelo método do cofator, deve-se aproveitar sempre a linha ou coluna que tiver a maior quantidade de zeros.*

**Métodos para cálculo da matriz inversa:**

- Método da adjunta; (visto anteriormente)
- Método de Shipley-Coleman;

O método de Shipley-Coleman usa os seguintes passos:

1. O processo inicia pela escolha do pivô, o qual deverá ser sempre diferente de zero e deve iniciar por $a_{11}$. O elemento pivô é modificado pela transformação

   $$a'_{kk} = -\frac{1}{a_{kk}}$$

2. Os elementos pertencentes à coluna de referência sofrem a transformação

   $$a'_{jk} = -\frac{a_{jk}}{a_{kk}} \qquad \text{para todo } j \neq k \qquad (a'_{jk} = a_{jk} \times a'_{kk})$$

3. Os elementos da linha de referência sofrem a transformação idêntica

   $$a'_{ki} = -\frac{a_{ki}}{a_{kk}} \qquad \text{para todo } i \neq k \qquad (a'_{ki} = a_{ki} \times a'_{kk})$$

4. Os demais elementos sofrem a transformação

   $$a'_{ij} = a_{ij} - a_{ki} \times a'_{jk} \qquad \text{para todo } j \neq k \text{ e } i \neq k$$

5. Repete-se o processo de 1 a 4, para todos os elementos da diagonal principal, um de cada vez;
6. Multiplicar a matriz resultante por –1.

**Métodos para cálculo de determinantes:**

- Método de Chió;
- Método do cofator;
- Método de Sarrus para determinante de ordem 3;
- Método por condensação de pivôs, usando a expansão de Laplace. O valor do determinante é obtido pelo produto dos elementos da diagonal principal;

**Propriedades dos determinantes:**

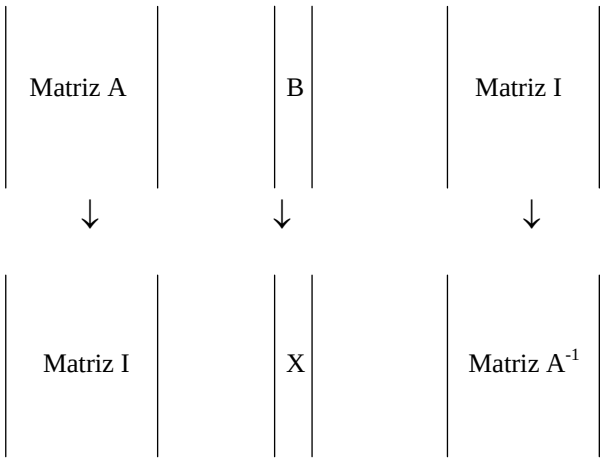
- Só é definido valor para determinantes em matrizes quadradas;
- Se todos os elementos de uma linha, ou uma coluna de uma matriz for igual a zero, o determinante será ***zero***;
- Trocando-se duas linhas ou duas colunas quaisquer, troca-se o sinal do determinante;
- Se todos os elementos de uma linha ou de uma coluna forem multiplicados por um valor constante, então o determinante fica multiplicado por esse valor;
- Se os elementos correspondentes de duas linhas ou colunas são iguais ou possuem uma razão constante, o determinante será igual a ***zero***;
- Se um múltiplo escalar de uma linha, ou uma coluna da matriz é adicionada à outra linha desta matriz, o determinante permanece inalterado;
- Se a matriz for triangular, o valor do determinante é obtido pelo produto de todos os elementos da diagonal principal;

## Solução de sistema de equações por métodos diretos

Vimos que é possível encontrar-se a matriz solução de um sistema, pelo emprego da matriz inversa, porém, embora tal método tenha solução aparentemente simples, envolve a inversão da matriz de coeficientes, o que dá mais trabalho do que resolver o sistema.
Num sistema de equações pode-se operar algumas transformações que não alteram os valores das incógnitas e a essas transformações se dá o nome de transformações elementares, sendo elas:

- ***Troca de posição entre duas equações***;
- ***Multiplicar toda uma equação por uma constante não nula;***
- ***Substituir uma equação por uma soma dela com qualquer outra equação do sistema, multiplicada ou não por uma constante qualquer.***

Pelo processo da matriz inversa percebe-se que ao multiplicar a matriz inversa pela matriz dos coeficientes esta se transforma em matriz identidade. Da mesma maneira, ao multiplicar a matriz inversa dos coeficientes pela matriz resultado (matriz B) a mesma se transforma na matriz solução (matriz X). Assim, se pudermos aplicar as transformações elementares para converter simultaneamente a matriz dos coeficientes em matriz identidade, a matriz dos resultados se converterá em matriz solução. Ampliando este método fazendo o mesmo processo com a matriz identidade, esta se converte automaticamente em matriz inversa da matriz dos coeficientes.




Os algoritmos de eliminação constam da execução sistemática das operações elementares sobre as equações, de maneira a anular os coeficientes quantos sejam necessários para que a matriz de coeficientes se torne a matriz identidade. Vários algoritmos são propostos na literatura.

- *Gauss*: Neste algoritmo transforma-se a matriz de coeficientes em matriz triangular. A matriz solução se obtém extraindo-se sucessivamente as incógnitas por substituição regressiva. Neste caso é possível encontrar o valor do determinante, uma vez que em matriz triangular ou diagonal, o mesmo é o produto de todos os elementos da diagonal.

- *Gauss-Jordan*: Consiste em transformar a matriz de coeficientes em matriz diagonal. Neste método, pode-se dividir cada valor da diagonal pelo próprio valor para transformá-lo na unidade, desde que esta divisão seja feita com toda linha do sistema, o valor da matriz B, irá ser convertido no valor solução.

Outros métodos diretos para resolução de sistema de equações:
- Método de Cramer. Usa operações com determinantes;
- Método pela redução de Crout;

Exemplo (método da eliminação de Gauss):
Seja o sistema:

$$\begin{aligned} 3x_1 + x_2 &= -2 \\ -2x_1 + 2x_3 &= 1 \\ x_1 + 3x_2 + 4x_3 &= 0 \end{aligned}$$

Em notação matricial, o sistema fica:

$$\begin{bmatrix} 3 & 1 & 0 \\ -2 & 0 & 2 \\ 1 & 3 & 4 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}$$

Como as operações listadas somente afetam os coeficientes e os segundos membros, é conveniente fazê-las sobre a matriz aumentada cujas linhas serão denotadas por $l_1$, $l_2$ e $l_3$.

$$\begin{matrix} \boldsymbol{l_1} \\ \boldsymbol{l_2} \\ \boldsymbol{l_3} \end{matrix} \quad \begin{bmatrix} 3 & 1 & 0 & -2 \\ -2 & 0 & 2 & 1 \\ 1 & 3 & 4 & 0 \end{bmatrix}$$

$$\boldsymbol{l'_{2j} = l_{1j} \text{ x } (2/3) + l_{2j}} \qquad \boldsymbol{l'_{3j} = l_{1j} \text{ x } (-1/3) + l_{3j}} \quad \begin{bmatrix} 3 & 1 & 0 & -2 \\ 0 & 0,667 & 2 & -0,333 \\ 0 & 2,667 & 4 & 0,667 \end{bmatrix} \qquad \boldsymbol{j = 1 \text{ até } 4}$$

$$\boldsymbol{l''_{3j} = l'_{2j} \text{ x } (-2,667/0,667) + l'_{3j}} \quad \begin{bmatrix} 3 & 1 & 0 & -2 \\ 0 & 0,667 & 2 & -0,333 \\ 0 & 0 & -4 & 2 \end{bmatrix}$$

Para se obter o valor das incógnitas, é suficiente calcular o valor de $x_3$ na terceira equação, substitui-lo na segunda equação, calculando então $x_2$ e então substituir ambos na primeira equação para calcular $x_1$. A esta parte da solução dá-se o nome de *substituição regressiva***.** *O valor do determinante se obtém pelo produto dos elementos da diagonal principal (3 x 2/3 x (-4) = -8).*

Da terceira equação: $x_3 = \dfrac{2}{-4} = -0,5$

Da segunda equação:

$$\begin{aligned} 0,667x_2 + 2(-0,5) &= -0,3333 \\ 0,667x_2 &= -0,333 + 1 = 0,667 \\ x_2 &= 1 \end{aligned}$$

Da primeira equação:

$$\begin{aligned} 3x_1 + 1 &= -2 \\ 3x_1 &= -3 \\ x_1 &= -1 \end{aligned}$$

Embora se possa terminar neste ponto, pode-se optar em continuar o processo, diagonalizando a matriz de coeficientes totalmente convertendo os elementos acima da diagonal principal em valores nulos, conforme abaixo:

$$\boldsymbol{l'_{1j} = l'_{2j} \text{ x } (-3/2) + l_{1j}} \quad \begin{bmatrix} 3 & 0 & -3 & -1,5 \\ 0 & 0,667 & 2 & -0,333 \\ 0 & 0 & -4 & 2 \end{bmatrix}$$

$$\boldsymbol{l''_{1j} = l''_{3j} \text{ x } (-3/4) + l'_{1j}} \qquad \boldsymbol{l''_{2j} = l''_{3j} \text{ x } 2/4 + l'_{2j}} \quad \begin{bmatrix} 3 & 0 & 0 & -3 \\ 0 & 0,667 & 0 & 0,667 \\ 0 & 0 & -4 & 2 \end{bmatrix}$$

Os elementos da diagonal da matriz de coeficientes, que são usados para se fazer às eliminações dos elementos da diagonal e que permanecem no final, são chamados *pivôs* da eliminação.
Dividindo-se cada elemento da diagonal por ele próprio, lembrando que ao multiplicar ou dividir um elemento da matriz, toda linha deverá sofrer a mesma transformação, convertem-se esses pivôs em unidades e conseqüentemente a coluna estendida da matriz irá ser transformada no vetor solução.

$$\begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -0,5 \end{bmatrix}$$

Vetor solução: **[ -1 1 -0,5 ]$^T$**

No exemplo anterior as contas foram feitas com elevada precisão, trabalhando com frações e não com as representações decimais de frações, e o resultado obtido foi exato. Entretanto, ao trabalhar com um número definido de algarismos em qualquer base numérica, os erros de arredondamento podem se acumular, levando a resultados totalmente sem significado.

Vejamos o exemplo:

$$\begin{array}{rcrcrcr} -x_1 & + & 2x_2 & + & 42x_3 & = & 83 \\ 72x_1 & - & 41x_2 & - & 14x_3 & = & 44 \\ 35x_1 & + & 10x_2 & - & 5x_3 & = & 25 \end{array}$$

cuja solução verdadeira é: $x_1 = 1,$ $x_2 = 0,$ $x_3 = 2$

Convertendo o sistema em matriz ampliada vem:

$$\begin{array}{l} \boldsymbol{l_1} \\ \boldsymbol{l_2} \\ \boldsymbol{l_3} \end{array} \quad \begin{bmatrix} -1 & 2 & 42 & 83 \\ 72 & -41 & -14 & 44 \\ 35 & 10 & -5 & 25 \end{bmatrix}$$

As várias etapas de solução pelo algoritmo de diagonalização **mantendo três algarismos significativos,** caso típico quando se processa esse tipo de cálculo eletronicamente, onde se fixa a quantidade de dígitos significativos, são:

$\boldsymbol{l'_{2j} = l_{1j} \times 72 + l_{2j}}$
$\boldsymbol{l'_{3j} = l_{1j} \times 35 + l_{3j}}$

$$\begin{bmatrix} -1 & 2 & 42 & 83 \\ 0 & 103 & 3010 & 6020 \\ 0 & 80 & 1460 & 2940 \end{bmatrix}$$

$\boldsymbol{l''_{3j} = l'_{2j} \times (-80/103) + l'_{3j}}$

$$\begin{bmatrix} -1 & 2 & 42 & 83 \\ 0 & 103 & 3010 & 6020 \\ 0 & 0 & -880 & -1740 \end{bmatrix}$$

$\boldsymbol{l'_{1j} = l'_{2j} \times (-2/103) + l_{1j}}$

$$\begin{bmatrix} -1 & 0 & -16,4 & -34 \\ 0 & 103 & 3010 & 6020 \\ 0 & 0 & -880 & -1740 \end{bmatrix}$$

$\boldsymbol{l''_{1j} = l''_{3j} \times (-16/880) + l'_{1j}}$
$\boldsymbol{l''_{2j} = l''_{3j} \times (3010/880) + l'_{2j}}$

$$\begin{bmatrix} -1 & 0 & 0 & -1,6 \\ 0 & 103 & 0 & 70 \\ 0 & 0 & -880 & -1740 \end{bmatrix}$$

donde se obtém: $x_1 = 1,6$ $x_2 = 0,68$ $x_3 = 1,98$

Solução com erros bastante grandes, especialmente em $x_1$ e $x_2$. Ainda mais, se a solução fosse feita pela transformação da matriz de coeficientes em unidade, ficaria:

$l_{1j} = l_{1j} / (-1)$
$l'_{2j} = l_{1j} \times (-72) + l_{2j}$
$l'_{3j} = l_{1j} \times (-35) + l_{3j}$

$$\begin{bmatrix} 1 & -2 & -42 & -83 \\ 0 & 103 & 3010 & 6020 \\ 0 & 80 & 1460 & 2940 \end{bmatrix}$$

$l'_{2j} = l'_{2j} / 103$
$l''_{3j} = l'_{2j} \times (-80) + l'_{3j}$

$$\begin{bmatrix} 1 & -2 & -42 & -83 \\ 0 & 1 & 29,2 & 58,4 \\ 0 & 0 & -880 & -1730 \end{bmatrix}$$

$l'_{1j} = l'_{2j} \times 2 + l_{1j}$

$l''_{3j} = l''_{3j} / (-880)$

$$\begin{bmatrix} 1 & 0 & 16,4 & 34 \\ 0 & 1 & 29,2 & 58,4 \\ 0 & 0 & 1 & 1,97 \end{bmatrix}$$

$l''_{1j} = l''_{3j} \times (-16,4) + l'_{1j}$
$l''_{2j} = l''_{3j} \times (-29,2) + l'_{2j}$

$$\begin{bmatrix} 1 & 0 & 0 & 1,7 \\ 0 & 1 & 0 & 0,9 \\ 0 & 0 & 1 & 1,97 \end{bmatrix}$$

Como se vê, a alteração mínima na ordem das contas alterou substancialmente o resultado, especialmente no valor de $x_2$, que passou de 0,68 para 0,9.

Para se evitar esse tipo de imprecisão, algumas técnicas podem ser adotadas entre as quais existem a **condensação pivotal**, e o **refino da solução pelo método dos resíduos**.

A técnica de **condensação pivotal,** que pode ser aplicada a qualquer um dos métodos apresentados, destina-se a evitar os erros de arredondamento e consta de, a cada eliminação, rearranjar primeiro as equações do sistema de maneira que o elemento da diagonal que intervirá nos cálculos (chamado pivô) seja o **maior valor absoluto** dos coeficientes restantes no sistema. No exemplo acima ficaria:

$$\begin{bmatrix} -1 & 2 & 42 & 83 \\ 72 & -41 & -14 & 44 \\ 35 & 10 & -5 & 25 \end{bmatrix}$$

Antes da primeira eliminação, rearranjar trocando a primeira e a segunda equação para:

A primeira eliminação dá:

$$\begin{bmatrix} 72 & -41 & -14 & 44 \\ -1 & 2 & 42 & 83 \\ 35 & 10 & -5 & 25 \end{bmatrix}$$

$l'_{2j} = l_{1j} \times (1/72) + l_{2j}$
$l'_{3j} = l_{1j} \times (-35/72) + l_{3j}$

$$\begin{bmatrix} 72 & -41 & -14 & 44 \\ 0 & 1,43 & 41,8 & 83,6 \\ 0 & 29,9 & 1,8 & 3,6 \end{bmatrix}$$

Antes da segunda eliminação, trocar a segunda e a terceira coluna (observar a troca de posição de $x_2$ e $x_3$) para:

$$\begin{bmatrix} 72 & -14 & -41 & 44 \\ 0 & 41,8 & 1,43 & 83,6 \\ 0 & 1,8 & 29,9 & 3,6 \end{bmatrix} \quad \begin{matrix} x_1 \\ x_3 \\ x_2 \end{matrix}$$

$l'_{1j} = l'_{2j}\ x\ (14/41{,}8) + l_{1j}$

$l''_{3j} = l'_{2j}\ x\ (-1{,}8/41{,}8) + l'_{3j}$

$$\begin{bmatrix} 72 & 0 & -40{,}5 & 72{,}0 \\ 0 & 41{,}8 & 1{,}43 & 83{,}6 \\ 0 & 0 & 29{,}8 & 0{,}0 \end{bmatrix} \quad \begin{matrix} x_1 \\ x3 \\ x2 \end{matrix}$$

$l''_{1j} = l''_{3j}\ x\ (40{,}5/29{,}8) + l'_{1j}$

$l''_{2j} = l''_{3j}\ x\ (-1{,}43/29{,}8) + l'_{2j}$

$$\begin{bmatrix} 72 & 0 & 0 & 72{,}0 \\ 0 & 41{,}8 & 0 & 83{,}6 \\ 0 & 0 & 29{,}8 & 0 \end{bmatrix}$$

Após a terceira eliminação, temos as três incógnitas com resultados exatos até o terceiro algarismo.

$x_1 = 72/72 = 1{,}00$ $\quad x_2 = 0/29{,}8 = 0{,}00$ $\quad x_3 = 83{,}6/41{,}8 = 2{,}00$

Um outro método, o refino da solução pelo método dos resíduos pode ser também usado. Depois de determinadas as raízes de um sistema de equações lineares por qualquer dos métodos já visto, **deve-se sempre substituir os valores obtidos no sistema e verificar se este é satisfeito**. Isto ocorrerá se os segundos membros forem reconstituídos exatamente. Se tal não ocorrer, o vetor ***diferença*** dos segundos membros obtidos e dos originais é chamado de ***resíduo*** da solução e é denotado por ***r.***

Usando o sistema anterior como exemplo, fazemos a verificação...

$x_1 = 1{,}7$ $\quad x_2 = 0{,}9$ $\quad x_3 = 1{,}97$

Fica:

$$\begin{bmatrix} -1 & 2 & 42 \\ 72 & -41 & -14 \\ 35 & 10 & -5 \end{bmatrix} \times \begin{bmatrix} 1{,}7 \\ 0{,}9 \\ 1{,}97 \end{bmatrix} = \begin{bmatrix} 82{,}840 \\ 57{,}920 \\ 58{,}650 \end{bmatrix}$$

As contas serão feitas com um algarismo a mais para que os resíduos possam ser calculados com melhor precisão. Seus valores são:

$$\begin{bmatrix} 82{,}84 \\ 57{,}92 \\ 58{,}65 \end{bmatrix} - \begin{bmatrix} 83 \\ 44 \\ 25 \end{bmatrix} = \begin{bmatrix} -0{,}16 \\ 13{,}92 \\ 33{,}75 \end{bmatrix}$$

Como se vê, os erros (enormes) nos resultados dão resíduos que são até maiores que o segundo membro correspondente.

Devido à linearidade do sistema, sua solução usando o resíduo ***r*** como segundo membro deverá dar um vetor de correções dos valores das raízes obtidas. Assim, é possível refinar a solução obtida resolvendo-se novamente o sistema. Tal nova solução consiste somente de operações envolvendo os segundos membros, pois, como os coeficientes não se alteram, as operações que os envolvem já estão feitas. O exemplo a seguir mostra a determinação das correções usando os resíduos obtidos acima.

$$\begin{bmatrix} -1 & 2 & 42 \\ 72 & -41 & -14 \\ 35 & 10 & -5 \end{bmatrix} \times \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} = \begin{bmatrix} -0{,}16 \\ 13{,}92 \\ 33{,}75 \end{bmatrix}$$

$$\begin{bmatrix} -1 & 2 & 42 & -0,16 \\ 72 & -41 & -14 & 13,92 \\ 35 & 10 & -5 & 33,75 \end{bmatrix}$$

$$\begin{bmatrix} 1 & -2 & -42 & 0,16 \\ 0 & 103 & 3010 & 2,40 \\ 0 & 80 & 1460 & 28,1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 16,4 & 0,207 \\ 0 & 1 & 29,2 & 0,0235 \\ 0 & 0 & -880 & 26,2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0,696 \\ 0 & 1 & 0 & 0,894 \\ 0 & 0 & 1 & -0,0298 \end{bmatrix}$$

$$\begin{bmatrix} 1,7 \\ 0,9 \\ 1,97 \end{bmatrix} - \begin{bmatrix} 0,696 \\ 0,894 \\ -0,0298 \end{bmatrix} = \begin{bmatrix} 1,00 \\ 0,00 \\ 2,00 \end{bmatrix}$$

*Como se verifica, foi agora obtida a solução exata do sistema. Isto não ocorre sempre no primeiro refinamento, podendo ser necessário continuar refinando até que o resíduo se anule, ou seja, inferior a uma tolerância pré-fixada, ou ainda, que a diferença entre sucessivos valores das incógnitas esteja abaixo de uma determinada tolerância.*

**Observações importantes:**

- Nos métodos diretos de resolução de sistemas de equações lineares, a quantidade de operações a serem realizadas é finita, e dependem da ordem do sistema;
- O erro de arredondamento é sempre uma função do método utilizado;

**Exercícios**:
*Resolver os seguintes sistemas de equações lineares:*
(empregar o método de Gauss-Jordan)

1.
$$\begin{aligned} x1 + 6x2 + 2x3 + 4x4 &= 8 \\ 3x1 + 19x2 + 4x3 + 15x4 &= 25 \\ x1 + 4x2 + 8x3 + 12x4 &= 18 \\ 5x1 + 33x2 + 9x3 + 3x4 &= 72 \end{aligned}$$

2.
$$\begin{aligned} 2x1 + 3x2 + 4x3 + 5x4 &= 14 \\ 4x1 + 6x2 + x3 + x4 &= 12 \\ 2x1 + x2 + x3 + x4 &= 5 \\ 4x1 - 2x2 - 2x3 + x4 &= 1 \end{aligned}$$

3.
$$\begin{aligned} 1,0234x1 - 2,4567x2 + 1,2345x3 &= 6,6728 \\ 5,0831x1 + 1,2500x2 + 0,9878x3 &= 6,5263 \\ -3,4598x1 + 2,5122x2 - 1,2121x3 &= -11,2784 \end{aligned}$$

## Aritmética em ponto flutuante em computadores.

Aos estudantes da escola primária são apresentados os números reais em etapas. Primeiro, os "números de contagem" (inteiros positivos); depois, os números negativos (completando o conjunto de todos os inteiros); finalmente, as frações e seus equivalentes decimais. Para demonstrar esses conceitos, os professores usam a ***reta numérica***, a qual é uma linha horizontal com setas em ambas extremidades apontando para o infinito negativo à esquerda e para o infinito positivo à direita e uma divisão no ponto médio indicando o zero.
Uma vez que os inteiros

...–3, –2, –1, 0, 1, 2, 3, .....

Sejam representados na reta numérica, é bastante fácil esclarecer a noção de números entre números, tendo como exemplo a fração, simbolizada por um ponto da reta numérica entre dois valores numéricos e depois compreender que pode ser encontrada uma quantidade infinita de números entre qualquer par de pontos da reta numérica, com a reta inteira correspondendo ao conjunto completo dos números reais.
Mais tarde, no curso secundário, se nossos estudantes hipotéticos escolherem álgebra, química ou física, lhes será ensinada a notação científica ou exponencial, a qual é uma forma mais genérica de representar números reais de qualquer tamanho ou precisão.
Como exemplo, toma-se o valor 0,25, o qual representado na notação científica é escrito como $2{,}5 \times 10^{-1}$. A parte 2,5 do número é chamada **mantissa, fração** ou **significando**. Sempre aparece um dígito diferente de zero à esquerda do ponto decimal, e o número de dígitos após o ponto decimal indica o ***grau de precisão*** com que o valor do número é conhecido. *Uma mantissa apresentada nesta forma é dita normalizada.* A parte $10^{-1}$ é chamada de **expoente** ou **característica**. Ela especifica a localização do ponto decimal dentro do número. A maioria dos professores também apresenta aos estudantes um conjunto completo de regras para as operações com números escritos em notação científica, incluindo instruções como: *"Para encontrar o produto de dois números, multiplique as mantissas e some os expoentes da parte característica"*.
Tendo dominado os números reais e os meios para manipulá-los, nossos meninos prodígios da escola secundária podem ser tentados a adotar uma atitude um pouco mais pretensiosa em relação à matemática de um modo geral. Evidentemente, essa visão prematuramente otimista será demolida quando eles se confrontarem com os números imaginários, os números irracionais, os números complexos, os infinitesimais e todos os outros monstros matemáticos não intuitivos.

Embora os métodos usados para lidar com números em ponto flutuante nos computadores sejam certamente baseados nas regras e algoritmos fundamentais que se aprende na escola, existem várias diferenças importantes. Para começar, a maioria dos computadores e bibliotecas de linguagens de alto nível suporta somente um número limitado de formatos em ponto flutuante, em geral apenas dois formatos e esses formatos têm naturalmente uma precisão e faixa de variação **finita**. Portanto, não se pode representar todos os números reais como número em ponto flutuante em um computador. Na verdade, a quantidade de números que não se pode representar é infinitamente maior que a quantidade de números que se pode representar. Os formatos foram escolhidos porque pareceram suficientes para representar a maior parte das aplicações normais e que podiam ser implementados de modo eficaz. Se os requisitos de algum programa qualquer ultrapassar os limites previstos pelos formatos normais, deve-se então se dispor de rotinas próprias para atender estes requisitos ou então simplesmente desistir da solução numérica do programa. Além disso, os números em ponto flutuantes num computador qualquer não estão mapeados de maneira uniforme em relação aos números reais. Por exemplo, quando se marca os números que podem ser representados por um inteiro em 32 bits na reta numérica real, se percebe um conjunto de pontos marcados uniformemente desde –2.147.483.648 ($-2^{31}$) até 2.147.483.647 ($2^{31}-1$). Porém, ao se marcar os números que podem ser marcados como números reais em 32 bits na reta numérica real, percebe-se que essa quantidade é exatamente igual a quantidade de inteiros que são representados, mas agora densamente agrupados em torno do valor zero, ficando cada vez mais esparsos a medida que aumenta a distância do número em relação a zero. Evidentemente, os números mais baixos e mais altos em ponto flutuantes são, respectivamente, muitos menores, e muito maior que os números inteiros mais baixo e mais alto, mas essa faixa dinâmica é obtida em prejuízo da precisão. Por fim, as pessoas inegavelmente preferem efetuar cálculos em base numérica 10, enquanto que os computadores o fazem pela base numérica 2. Este fato, força a todo cálculo, sempre duas conversões obrigatórias. Uma, de decimal para binário no inicio e outra de binário para decimal no final. Toda essa transformação raramente causa problemas com números inteiros, mas se torna um problema espinhoso quando os números estão em ponto flutuantes, porque alguns números decimais não podem ser expressos exatamente na forma binária em ponto flutuante, sendo exemplo deste fato o número $1{,}0 \times 10^{-1}$.

**Formato de dados binários em ponto flutuante:**
Depois que um número decimal em ponto flutuante é convertido em um número binário em ponto flutuante normalizado, ele terá a forma mostrada abaixo, onde cada bit b na mantissa é 0 ou 1.

$$1.bbbbbbbb.... \times 2^{n}$$

A mantissa é normalizada ajustando-se o expoente de forma que o bit 1 mais significativo fique à esquerda do ponto binário. Em outras palavras, a mantida é sempre maior ou igual a 1 e menor que 2. Mas como esses números são realmente armazenados na memória do computador?...Nos primeiros dias da computação, praticamente cada compilador e cada UCP usavam um formato de dados em ponto flutuante diferente, o que tornava difícil transportar dados de um computador para outro ou até mesmo de um programa em uma linguagem de alto nível para um programa escrito em outra linguagem. No final dos anos 70, começou um esforço para padronizar a aritmética binária em ponto flutuante, primeiro sob os auspícios da ACM e depois patrocinado pela IEEE Computer Society. Esse empreendimento delineou diversas posições, sendo a mais importante aquela que representava a integração de conceitos e técnicas que provinham dos primórdios da informática. Em 1981, a IEEE publicou o trabalho IEEE-754 (Normas para aritmética em ponto flutuante binária) que foi modificada e adotada em 1985 como padrão oficial ANSI. Mais tarde remodelada mais uma vez de uma forma generalizada e apresentada como IEEE-854, a qual perdura como padrão oficial desde 1987. O padrão ANSI/IEEE –754 era principalmente dirigido à elaboração de cálculos em ponto flutuante de forma segura e previsível, para programadores que possuíam treinamento em análise numérica. Ele especificava o grau de precisão com que os cálculos deviam ser feitos, o comportamento durante o arredondamento, o tratamento de erros e exceções e os resultados das operações básicas, comparações e conversões em ponto flutuante. Ele também especificava os formatos binários para os números em ponto flutuante, que foram rapidamente adotados pela indústria e que são agora amplamente suportados, em nível de hardware e de software.

Os dois formatos de dados em ponto flutuante mais importantes descritos pelo padrão ANSI/IEEE 754 estão mostrados abaixo. O formato de precisão simples ocupa 32 bits. O formato de precisão dupla exige 64 bits. Ambos os formatos consistem em três campos: um bit de sinal, sendo sempre o mais significativo, seguido pelo expoente binário e a mantissa nos bits restantes menos significativos. Os números em precisão simples podem corresponder a valores na faixa aproximada de $\pm 1{,}18 \times 10^{-38}$ até
$\pm 3{,}40 \times 10^{38}$, enquanto que os de precisão dupla residem na faixa de $\pm 2{,}23 \times 10^{-308}$ até $\pm 1{,}80 \times 10^{308}$.

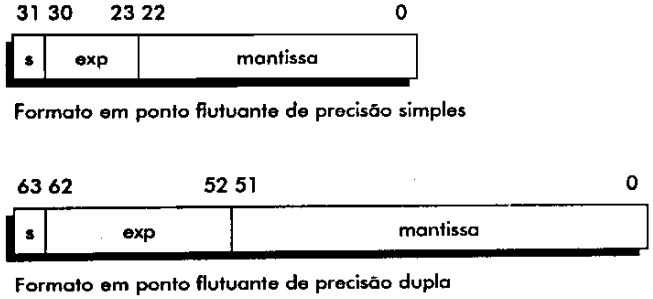



**Figura 13-1.** *Os formatos em ponto flutuante de precisão simples dupla especificados pelo padrão ANSI/IEEE 754.*

O bit de sinal tem valor 1 se o número for negativo e 0 se for positivo. A mantissa é sem sinal e não muda com o sinal do número em ponto flutuante. Como a mantissa é normalizada à esquerda, seu bit mais significativo é por definição sempre 1. Consequentemente, os projetistas do IEEE 754 criaram um artificio elegante: especificaram que a mantissa tem um *bit inicial implicito* que é sempre 1 e não está presente nos dados reais. Isso permite que seja extraído um bit a mais de precisão de cada formato em ponto flutuante.

O campo do expoente dos tipos de números em ponto flutuante é predifinido com deslocamento a partir de 0 em uma quantidade fixa. Para números de precisão simples que têm uma característica de 8 bits, um valor 127 (7F em hexadecimal) no campo do expoente corresponde a um valor real 0 para expoente. Para os números em precisão dupla, o expoente usa um campo de 11 bits, e o desvio aplicado ficou em 1023 (3FF em hexadecimal). Esse desvio permite que o inverso de qualquer número em ponto flutuante normalizado seja representado sem estouro de campo (underflow). Os tamanhos relativos dos campos de expoentes nos dois formatos foram escolhidos de modo

a permitir que um número em precisão dupla acomode o produto de até oito números de precisão simples, sem possilbilidade de estouro (overflow).

O expoente de um número em ponto flutuante com base no ANSI/IEEE 754 também pode se basear em dois ***valores mágicos*** que fazem com que o número seja tratado de uma forma especial. Se todos os bits do expoente são iguais a 0, então o número em ponto flutuante é zero ou um número ***desnormalizado*** (resultado de um underflow gradual). Se todos os bits do expoente são 1, então o número em ponto flutuante representa o infinito ou um valor de sinalização especial chamado de NaN (não é um número) conforme mostrado na tabela abaixo.

| Bits de expoente | Bits de mantissa | Significado especial |
|---|---|---|
| Não 0 | Não 0 | Número em ponto flutuante |
| Todos 0 | Todos 0 | Zero em ponto flutuante |
| Todos 0 | Não 0 | Número em ponto flutuante |
| Todos 1 | Todos 0 | Infinito |
| Todos 1 | Não 0 | NaN – Não é um número. |

Ao examinar alguns exemplos práticos de dados binários em ponto flutuantes temos:
Em precisão simples, observar o número, em notação hexadecimal: 41 20 00 00, onde o byte 41 é o byte mais significativo e o byte 00 último é o byte menos significativo.

Convertendo estes bytes ao equivalente em binário tem-se:

0100 0001 0010 0000 0000 0000 0000 0000.

O primeiro bit, o bit de sinal é 0. Em seguida vem os 8 bits do expoente, 1000 0010 e a mantissa depois de restaurado o bit inicial implícito, resulta em:

1010 0000 0000 0000 0000 0000.

Corrigindo-se o desvio do expoente, resulta em binário o valor $1{,}01 \times 2^3$ que em notação decimal é o valor 10.

Outro exemplo, agora em precisão dupla, observar o número em hexadecimal: BF E0 00 00 00 00 00 00, onde BF é o byte mais significativo e o último 00 é o byte menos significativo.

Convertendo-se esse valor para a notação em binário obtém-se:

1011 1111 1110 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000

O bit de sinal é 1, o expoente predefinido em binário é 01111111110 e a mantissa depois da inserção do bit inicial implícito é

10000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000

A correção do desvio do expoente produz em binário o valor $-1{,}0 \times 2^{-1}$ que em notação decimal é escrito como –0,5.

## Exemplos de programas em FORTRAN aplicados ao cálculo numérico:

### 1. Para produto entre duas matrizes:

Algoritmo do produto entre duas matrizes
Condição: O número de colunas da primeira matriz deve ser igual ao número de colunas da segunda matriz;
A(m,n) * B(n,l) = C(m,l)

$$C_{ij} = \sum_{k=1}^{n} a_{ik} * b_{kj}$$ (i varia de 1 até *m* e j varia de 1 até *l*)

Trechos do programa:

**Definições iniciais**

```
Dimension a(m,n),b(n,l),c(m,l)
.....
......
......
```

**Leitura de matrizes**

```
      Read (2,50)((a(i,j),j=1,n),i=1,m)
30    Format(10F8,2)
......
......
      do 10 i=1,m
      read(2,30) (a(i,j),j=1,n)
      write(5,40)(a(i,j),j=1,n)
10    continue
30    format(10F8,2)
40    format(10x,10F8,2)
```

*rotina de multiplicação de matrizes*

```
      soma=0
      do 5 i=1,m
      do 6 j=1,l
      do 7 k=1,n
      soma=a(i,k)*b(k,j)+soma
7     continue
      c(i,j)=soma
      soma=0
6     continue
5     continue
```

**2. Para inversão de matrizes pelo método de Shipley-Coleman:**

```
      DIMENSION A(15,15)
      WRITE (5,9)
9     FORMAT(‘ 1’)
      WRITE(5,6)
6     FORMAT(1X,’INVERSÃO DE MATRIZ PELO METODO DE SHIPLEY COLEMAN’,/,)
      READ(2,20) M1
      DO 25 M=1,M1
      WRITE(5,10)
10    FORMAT(10X, ‘ MATRIZ ORIGINAL’,/)
      READ(2,20) N
20    FORMAT(I2)
      DO 30 I=1,N
      READ(2,40)(A(I,J),J=1,N)
      WRITE(5,50)(I,(A(I,J),J=1,N)
30    CONTINUE
40    FORMAT(8f10,1)
50    FORMAT(‘0’,I2,4X,10F11,3)
      DO 100 K=1,2
      WRITE(5,60) K
60    FORMAT(////,11X,’MATRIZ DEPOIS DA INVERSÃO N.’,I2,//)
C     ***************************************************************************
      DO 1 L=1,N
      A(L,L)=–1/A(L,L)
      DO 3 I=1,N
      IF(I–L)24,3,24
24    A(I,L)=A(I,L)*A(L,L)
3     CONTINUE
      DO 4 J=1,N
      B=A(L,J)
      DO 4 I=1,N
      IF(J–L)21,4,21
21    IF(I–L)22,23,22
23    A(L,J)=A(L,J)*A(L,L)
      GO TO 4
22    A(I,J)=A(I,J)+B*A(I,L)
4     CONTINUE
1     CONTINUE
      DO 55 I=1,N
      DO 55 J=1,N
55    A(I,J)=–A(I,J)
C     ***************************************************************************
      DO 100 I=1,N
100   WRITE(5,50) I,(A(I,J),J=1,N)
25    CONTINUE
      STOP
      END
```

**Solução de sistema de equações lineares pela redução de Crout**

A resolução de um sistema pelo método de Crout baseia-se na substituição da matriz expandida de acordo com as seguintes transformações com *i* e *j* variando de 1 até n:

- Para os elementos onde $i \geq j$:

$$a'_{ij} = a_{ij} - \sum_{k=1}^{j-1} a'_{ik} \times a'_{kj}$$

- Para os elementos onde $i < j$:

$$a'_{ij} = \frac{1}{a'_{ii}} \left[ a_{ij} - \sum_{k=1}^{i-1} a'_{ik} \times a'_{kj} \right]$$

- Para os termos da matriz b:

$$b'_i = \frac{1}{a'_{ii}} \left[ b_i - \sum_{k=1}^{i-1} a'_{ik} \times b'_k \right]$$

- Os valores de x são então obtidos pela transformação:

$$x_i = b'_i - \sum_{k=i+1}^{n} a'_{ik} \times x_k \qquad (i \text{ deve variar de } n \text{ até } 1)$$

Exemplo:
Seja o sistema

$$\begin{array}{rcrcrcr} 2x_1 & - & 3x_2 & + & x_3 & = & 1 \\ x_1 & + & 2x_2 & + & 3x_3 & = & 3 \\ 4x_1 & - & x_2 & - & x_3 & = & 1 \end{array}$$

A matriz expandida fica:

$$\begin{bmatrix} 2 & -3 & 1 & 1 \\ 1 & 2 & 3 & 3 \\ 4 & -1 & -1 & 1 \end{bmatrix}$$

Fazendo as transformações, vem:

$a'_{11} = a_{11} = 2 \qquad (i = j)$

$a'_{12} = \frac{1}{a'_{11}} \times a_{12} = \frac{1}{2} \times (-3) = -\frac{3}{2} \qquad (i < j)$

$a'_{13} = \frac{1}{a'_{11}} \times a_{13} = \frac{1}{2} \times (1) = \frac{1}{2} \qquad (i < j)$

$b'_1 = \frac{1}{a'_{11}} \times b_1 = \frac{1}{2} \times 1 = \frac{1}{2}$

$a'_{21} = a_{21} = 1 \qquad (i \geq j)$

$$a'_{22} = a_{22} - a'_{21} \times a'_{12} = 2 - 1 \times (-\frac{3}{2}) = \frac{4}{2} + \frac{3}{2} = \frac{7}{2} \qquad (i \geq j)$$

$$a'_{23} = \frac{1}{a'_{22}} \times \left[a_{23} - a'_{21} \times a'_{13}\right] = \frac{2}{7} \times \left[3 - 1 \times \frac{1}{2}\right] = \frac{2}{7} \times \frac{5}{2} = \frac{10}{14} = \frac{5}{7} \qquad (i < j)$$

$$b'_2 = \frac{1}{a'_{22}} \times \left[b_2 - a'_{21} \times b'_1\right] = \frac{2}{7} \times \left[3 - 1 \times \frac{1}{2}\right] = \frac{2}{7} \times \frac{5}{2} = \frac{10}{14} = \frac{5}{7}$$

$$a'_{31} = a_{31} = 4 \qquad (i \geq j)$$

$$a'_{32} = a_{32} - a'_{31}.a'_{12} = -1 - 4 \times (-\frac{3}{2}) = -1 + \frac{12}{2} = -1 + 6 = 5 \qquad i \geq j)$$

$$a'_{33} = a_{33} - a'_{31} \times a'_{13} - a'_{32} \times a'_{23} =$$

$$b'_3 = \frac{1}{a'_{33}} \times \left[b_3 - a'_{31} \times b'_1 - a'_{32} \times b'_2\right] =$$

Os valores de $x_i$ podem agora ser obtidos, em ordem reversa:

$$x_3 = b'_3 = \frac{16}{23}$$

$$x_2 = b'_2 - a'_{23}.x_3 = \frac{5}{23}$$

$$x_1 = b'_1 - a'_{12}.x_2 - a'_{13}.x_3 = \frac{11}{23}$$

Resumo das operações:

- Manter todos os elementos da 1ª coluna sem transformação;
- Dividir a 1ª linha, incluindo o valor $b_1$, exceto o 1º elemento pelo pivô (elemento 1,1)
- Todos os demais elementos abaixo da diagonal principal são transformados tomando-se o elemento em questão e subtraindo dele a soma dos produtos a(i,k) pelo correspondente a(k,j) variando k desde 1 até o valor j –1;
- Todos os demais elementos acima da diagonal principal, inclusive os valores $b_i$ , sofrem o mesmo procedimento, exceto que k varia desde 1 até i –1, e ainda são divididos pelo valor da diagonal principal correspondente obtido na operação anterior;
- Uma vez terminada a operação anterior, proceder ao cálculo das variáveis, retornando, isto é, iniciando o cálculo das variáveis pelo ultimo valor de $b_i$.

Exercícios propostos:

Resolver pelo método de Crout, os seguintes sistemas de equações lineares:

1.

$$\begin{aligned} 3x_1 + x_2 + &= -2 \\ -2x_1 + 2x_3 &= 1 \\ x1 + 3x_2 + 4x_3 &= 0 \end{aligned}$$

2.

$$\begin{aligned} 2x_1 + 8x_2 + 10x_3 + 1x_4 + 2x_5 + 18x_6 &= 68 \\ 4x_1 + 15x_2 + 220x_3 + 1x_4 + 3x_5 + 30x_6 &= 121 \\ 2x_1 + 6x_2 + 13x_3 + 3x_4 + 2x_5 + 16x_6 &= 67 \\ 6x_1 + 20x_2 + 26x_3 + 2x_4 + 5x_5 + 50x_6 &= 181 \\ 1x_1 + 3x_2 + 4x_3 + 5x_4 + 1x_5 + 19x_6 &= 60 \\ 2x_1 + 6x_2 + 5x_3 + 1x_4 + 2x_5 + 18x_6 &= 59 \end{aligned}$$

3. Com o sistema já na forma de matriz estendida:

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

Resultados: 1.- x1=
x2=
x3=

2.- x1=
x2=
x3=
x4=
x5=
x6=

3.- x1=
x2=
x3=
x4=

4.

$$\begin{aligned} 18{,}7492x_1 + 6{,}0832x_2 - 4{,}8742x_3 &= 18{,}4666 \\ 6{,}0832x_1 + 12{,}3664x_2 + 2{,}4326x_3 &= 16{,}4098 \\ -4{,}8742x_1 + 2{,}4326x_2 + 16{,}8858x_3 &= 7{,}8678 \end{aligned}$$

x1=
x2=
x3=

**Solução de sistema de equações lineares por métodos iterativos:**

- Método dos deslocamentos simultâneos (Jacobi);
- Métodos dos deslocamentos sucessivos (Gauss-Seidel);
- Método de relaxação;

**Método dos deslocamentos simultâneos:**
Sistemas de equações lineares podem também ser resolvidos por vários esquemas iterativos, nos quais, partindo-se de uma estimativa inicial da solução, faz-se refinamentos sucessivos utilizando uma fórmula iterativa até que duas soluções sucessivas difiram de um valor inferior a uma tolerância pré-estabelecida.
Admitindo a existência de um sistema de equações lineares, que em termos matriciais é escrita como:

$$[A_{(n,\,n)}]x = [b_n]$$

Ou na forma desenvolvida, como:

$$\begin{array}{lllllllll}
a_{11}\,x_1 & + & a_{12}\,x_2 & + & a_{13}\,x_3 & + & \ldots\ldots & + & a_{1n}\,x_n & = & b_1 \\
a_{21}\,x_1 & + & a_{22}\,x_2 & + & a_{23}\,x_3 & + & \ldots\ldots & + & a_{2n}\,x_n & = & b_2 \\
a_{31}\,x_1 & + & a_{32}\,x_2 & + & a_{33}\,x_3 & + & \ldots\ldots & + & a_{3n}\,x_n & = & b_3 \\
\ldots\ldots\ldots & \ldots & \ldots\ldots\ldots & \ldots & \ldots\ldots\ldots & \ldots & \ldots\ldots & \ldots & \ldots\ldots\ldots & \ldots & \ldots\ldots \\
\ldots\ldots\ldots & \ldots & \ldots\ldots\ldots & \ldots & \ldots\ldots\ldots & \ldots & \ldots\ldots & \ldots & \ldots\ldots\ldots & \ldots & \ldots\ldots \\
\ldots\ldots\ldots & \ldots & \ldots\ldots\ldots & \ldots & \ldots\ldots\ldots & \ldots & \ldots\ldots & \ldots & \ldots\ldots\ldots & \ldots & \ldots\ldots \\
 & & & & & & & & & & \\
a_{n1}\,x_1 & + & a_{n2}\,x_2 & + & a_{n3}\,x_3 & + & \ldots\ldots & + & a_{nn}\,x_n & = & b_n
\end{array}$$

Admitir que os coeficientes da diagonal $a_{ii}$ (i=1,n) não se anulem. Se esta condição não for a situação real, deve-se reorganizar as equações, para ter esta condição.

O sistema acima pode ser rescrito na forma:

$$x_1 = -\frac{1}{a_{11}}(a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n) + \frac{b_1}{a_{11}}$$

$$x_2 = -\frac{1}{a_{22}}(a_{21}x_1 + a_{23}x_3 + \ldots + a_{2n}x_n) + \frac{b_2}{a_{22}}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$x_n = -\frac{1}{a_{nn}}(a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{n,\,n-1}x_{n-1}) + \frac{b_n}{a_{nn}}$$

Que na forma matricial assume o formato

$$x = [B_n]x + c$$

Onde:

$$B = \begin{bmatrix}
0 & -\frac{a_{12}}{a_{11}} & -\frac{a_{13}}{a_{11}} & \ldots & -\frac{a_{1n}}{a_{11}} \\
-\frac{a_{21}}{a_{22}} & 0 & -\frac{a_{23}}{a_{22}} & \ldots & -\frac{a_{2n}}{a_{22}} \\
\ldots\ldots & \ldots & \ldots & \ldots & \ldots \\
-\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \ldots & -\frac{a_{n,\,n-1}}{a_{nn}} & 0
\end{bmatrix}$$

E os elementos de c, são:

$$c_i = \frac{b_i}{a_{ii}}$$

Para resolver este sistema é considerado que cada elemento inicial de x, ou seja, o conjunto $x^{(0)}$, sendo estes uma aproximação para o vetor solução, e assim, estes valores são introduzidos no segundo membro de cada equação. A solução das equações fornecerá então o conjunto de valores $x^{(1)}$ que sem dúvida, constitui uma melhor aproximação à solução do que $x^{(0)}$. Substituindo-se $x^{(1)}$ no segundo membro obtém-se uma outra aproximação agora chamada $x^{(2)}$, continuando desta maneira até que, após k iterações, os valores $x^{(k)}$ tenham convergido para a solução satisfatória do sistema, considerando-se para tal um número preestabelecido de algarismos significativos. O teste de erro que se deve usar para determinar a convergência é:

$$\frac{\left|x_i^{k+1} - x_i^k\right|}{\left|x_i^{k+1}\right|} < \varepsilon$$

Este método é conhecido como iteração linear, também conhecido como ***método dos deslocamentos simultâneos.***

Resumindo:
Para resolver um sistema de equações lineares por este método fazer:

- Escrever o sistema na forma *x=Bx+c;*
- Escolher um vetor aproximação inicial $x^{(0)}$ arbitrário. Se não estiver disponível uma alternativa melhor, fazer $x_i^{(0)}=0$ para todos os valores *i*;
- Gerar aproximações sucessivas $x^{(k)}$ a partir da iteração $x^{(k+1)}=Bx^{(k)}+c$ k=1,m;
- Continuar até que seja satisfeita uma das duas condições:

$$\max_{1\le i\le n} \frac{\left|x_i^{(k+1)} - x_i^k\right|}{\left|x_i^{k+1}\right|} < \varepsilon \qquad \text{ou} \qquad k>m$$

Escolher sempre o valor máximo dos valores ***x*** na iteração ***i***

Exemplo numérico:
Seja o sistema:

$$\begin{aligned} 10x_1 + x_2 + x_3 &= 12 \\ x_1 + 10x_2 + x_3 &= 12 \\ x_1 + x_2 + 10x_3 &= 12 \end{aligned}$$

Cuja solução verdadeira é o conjunto: **[1, 1, 1].** Para resolver por iteração, rescreve-se o sistema, ficando assim:

$$\begin{aligned} x_1 &= -0{,}1x_2 - 0{,}1x_3 + 1{,}2 \\ x_2 &= -0{,}1x_1 - 0{,}1x_3 + 1{,}2 \\ x_3 &= -0{,}1x_1 - 0{,}1x_2 + 1{,}2 \end{aligned}$$

Iniciando o procedimento anteriormente descrito com o conjunto solução estimada $x_1 = x_2 = x_3 = 0$ e processando os devidos cálculos têm-se os valores conforme a tabela abaixo:

| *k* | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 1,2 | 1,2 | 1,2 |
| 2 | 0,96 | 0,96 | 0,96 |
| 3 | 1,108 | 1,108 | 1,108 |
| 4 | 0,9792 | 0,9792 | 0,9792 |
| 5 | 1,00416 | 1,00416 | 1,00416 |
| 6 | 0,999168 | 0,999168 | 0,999168 |

Nota-se que a cada iteração, há um ganho de um algarismo significativo exato.
*Observações: são executados os cálculos em todas as equações existentes com os valores de x atribuídos. Só então se reinicia nova iteração.*

**Método dos deslocamentos sucessivos (Método de Gauss-Seidel):**
Uma simples modificação da iteração linear conduz, algumas vezes, a uma convergência consideravelmente mais rápida. O processo consiste em calcular o valor de $x^{(0)}_1$ na primeira equação e já utilizar este valor na segunda equação e desta obter o valor de $x^{(0)}_2$ que será logo utilizado na terceira equação e assim sucessivamente. Isto eqüivale a usar um componente aperfeiçoado tão logo o mesmo esteja disponível. A iteração linear quando modificada desta maneira é denominada de ***método dos deslocamentos sucessivos***.

O mesmo exemplo anterior quando resolvido por este método produz os seguintes resultados

| $k$ | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 1,2 | 1,08 | 0,972 |
| 2 | 0,9948 | 1,0033 | 1,00019 |
| 3 | 0,99965 | 1,000016 | 1,000033 |

*Observa-se claramente que sucessivas iterações estão convergindo muito mais rapidamente usando este método. Ao final de três iterações, a precisão é melhor do que a obtida pelo método anterior obtida após 6 iterações.*

Os processos iterativos são usualmente aplicados para grandes sistemas lineares, nos quais a matriz dos coeficientes é ***esparsa***. Para matrizes esparsas é pequeno o número de elementos não nulos e, portanto, o número de operações a serem realizadas por iterações é, também, pequeno. Além disso, os métodos iterativos são menos vulneráveis ao crescimento do erro de arredondamento. O único arredondamento que ocorre é o gerado numa iteração isolada. Por outro lado, ***os métodos iterativos não convergem na totalidade das vezes***, sendo que, no caso em que ocorra a convergência, podem ainda requerer um número proibitivamente grande de iterações. Quando isto acontece, é necessário introduzir alguma técnica para acelerar a convergência. Um método muito usado consiste em substituir o valor $x^{(k)}_i$ pelo valor obtido pela transformação abaixo. Considere que $x^{(k)}$, $x^{(k+1)}$ e $x^{(k+2)}$ sejam três valores obtidos sucessivamente por qualquer método interativo. Assim $x^{(*)}$ fica:

$$x_i^{(*)} = \frac{\begin{vmatrix} x_i^k & x_i^{k+1} \\ x_i^{k+1} & x_i^{k+2} \end{vmatrix}}{x_i^k - x_i^{k+1} + x_i^{k+2}} \qquad \text{i=1,2,3.....n}$$

Na prática, é costume iterar um número fixo de vezes e, então, aplicar a aceleração. Contudo, pode as vezes, acontecer de se obter resultados menos precisos, de modo que este método não pode ser usado indiscriminadamente. ***Um outro método para tentar a convergência consiste basicamente em permutar o valor não convergente pela média aritmética dos dois últimos valores obtidos.***

**Convergência em métodos interativos:**
A cada iteração do algoritmo de Gauss-Seidel, espera-se obter valores das incógnitas mais próximos da solução como no exemplo acima mostrado. Quando isto ocorre, diz-se que o processo iterativo ***converge***. Há casos, entretanto, em que isto não acontece, quando então, o método, *não levando à solução, não resolve o problema dado.* Neste caso, se diz que o processo ***diverge***. Em todo método iterativo é importante obter-se critérios de convergência que possam servir para, *antes da solução de um problema, prever se ele convergirá ou não.*
Para o algoritmo de Gauss-Seidel existe uma condição suficiente de convergência muito simples de aplicar, que consiste em verificar se em cada fila da matriz de coeficientes, *o valor absoluto do elemento da diagonal for maior que a soma dos valores absolutos de todos os outros elementos*. Quando isto acontece se diz que a matriz dos coeficientes é ***diagonalmente dominante*** e neste caso ***sempre convergirá***.
Assim, antes da aplicação do algoritmo de Gauss-Seidel, o sistema de equações deve ser rearranjado de maneira que fiquem na diagonal os elementos de maior valores absolutos. Aí, o teste de convergência pode ser aplicado. *Observe que a condição dada é suficiente,* mas não necessária, ou seja, pode ser que sistemas que não a satisfazem tenham solução convergente. Entretanto a convergência será muito mais rápida se a condição for cumprida.

*É importante também saber que a convergência do algoritmo de Gauss-Seidel não depende da iniciativa inicial. Se o método converge para um determinado sistema, ele convergirá para qualquer que seja a estimativa inicial, onde é sempre comum adotar-se valores zero para todas as variáveis.*

**Sistemas esparsos:**
*Sistemas de equações lineares que possuem muitos coeficientes nulos são ditos* ***esparsos*** e as matrizes de seus coeficientes são também chamadas *matrizes esparsas* É muito usual que sistemas de ordens elevadas sejam esparsos e é na solução de tais sistemas que o método de Gauss-Seidel apresenta a maior vantagem, pois, como ele não altera os coeficientes, é feita máxima utilização da existência de coeficientes nulos na redução do número de cálculos. Para sistemas esparsos, ao contrário, os métodos de eliminação, apesar de também terem reduzido o número de eliminações a fazer, durante sua execução substituem vários coeficientes nulos por coeficientes não nulos, complicando desnecessariamente a solução.

Exemplo numérico:
Seja o sistema:

$$\left[\begin{array}{ccccccc} \frac{17}{60} & -\frac{1}{20} & -\frac{1}{30} & 0 & 0 & 0 & 10 \\ -\frac{1}{20} & \frac{17}{100} & 0 & -\frac{1}{50} & 0 & 0 & 0 \\ -\frac{1}{30} & 0 & \frac{17}{120} & 0 & -\frac{1}{40} & 0 & 0 \\ 0 & -\frac{1}{50} & 0 & \frac{23}{150} & 0 & -\frac{1}{20} & 0 \\ 0 & 0 & -\frac{1}{40} & 0 & \frac{19}{120} & -\frac{1}{30} & 0 \\ 0 & 0 & 0 & -\frac{1}{20} & -\frac{1}{30} & \frac{17}{60} & 8 \end{array}\right]$$

*Este sistema é diagonalmente dominante e ao mesmo tempo, esparso, sendo portanto, vantajoso o uso do método de Gauss-Seidel. Usando oito casas decimais o sistema na forma decimal e na forma iterativa fica:*

$x_1 = 35{,}29411765 + 0{,}17647059x_2 + 0{,}11764706x_3$
$x_2 = 0{,}29411765x_1 + 0{,}11764706x_4$
$x_3 = 0{,}23529412x_1 + 0{,}17647059x_5$
$x_4 = 0{,}13043478x_2 + 0{,}32608696x_6$
$x_5 = 0{,}15789474x_3 + 0{,}21052632x_6$
$x_6 = 28{,}23529412 + 0{,}17647059x_4 + 0{,}11764706x_5$

os resultados das iterações

| it. | x1 | x2 | x3 | x4 | x5 | x6 |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 35,29411765 | 10,38062294 | 8,30449835 | 1,35399427 | 1,31123661 | 28,62849742 |
| 2 | 38,10299213 | 11,36605595 | 9,20435017 | 10,81790870 | 7,48037069 | 31,02438047 |
| 3 | 38,38275699 | 12,56174144 | 10,35130246 | 11,75513389 | 8,16586486 | 31,27041952 |
| 4 | 38,72869587 | 12,77374996 | 10,55366940 | 11,86301731 | 8,24961523 | 31,29931076 |
| 5 | 38,78991702 | 12,80444835 | 10,58285386 | 11,87644250 | 8,26030567 | 31,30293762 |
| 6 | 38,79876785 | 12,80863096 | 10,58682295 | 11,87817073 | 8,26169592 | 31,30340616 |
| 7 | 38,79997291 | 12,80918872 | 10,58735184 | 11,87839626 | 8,26187807 | 31,30346738 |
| 8 | 38,80013356 | 12,80926250 | 10,58742178 | 11,87842596 | 8,26190200 | 31,30347542 |
| 9 | 38,80015481 | 12,80927223 | 10,58743100 | 11,87842975 | 8,26190515 | 31,30347648 |
| 10 | 38,80015761 | 12,80927353 | 10,58743222 | 11,87843025 | 8,26190557 | 31,30347661 |

*Observe que os valores irão sempre convergindo mais para a solução verdadeira do sistema.*

**Programação:**
Para programar a resolução de um sistema de equações pelo método de Gauss-Seidel, não é necessário rescrever a matriz na forma de equações. Basta observar que uma vez montada a matriz expandida, inicia-se o processo definindo-se outra matriz do tipo matriz coluna, que irá conter os valores estimados para as variáveis do sistema, as quais podem ser iniciadas geralmente com o valor *zero* e fixando como pivô o primeiro elemento da diagonal principal. Este elemento irá ser usado para dividir toda a sua linha, sendo que os elementos da matriz coeficiente, exceto pelo pivô, terão seu sinal trocado e somados entre si, incluindo o valor ***b***, o qual também é dividido pelo elemento do pivô, exceto pelo sinal, que não é trocado. Fazer este processo com todas as linhas do sistema a cada iteração, e a cada linha usar o elemento da diagonal principal correspondente como pivô. Usar um controle de convergência para sair do "***loop"*** de iteração sucessiva, sempre no último pivô. Deve-se prever um número limite de verificação de convergência para, no caso de não haver convergência, fazer alguma correção, ou parar o processamento. Pode-se como prevenção, se processar a verificação se a matriz dos coeficientes é do tipo diagonal dominante, emitindo um aviso para parar ou continuar. De qualquer maneira, a matriz dos coeficientes nunca é alterada. Os únicos valores que mudam na memória do computador são os valores das variáveis.

**Exemplo de aplicação técnica: Cálculo de esforços em treliça:**
(inserir aqui o desenho da treliça com os dados)

Seja a treliça conforme mostrada na figura abaixo, onde os triângulos são todos equiláteros. Admitindo uma carga concentrada de 10 toneladas atuando na vertical para baixo na junta 3 e denotando os esforços de tração positivos, os balanços verticais e horizontais nas juntas dão as equações a seguir, onde $f_i$ indica a tração em cada barra da treliça

Esforço horizontal na junta 2: -f1.sen 30$^o$ + f3.sen 30$^o$ + f4 = 0
Esforço vertical na junta 2: -f1.cos 30$^o$ + f3.cos 30$^o$ = 0
Esforço horizontal na junta 3: -f2 – f3.sen 30$^o$ + f5.sen 30$^o$ + f6 = 0
Esforço vertical na junta 3: -f3.cos 30$^o$ + f5.cos 30$^o$ = 10
Esforço horizontal na junta 4: -f4 – f5.sen 30$^o$ + f7.sen 30$^o$ = 0
Esforço vertical na junta 4: -f5.cos 30$^o$ + f7.cos 30$^o$ = 0
Esforço horizontal na junta 5: -f6 – f7.sen 30$^o$ = 0

Que em forma matricial fica:

$$\begin{bmatrix} -1/2 & 0 & 1/2 & 1 & 0 & 0 & 0 \\ \sqrt{3/2} & 0 & \sqrt{3/2} & 0 & 0 & 0 & 0 \\ 0 & -1 & -1/2 & 0 & 1/2 & 1 & 0 \\ 0 & 0 & \sqrt{3/2} & 0 & \sqrt{3/2} & 0 & 0 \\ 0 & 0 & 0 & -1 & -1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & \sqrt{3/2} & 0 & \sqrt{3/2} \\ 0 & 0 & 0 & 0 & 0 & -1 & -1/2 \end{bmatrix} \times \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \\ f_7 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 10 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Este sistema é esparso, mas não é diagonalmente dominante (tente rearranjá-lo para que seja!) e a solução então é melhor pelo sistema de Gauss-Jordan.

Respostas:
***f1= -5,774 ton***
***f2 = 2,887 ton***
***f3 = 5,774 ton***
***f4 = -5,774 ton***
***f5 = 5,774 ton***
***f6 = 2,887 ton***
***f7 = - 5774 ton***

**Método de relaxação:**
Para o método de relaxação estima-se o valor inicial para as incógnitas $\boldsymbol{x}$ e calculam-se então os resíduos resultantes da substituição dos valores na matriz transformada efetuando-se em seguida o teste de convergência selecionando o maior resíduo entre todos e comparando-o com a tolerância exigida. Se não satisfizer a condição da tolerância, alterar o valor da incógnita $\boldsymbol{x_i}$ e continuar o processo. Para melhor entender o método, considerar o sistema genérico abaixo, de ordem 3. (pode ser de qualquer ordem):

$$\begin{aligned} a_{11}\,x_1 &+ a_{12}\,x_2 + a_{13}\,x_3 = b_1 \\ a_{21}\,x_1 &+ a_{22}\,x_2 + a_{23}\,x_3 = b_2 \\ a_{31}\,x_1 &+ a_{32}\,x_2 + a_{33}\,x_3 = b_3 \end{aligned}$$

Sistema que pode ser rescrito para

$$\begin{aligned} \mathbf{x_1} &+ c_{12}\,x_2 + c_{13}\,x_3 - z1 = 0 \\ c_{21}\,x_1 &+ \mathbf{x_2} + c_{23}\,x_3 - z2 = 0 \\ c_{31}\,x_1 &+ c_{32}\,x_2 + \mathbf{x_3} - z3 = 0 \end{aligned}$$

Os termos $c_{ij}$ são modificados pela operação

$$c_{ij} = \frac{a_{ij}}{a_{ii}} \qquad \text{Para i = 1 até n}$$

E $b_{ij}$, por

$$z_i = \frac{b_i}{a_{ii}} \qquad \text{Para i = 1 até n}$$

Ao resolver o sistema modificado acima, primeiramente é necessário atribuir valores para as incógnitas $x_1$, $x_2$ e $x_3$.

- Atribui-se qualquer valor para cada um dos $\boldsymbol{x_i}$, normalmente escolhe-se 1 para todos;
- Substituem-se os valores atribuídos no sistema acima e resolve-se o mesmo anotando o resultado em cada equação. Estes resultados são chamados de resíduos, representados por $\boldsymbol{R_i}$;

$$\begin{aligned} \mathbf{x_1} &+ c_{12}\,x_2 + c_{13}\,x_3 - z1 = R_1 \\ c_{21}\,x_1 &+ \mathbf{x_2} + c_{23}\,x_3 - z2 = R_2 \\ c_{31}\,x_1 &+ c_{32}\,x_2 + \mathbf{x_3} - z3 = R_3 \end{aligned}$$

- Após concluir os cálculos em todas as equações, escolher entre os $\boldsymbol{i}$'s resíduos o de maior valor e comparar o mesmo com a tolerância estabelecida no início, onde:

$$R_{\max}^{k} \le \xi \qquad \text{(precisão mínima)}$$

- Se o resíduo máximo não satisfizer a condição, calcular a variação na incógnita $\boldsymbol{x_i}$ necessária para reduzir o resíduo $\boldsymbol{R_i}$ a zero, segundo o esquema abaixo:

$$\Delta x_i^k = -R_i^k$$
$$x_i^{k+1} = x_i^k + \Delta x_i^k = x_i^k - R_i^k$$

- Calcular os demais resíduos restantes pela modificação:

$$R_j^{k+1} = R_j^k + c_{ji} \times \Delta x_i^k \qquad \text{Para todo j}\neq\text{i}$$

- Com relação à convergência, vale os mesmos princípios aplicados aos métodos iterativos já vistos, ou seja, se a matriz for diagonalmente dominante ela sempre convergirá. Nos outros casos, se não convergir deve-se procurar resolver o sistema por um método direto.

Exemplo:
Seja o sistema abaixo:

$$\begin{array}{rcrcrcr} 0{,}627\,x_1 & + & 0{,}193\,x_2 & + & 0{,}010\,x_3 & = & 1 \\ 0{,}193\,x_1 & + & 0{,}484\,x_2 & + & 0{,}171\,x_3 & = & 1 \\ 0{,}010\,x_1 & + & 0{,}171\,x_2 & + & 0696\,x_3 & = & 1 \end{array}$$

Reescrevendo o sistema, fica:

$$\begin{array}{rcrcrcrcr} x_1 & + & 0{,}3078\,x_2 & + & 0{,}0159\,x_3 & -1{,}5949 & = & 0 \\ 0{,}3988\,x_1 & + & x_2 & + & 0{,}3533\,x_3 & -2{,}0661 & = & 0 \\ 0{,}0144\,x_1 & + & 0{,}2457\,x_2 & + & x_3 & -1{,}4368 & = & 0 \end{array}$$

Atribuindo valores para x: $x_1 = x_2 = x_3 =$ **1** e fazendo $\xi = 0{,}01$ e calculando-se os resíduos:

$$\begin{array}{rcrcrcrcr} 1 & + & 0{,}3078 \times 1 & + & 0{,}0159 \times 1 & -1{,}5949 & = & -0{,}2711 \\ 0{,}3988 \times 1 & + & 1 & + & 0{,}3533 \times 1 & -2{,}0661 & = & -0{,}3140 \\ 0{,}0144 \times 1 & + & 0{,}2457 \times 1 & + & 1 & -1{,}4368 & = & -0{,}1767 \end{array}$$

Montando uma tabela, conforme mostrado abaixo, vem:

| *k* | $x_1$ | $x_2$ | $x_3$ | $R_1$ | $R_2$ | $R_3$ |
|---|---|---|---|---|---|---|
| **0:** | 1 | 1 | 1 | -0,2711 | <u>-0,3140</u> | -0,1767 |

$$x_2^1 = x_1^0 - R_2^0 = 1 - (-0{,}3140) = 1{,}3140$$
$$R_1^1 = -0{,}2711 + 0{,}3078 \times 0{,}3140 = -0{,}1745$$
$$R_3^1 = -0{,}1767 + 0{,}2457 \times 0{,}3140 = 0{,}0996$$

| | | | | | | |
|---|---|---|---|---|---|---|
| **1:** | 1 | **1,3130** | 1 | <u>-0,1745</u> | 0 | -0,0996 |

$$x_1^2 = x_1^1 - R_1^1 = 1 - (-0{,}1745) = 1{,}1745$$
$$R_2^2 = 0 + 0{,}3988 \times 0{,}1745 = 0{,}0696$$
$$R_3^2 = -0{,}0969 + 0{,}0144 \times 0{,}1745 = -0{,}0971$$

| | | | | | | |
|---|---|---|---|---|---|---|
| **2:** | **1,1745** | 1,3140 | 1 | 0 | 0,0696 | <u>-0,0971</u> |

$$x_3^3 = x_3^2 - R_3^2 = 1 - (-0{,}0971) = 1{,}0971$$
$$R_1^3 = 0 + 0{,}0159 \times 0{,}0971 = 0{,}0015$$
$$R_2^3 = -0{,}0696 + 0{,}3533 \times 0{,}0971 = 0{,}1039$$

| | | | | | | |
|---|---|---|---|---|---|---|
| **3:** | 1,1745 | 1,3140 | **1,0971** | 0,0015 | <u>0,1039</u> | 0 |

$$x_2^4 = x_2^3 - R_2^3 = 1{,}3140 - (0{,}1030) = 1{,}2102$$
$$R_1^4 = 0{,}0015 + 0{,}3078 \times (-0{,}1030) = -0{,}0304$$
$$R_3^4 = 0 + 0{,}2457 \times (-0{,}1030) = -0{,}0255$$

| | | | | | | |
|---|---|---|---|---|---|---|
| **4:** | 1,1745 | **1,2102** | 1,0971 | -<u>0,0304</u> | 0 | -0,0255 |
| **5:** | **1,2049** | 1,2102 | 1,0971 | 0 | 0,0121 | <u>-0,0251</u> |
| **6:** | 1,2049 | 1,2102 | **1,1221** | 0,0004 | <u>0,0210</u> | 0 |
| **7:** | 1,2049 | **1,1892** | 1,1221 | <u>-0,0061</u> | 0 | -0,0052 |
| **8:** | **1,2109** | 1,1892 | 1,1221 | 0 | 0,0024 | <u>-0,0051</u> |
| **9:** | 1,2109 | 1,1892 | **1,1272** | 0,0001 | <u>0,0042</u> | 0 |
| **10:** | 1,2109 | **1,1850** | 1,1272 | <u>-0,0012</u> | 0 | -0,0010 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **11:** | **1,2122** | 1,1850 | 1,1272 | 0 | 0,0005 | -0,0010 |
| **12:** | 1,2122 | 1,1850 | **1,1282** | 0 | 0,0008 | 0 |
| **13:** | 1,2122 | **1,1841** | 1,1282 | -0,0002 | 0 | -0,0002 |
| **14:** | **1,2124** | 1,1841 | 1,1282 | 0 | 0,0001 | -0,0002 |
| **15:** | 1,2124 | 1,1841 | **1,1284** | 0,0000 | 0,0002 | 0 |
| **16:** | 1,2124 | **1,1840** | 1,1284 | 0,0000 | 0 | 0,0000 |
| **17:** | **1,2125** | 1,1840 | 1,1284 | 0 | 0,0000 | 0,0000 |
| **18** | 1,2125 | 1,1840 | **1,1285** | 0,0000 | 0,0000 | 0 |

Percebe-se claramente que a cada iteração o sistema converge cada vez mais refinando os valores da verdadeira solução, ou seja, *para os valores reais de* $\boldsymbol{x_i}$ *o resíduo deve ser exatamente igual a zero.*

Num outro exemplo como o sistema abaixo:

| | | |
|---|---|---|
| 1 | 2 | 3 |
| 4 | 5 | 6 |

Cuja solução verdadeira é $x_1 = -1$ e $x_2 = 2$

Aplicando o método de relaxação tem-se:

A matriz modificada fica:

| | | |
|---|---|---|
| 1.0000 | 2.0000 | -3.0000 |
| .8000 | 1.0000 | -1.2000 |

| *k* | $x_1$ | $x_2$ | $R_1$ | $R_2$ |
|---|---|---|---|---|
| 0 | 1.0000 | 1.0000 | .0000 | .6000 |
| 1 | 1.0000 | .4000 | -1.2000 | .0000 |
| 2 | 2.2000 | .4000 | .0000 | .9600 |
| 3 | 2.2000 | -.5600 | -1.9200 | .0000 |
| 4 | 4.1200 | -.5600 | .0000 | 1.5360 |
| 5 | 4.1200 | -2.0960 | -3.0720 | .0000 |
| 6 | 7.1920 | -2.0960 | .0000 | 2.4576 |
| 7 | 7.1920 | -4.5536 | -4.9152 | .0000 |

*Percebe-se que a mesma não converge. Pode-se tentar mudar a ordem das equações, como abaixo:*

Fazendo:

| | | |
|---|---|---|
| 4 | 5 | 6 |
| 1 | 2 | 3 |

A matriz modificada fica:

| | | |
|---|---|---|
| 1.0000 | 1.2500 | -1.5000 |
| .5000 | 1.0000 | -1.5000 |

| *k* | $x_1$ | $x_2$ | $R_1$ | $R_2$ |
|---|---|---|---|---|
| 0 | 1.0000 | 1.0000 | .7500 | .0000 |
| 1 | .2500 | 1.0000 | .0000 | -.3750 |
| 2 | .2500 | 1.3750 | .4688 | .0000 |
| 3 | -.2188 | 1.3750 | .0000 | -.2344 |
| 4 | -.2188 | 1.6094 | .2930 | .0000 |
| 5 | -.5117 | 1.6094 | .0000 | -.1465 |
| 6 | -.5117 | 1.7559 | .1831 | .0000 |
| 7 | -.6948 | 1.7559 | .0000 | -.0916 |
| 8 | -.6948 | 1.8474 | .1144 | .0000 |

| | | | | |
|---|---|---|---|---|
| 9 | -.8093 | 1.8474 | .0000 | -.0572 |
| 10 | -.8093 | 1.9046 | .0715 | .0000 |
| 11 | -.8808 | 1.9046 | .0000 | -.0358 |
| 12 | -.8808 | 1.9404 | .0447 | .0000 |
| 13 | -.9255 | 1.9404 | .0000 | -.0224 |
| 14 | -.9255 | 1.9627 | .0279 | .0000 |
| 15 | -.9534 | 1.9627 | .0000 | -.0140 |
| 16 | -.9534 | 1.9767 | .0175 | .0000 |
| 17 | -.9709 | 1.9767 | .0000 | -.0087 |
| 18 | -.9709 | 1.9854 | .0109 | .0000 |
| 19 | -.9818 | 1.9854 | .0000 | -.0055 |
| 20 | -.9818 | 1.9909 | .0068 | .0000 |
| 21 | -.9886 | 1.9909 | .0000 | -.0034 |

*Percebe-se agora que os valores de* ***x*** *convergem, porém, devido ao fato da matriz não ser diagonalmente dominante, a convergência do cálculo é mais lenta, necessitando aproximadamente 50 iterações para se conseguir uma precisão de 0,001*.

## Métodos gerais para resolução de equações não lineares de uma única variável

1- *Por substituição sucessiva*. Seja uma equação não linear f(x) = 0 para ser resolvida. Se a mesma for rescrita como f(x) = x, então um esquema iterativo pode ser empregado obtendo-se $x_{k+1} = f(x_k)$. O valor inicial de x deve ser estimado, valor este que pode ser eventualmente obtido de forma aproximada graficamente ou simplesmente escolher-se qualquer valor. A convergência do cálculo depende da escolha inicial do valor de x e da própria função f(x). Não existe uma regra geral para garantir a convergência, entretanto, se a é a raiz de f(x) = 0, a condição necessária para a convergência é que $|f'(x)| < 1$ (*a derivada da função no intervalo, tomada em módulo, testado deve ser menor que 1*). Este processo é tido como um processo de primeira ordem, porque o erro em $x_{k+1}$ é sempre proporcional à primeira potência do erro em $x_k$.

Exemplo: seja a $f(x) = x^3 - x - 1 = 0$. A solução gráfica para esta equação mostra uma raiz real em aproximadamente 1,3. Ao escrever esta equação na forma x = f(x), tem-se:

Opção 1: $x = x^3 - 1$ ou

Opção 2: $x^3 = 1 + x$ ou $x = \sqrt[3]{1+x}$

Opção 3: $x.x^2 - x - 1 = 0$ $x(x^2 - 1) = 1$ $x = \frac{1}{(x^2 - 1)}$

Efetuando-se os cálculos em cada uma das formas encontradas, vem:

| k | $x = x^3 - 1$ | $x = \sqrt[3]{1+x}$ | $x = \frac{1}{(x^2 - 1)}$ |
|---|---|---|---|
| 0 | 1,3 | 1,3 | 1,3 |
| 1 | 1,4493 | 1,197 | 1,32 |
| 2 | 0,9087 | 0,715 | 1,3238 |
| 3 | -5,737 | -0,6345 | 1,3247 |
| 4 | ..... | ..... | 1,3247 |

Conferindo as derivadas, observa-se:
Na primeira forma, a derivada da equação, no ponto 1,3, fica em 3x2= 5,07
Na segunda forma, a derivada da equação, no ponto 1,3, fica em 5,46 e
Na terceira forma, a derivada, no ponto 1,3, fica menor que 1. Assim observa-se também que apenas a terceira equação convergiu. Para forçar a convergência pode-se adotar alguns melhoramentos como, por exemplo, o método de Wegstein. (pesquisar literatura especializada ou o manual do engenheiro químico).
Observa-se também que a maior dificuldade deste método é encontrar a função que satisfaça à condição de convergência.

2- *Método de bisseção*. Baseia-se exclusivamente no fato que entre dois valores para a f(x) com sinais trocados devem pelo menos existir uma raiz ou conforme diz a regra de Descartes, um número impar de raízes. Uma vez encontrados dois valores de x onde f(x) tenham sinais trocados, o próximo valor de ***x*** se obtém através da média aritmética entre estes dois valores. O novo valor de x irá determinar um novo valor para f(x). O sinal da função neste ponto deve ser comparado com os dois pontos anteriores e a nova média deverá ser obtida entre dois valores de ***x***, onde, um deles é último ***x*** obtido e o outro é um dos dois anteriores ao último. Aquele que dá a f(x) o sinal contrário a f(x) do último ***x***. De aplicação bastante simples, este método não necessita das derivadas, mas tem a desvantagem de ter uma convergência bastante lenta.

Exemplo: Seja a $f(x) = x^3 - x - 1 = 0$. A solução gráfica para esta equação mostra uma raiz real em aproximadamente 1,3. Escolhe-se um valor inicial para começar os cálculos. Normalmente deve-se procurar escolher um valor o mais próximo possível, o qual pode ser obtido aproximadamente de maneira gráfica.

No caso, escolher-se-á o valor 1.

| K | x | f(x) |
|---|---|---|
| 0 | 1 | $1^3$-1-1=-1 |
| 1 | 2 | $2^3$-2-1=5 |

Observa-se aqui que houve inversão no sinal de f(x), indicando que existe pelo menos uma raiz entre x=1 e x=2.

| | | | |
|---|---|---|---|
| 2 | $(x_0+x_1)/2=$ | (1+2)/2=1,5 | $(1,5)^3$-1,5-1=0,875 |
| 3 | $(x_0+x_2)/2=$ | (1+1,5)/2=1,25 | $(1,25)^3$-1,25-1=-0296875 |

Houve inversão de sinal em f(x), portanto:

| | | | |
|---|---|---|---|
| 4 | $(x_2+x_3)/2=$ | (1,5+1.25)/2=1,375 | $(1.375)^3$-1.375-1=0,224609 |

5 continuar....

6

3- *Método da falsa posição*. Este método também é conhecido como método de Illinois ou método pegassus e também parte do princípio de se encontrar dois pontos onde f(x) apresenta sinais contrários. Então a linha secante que une os pontos ($x_k$, $f(x_k)$ e ($x_{k+n}$, $f(x_{k+n})$) irá cortar a reta das abcissas determinando um novo ponto para ***x***. Este método é um melhoramento do método das cordas. Não exige o conhecimento das derivadas como o método das cordas exige e tem uma convergência só superada pelo método de Newton-Raphson. O valor do novo ponto, uma vez conhecidos dois pontos $x_{k-1}$ e $x_k$, onde f(x) tenha sinais contrários, é obtido pela relação):

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

Ou seja, considerando o triângulo formado entre os dois valores de x e os dois valores da função, a correção de $x_{k+1}$ é diretamente proporcional à variação da função, através da regra de três abaixo.

| | |
|---|---|
| $x_k - x_{k-1}$ está para | $f(x_k) - f(x_{k-1})$ |
| $x_{k+1} - x_k$ | $f(x_k) - 0$ |

Exemplo: Seja a $f(x) = x^3 - x - 1 = 0$. A solução gráfica para esta equação mostra uma raiz real em aproximadamente 1,3. Ao escrever esta equação na forma x = f(x), tem-se:

Escolhe-se um valor inicial para começar os cálculos. Normalmente deve-se procurar escolher um valor o mais próximo possível, o qual pode ser obtido aproximadamente de maneira gráfica.

No caso, escolher-se-á o valor 1.

| k | x | f(x) | |
|---|---|---|---|
| 0 | 1 | -1 | |
| 1 | 2 | 5 | |
| 2 | 2-[(2-1)/(5-(-1)]x5=1,1667 | -0,5786 | |
| 3 | 1,2531 | | (continuar o cálculo |
| 4 | | | usando calculadora) |
| 5 | | | |

4- *Método de Newton-Raphson*. Dada uma equação f(x) = 0, o cálculo iterativo é obtido pela expressão:

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

Onde:
$f(x_k)$ = valor da expressão não linear a ser resolvida, no intervalo k;
$f'(x_k)$ = valor da derivada da expressão no intervalo k;

Este método é muito mais rápido que o método da bisseção, mas pode não levar à raiz. Esta é uma característica de muitos métodos iterativos. Eles podem, a partir de uma estimativa inicial, afastar-se da solução. Quando isto acontece, diz-se que o método diverge. Quando o contrário acontece, então converge. O método de Newton-Raphson pode não ser seguro devido ao uso da tangente. O método em si, consiste em, a partir de uma estimativa inicial, determinar a tangente à curva da função neste ponto e adotar como próxima estimativa a interseção desta tangente com o eixo das abcissas, ponto esse calculado pela fórmula apresentada acima.

Exemplo: Seja a $f(x) = x^3 - x - 1 = 0$. A solução gráfica para esta equação mostra uma raiz real em aproximadamente 1,3. Escolhe-se um valor inicial para começar os cálculos. Normalmente deve-se procurar escolher um valor o mais próximo possível, o qual pode ser obtido aproximadamente de maneira gráfica. No caso foi escolhido o valor 2.

A derivada da função acima fica: $f'(x) = 3x^2 - 1$

| K | x | f(x) | f´(x) |
|---|---|---|---|
| 0 | 2 | -5 | $3x2^2 - 1 = 11$ |
| 1 | 2 – (-5)/11= 1,5454 | 1,1454 | 6,1648 |
| 2 | 1,3596 | 0,1536 | 4,5455 |
| 3 | 1,3258 | 0,0046 | 4,2732 |
| 4 | 1,3247 | - | - |

Percebe-se que já convergiu para a raiz com a precisão de 0,0001.

**Exercício:**
Uma loja oferece dois planos de financiamento para a compra de um produto cujo preço à vista é de R$ 16200,00, a saber:

Plano A: entrada de R$ 2200,00 + 9 prestações de R$ 2652,52

Plano B: entrada de R$ 2200,00 + 12 prestações de R$ 2152,27

Qual é o melhor dos dois planos?

A expressão que rege o pagamento de prestações é:

$$\frac{VP}{PM} = \frac{1-(1+i)^{-n}}{i} = \frac{(1+i)^n - 1}{i.(1+i)^n}$$

Onde:
VP = valor a ser a ser financiado
PM = valor da prestação mensal
I = taxa de juros do empréstimo.

Obviamente, a melhor opção é aquela que oferece a menor taxa de juros i, portanto, deve-se aplicar um dos métodos propostos para determinar em cada uma das opções e determinar-se a taxa ***i*** correspondente.

5- *Método das raízes quadradas de Graeffe*. (*Método especial para polinômios)*
Considerar a equação polinomial de grau n

$P(x) = a_0.x^n + a_1.x^{n-1} + a_2.x^{n-2} + .....+ a_{n-1}.x + a_n = 0$

Considerar que esta equação tenha pelo menos uma raiz real. (este polinômio pode ter n raízes, as quais podem ser reais ou complexas). Se todos os coeficientes de P(x) forem inteiros então qualquer raiz racional, ou seja, na forma ***r/s***, onde r e s são inteiros, não tendo divisor comum, devem ser tais que *r* é sempre um divisor comum de $a_n$ e ***s*** é sempre um divisor de $a_0$. Qualquer polinômio com coeficientes racionais podem ser convertidos em coeficientes inteiros através de multiplicação dos mesmos pelo mínimo múltiplo comum entre todos os denominadores presentes no polinômio.

Exemplo: dada a equação $3x^4 - \frac{5}{3}x^2 + \frac{1}{5}x - 2 = 0$

Retirando-se os denominadores tem-se

$$45x^4 - 25x^2 + 3x - 30 = 0$$

Onde as únicas possíveis soluções racionais ***r/s*** são tais que ***r*** devem ter valores ±30, ±15, ±10, ±6, ±3, ±2 ou ±1 e ***s*** pode ter os valores ±45, ±15, ±9, ±5, ±3 ou ±1. As possíveis raízes podem ser formadas a partir dos possíveis quocientes, tendo fatores não comuns.
Outras propriedades podem ser usadas para estimar a quantidade possível de raízes reais, como por exemplo, a regra de Descartes, que diz que a quantidade de raízes reais positivas de um polinômio com coeficientes reais é tão igual quanto às trocas de sinal do polinômio ou menor deste por um número sempre inteiro e positivo. Outras propriedades mais podem ser usadas e para isto basta consultar a teoria sobre polinômios. Para encontrar numericamente as raízes de uma equação algébrica pode-se usar, além dos métodos já vistos, usar-se também o método das raízes quadradas de Graeffe. Seja a equação polinomial abaixo:

$$f(x) = a_0 + a_1 x^{p-1} + .... + a_{p-1}x + a_p = 0$$

Se as raízes são $r_1$, $r_2$,....., $r_p$, então se pode escrever

$$S_p = r_1^p \left( 1 + \frac{r_2^p}{r_1^p} + \frac{r_3^p}{r_1^p} + \frac{r_4^p}{r_1^p} + .... \right)$$

O método de Graeffe fornece uma maneira bastante eficiente para se obter $S_p$ da equação acima através da seqüência de equações ***tal que as raízes de cada equação são os quadrados das raízes das equações precedentes na seqüência***. Pela análise da equação acima se pode verificar que se $r_1$ for suficientemente muito maior que todas as demais raízes, o termo dentro do parêntesis fica próximo de 1, e obviamente o valor da raiz $r_1$ é aproximadamente o valor de $S_p$, ou generalizando:

$$\lim_{p->\infty} S_p^{1/p} = r_1$$

O processo iterativo consiste em aplicar a seguinte seqüência de fórmulas, cuja demonstração pode ser obtida em literatura especializada.

$$\frac{a_1}{a_0} = -\sum_{i=1}^{p} r_i \qquad \frac{a_1^1}{a_0^1} = -\sum_{i=1}^{p} r_i^2 \qquad \frac{a_1^2}{a_0^2} = -\sum_{i=1}^{p} r_i^4 \qquad \frac{a_1^p}{a_0^p} = -\sum_{i=1}^{p} r_i^{2p}$$

Se as raízes são todas distintas e $r_1$ é a maior delas em magnitude, então eventualmente...

$$r_1^{2p} \approx -\frac{a_1^p}{a_0^p}$$

E se $r_2$ for a próxima maior em magnitude, também...

$$r_2^{2p} \approx \frac{a_2^p}{a_1^p}$$

E generalizando vem

$$(-1)^n r_n^{2p} \approx \frac{a_n^p}{a_{n-1}^p}$$

Este procedimento pode ser facilmente generalizado para polinômios de qualquer grau e otimizado para incluir múltiplas raízes complexas. Os sinais das raízes obtidas são indeterminados, mas estes podem ser determinados testando ambas as possibilidades.

Exemplo: $f(x) = x^4 - 7x^3 + 9x^2 + 7x - 10 = 0$ , cujas raízes são 5, 2, 1 e –1.

Elevando a equação acima ao quadrado duas vezes, vem:

1ª vez:

$$a_{0(1)} = 1 \rightarrow a_0^2 = 1^2 = 1$$

$$a_{1(1)} = -7 \rightarrow (2a_0.a_2 - a_1^2) = 2\times1\times9 - (-7)^2 = -31$$

$$a_{2(1)} = 9 \rightarrow (2a_0.a_4 - 2a_1.a_3 + a_2^2) = 2\times1\times(-10) - 2\times(-7)\times7 + 9^2 = 159$$

$$a_{3(1)} = 7 \rightarrow (2a_2.a_4 - a_3^2) = 2\times9\times(-10) - 7^2 = -229$$

$$a_{4(1)} = -10 \rightarrow a_4^2 = (-10)^2 = 100$$

2ª vez: repete-se o mesmo cálculo usando agora os coeficientes obtidos no cálculo anterior.

$$a_{0(2)} = 1 \rightarrow a_0^2 = 1^2 = 1$$

$$a_{1(2)} = -31 \rightarrow (2a_0.a_2 - a_1^2) = 2\times1\times159 - (-31)^2 = -643$$

$$a_{2(2)} = 159 \rightarrow (2a_0.a_4 - 2a_1.a_3 + a_2^2) = 2\times1\times(100) - 2\times(-31)\times(-229) + 159^2 = 11283$$

$$a_{3(2)} = -229 \rightarrow (2a_2.a_4 - a_3^2) = 2\times159\times(100) - (-229)^2 = -20641$$

$$a_{4(2)} = 100 \rightarrow a_4^2 = (100)^2 = 10000$$

| Passo (p) | $a_0$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ |
|---|---|---|---|---|---|
| 0 | 1 | –7 | 9 | 7 | –10 |
| 1 | 1 | –31 | 159 | –229 | 100 |
| 2 | 1 | –643 | 11283 | –20641 | 10000 |

assim: $r_1^4 \approx -\frac{a_1^2}{a_0^2} = -\frac{-643}{1} = 643$ $\qquad r_1 = (643)^{1/4} = 5{,}04$

$$r_2^4 \approx -\frac{a_2^2}{a_1^2} = -\frac{11283}{-643} = 17{,}6 \qquad r_2 = (17{,}6)^{1/4} = 2.05$$

E assim sucessivamente até $r_4$

$$r_3^4 \approx -\frac{a_3^2}{a_2^2} = -\frac{-20641}{11283} = 1{,}83 \qquad r_3 = (1{,}83)^{1/4} = 1.16$$

$$r_4^4 \approx -\frac{a_4^2}{a_3^2} = -\frac{10000}{-20641} = 0{,}83 \qquad r_4 = (0{,}83)^{1/4} = 0{,}95$$

Após conseguir-se as raízes do polinômio, estas podem ser refinadas pelo método de Newton-Raphson ou outro qualquer.

Para elevar um polinômio ao quadrado procede-se da seguinte maneira:

Seja o polinômio:

$$P(x) = a_0.x^4 + x_1.x^3 + a_2.x^2 + a_3.x + a_4 = 0$$

Desmembram-se os termos pares dos impares, mandando-os para o outro lado da equação. Fica:

$$a_0.x^4 + a_2.x^2 + a_4 = -(x_1.x^3 + a_3.x)$$

Elevando ao quadrado ambos os lados, vem:

$$(a_0.x^4 + a_2.x^2 + a_4)^2 = -(x_1.x^3 + a_3.x)^2$$

Desenvolvendo o quadrado em ambos os termos, vêm:

$$a_0^2.x^8 + 2a_0a_2.x^6 + 2a_0a_4.x^4 + a_2^2.x^4 + a_2a_4.x^2 + a_4^2 = a_1^2.x^6 + 2a_1a_3.x^4 + a_3^2.x^2$$

Reagrupando os termos semelhantes:

$$a_0^2.x^8 + (2a_0a_2 - a_1^2)x^6 + (2a_0a_4 - 2a_1a_3 + a_2^2)x^4 + (a_2a_4 - a_3^2)x^2 + a_4^2 = 0$$

Fazendo a substituição da variável x por outra y, tal que $y=x^2$, vem:

$$a_0^2.y^4 + (2a_0a_2 - a_1^2.x^6)y^3 + (2a_0a_4 - 2a_1a_3 + a_2^2)y^2 + (a_2a_4 - a_3^2)y \ + a_4^2 = 0$$

Percebe-se que se retorna ao formato de uma equação algébrica do 4º grau, onde os novos coeficientes são agora os quadrados dos coeficientes anteriores e obviamente as raízes também foram elevadas ao quadrado. Este procedimento pode ser estendido a qualquer equação algébrica, desde o segundo grau até grau n. Desta forma se pode então obter uma forma generalizada para qualquer grau. Uma vez elevada ao quadrado, um número de vezes igual à metade do grau do polinômio, estes valores podem ser aplicados para se obter os valores aproximados das raízes, que após o refinamento seriam determinadas com a precisão estabelecida pela tolerância desejada.
Dentro deste princípio, fazendo estes cálculos para cada tipo de polinômio vêm as seguintes modificações:

Para equação do 2º grau:

$$a_0 \rightarrow a_0^2$$
$$a_1 \rightarrow 2a_0a_2 - a_1^2$$
$$a_2 \rightarrow a_2^2$$

Para equação do 3º grau:

$$a_0 \rightarrow a_0^2$$
$$a_1 \rightarrow 2a_0a_2 - a_1^2$$
$$a_2 \rightarrow -2a_1a_3 + a_2^2$$
$$a_3 \rightarrow a_3^2$$

Para equações do 4º grau:

$$a_0 \rightarrow a_0^2$$
$$a_1 \rightarrow 2a_0a_2 - a_1^2$$
$$a_2 \rightarrow 2a_0a_4 - 2a_1a_3 + a_2^2$$
$$a_3 \rightarrow 2a_2a_4 - a_3^2$$
$$a_4 \rightarrow a_4^2$$

*O mesmo pode ser feito para os polinômios de 5 e 6 graus. Pode-se usar estes exemplos em seguida para se extrair uma fórmula genérica para grau n. (exercício: tentar a generalização desta fórmula. Perceber que os termos que têm coeficientes ímpares têm sempre o sinal negativo).*

**Localização das raízes de uma equação:**

A determinação dos valores das raízes das equações algébricas e transcendentes é um dos mais importantes problemas na matemática, ocorrendo tanto por si só como também integrado na solução de muitos outros problemas. Exceto no caso particular das equações do segundo grau, de algumas de terceiro e de quarto graus e de certos casos particulares de equações transcendentes, a solução deve ser obtida por métodos gráficos ou numéricos. A maioria dos métodos numéricos é do tipo iterativo e exige inicialmente uma estimativa da raiz, ou pelo menos, a determinação de um pequeno intervalo onde se pode encontrar a raiz. As mais importantes equações não lineares em aplicações à ciência e à engenharia são equações algébricas. Assim, uma boa estratégia para se obter uma raiz é:

- Localização: A localização das raízes exige uma pesquisa ao longo de toda a função, até encontrar-se valores sucessivos da função com sinais contrários. É sabido que a cada mudança de sinal, a função passa por uma raiz real. Para polinômios existem alguns métodos especiais que fornecem uma boa aproximação de todas as raízes existentes. A localização das raízes de forma aproximada pode ser feita através de esboço do gráfico da função e por análise em tabela de valores da função, sempre verificando onde ocorrem mudanças de sinal nestes valores, o que indica que houve uma passagem pelo eixo das abcissas, caracterizando uma raiz real;
- Pré-refinamento: Localizando o subintervalo contendo uma raiz, pode-se usar o método de bisseção para se aproximar da raiz até atender uma primeira tolerância;
- Refinamento: Após estreitar o intervalo onde se encontra a raiz, pode usar o método de Newton-Raphson para encontrar a raiz com a precisão dentro da tolerância desejada.

*Localização de raízes complexas*. Além das raízes reais, as equações não lineares podem apresentar raízes complexas e em muitas aplicações é necessária a determinação destas raízes, especialmente nas equações algébricas. Sabe-se, da teoria, que as raízes complexas ocorrem sempre em ***pares conjugados***, o que simplifica o problema. A localização das raízes complexas é sempre mais trabalhosa que as raízes reais. Existem muitos métodos de localização que funcionam bem, porém, duas maneiras que são usadas para localizar as raízes reais podem também ser empregadas para se determinar as raízes complexas. São:

*Localização por esboço gráfico*. A traçar um gráfico da função as raízes reais aparecem cortando o eixo das abcissas. As raízes complexas implicam numa reversão da função sem cortar o eixo das abcissas, entretanto, localizar essa reversão por si só não dá base para estimar o valor das raízes, exceto talvez no caso em que a reversão seja tão próxima do eixo x que se possa estimar um valor para as raízes com uma parte real correspondente à abcissa da reversão e uma pequena parte imaginária. Uma reversão na função significa uma passagem da primeira derivada por zero, ou seja, ocorre uma inversão no sinal da derivada da função. Caso a reversão não seja próxima do eixo x, pode-se ainda recorrer à localização gráfica das raízes complexas usando o artificio seguinte:

A equação em questão é escrita com variável complexa f(z) = 0, substituindo-se ***z*** por ***x+iy***, obtendo-se f(x+iy) = 0. Nesta equação, separam-se os termos reais dos termos imaginários, obtendo-se uma função composta pela soma de duas outras funções: u(x,y) + i.v(x,u) = 0. Para que seja satisfeita esta equação, é necessário que ambas apresentem a condição: u(x,y) = 0 e v(x,y) = 0, ***implicando na resolução de um sistema de equações não lineares***, cuja solução serão os valores das partes reais e imaginárias das raízes complexas de f(z) = 0.

*Localização das raízes por tabela de valores da função*. Também é possível localizar as raízes complexas examinando os valores complexos da função f(z) para vários valores de ***z*** (também complexos), quando ambas as partes real e imaginária do valor de f(z) trocarem de sinal simultaneamente, sendo, ainda mais, seus valores absolutos pequenos. *Os valores de **z** quando isto acontece podem ser tomados como uma aproximação de uma raiz complexa*. A preparação de uma tabela de valores, neste caso, dá um número de cálculos tão grande que o auxílio de uma combinação calculadora/cabeça é muito melhor que o uso de um computador.

*Refinamento de raízes complexas*. Uma vez localizadas as raízes complexas estas devem ser refinadas e para este caso também existem muitos métodos especiais. Porém, pode-se também usar os métodos acima já vistos devido a sua aplicabilidade mais geral. No caso das raízes complexas, a aplicação do método da bisseção fica ainda mais trabalhosa e a possibilidade de o método de Newton-Raphson não convergir aumenta, de maneira que o uso do pré-refinamento por bisseção fica quase imperativo. Ao usar o método de bisseção, o qual é caracterizado em ir diminuindo o intervalo por reduções sempre ao meio, lembrar que uma raiz complexa é formada de duas partes, portanto, o intervalo também é de duas dimensões, e isto significa manter a raiz enquadrada dentro de um quadrado

(ou um retângulo) com cada vértice em um quadrante. Ao refinar pelo método de Newton-Raphson, a única diferença do caso das raízes reais é que a estimativa inicial deve ser complexa e toda a aritmética deve também ser complexa. *Para ambos casos devem-se usar os métodos pertinentes a um sistema de equações não lineares*.

**Exercícios:**
Encontrar todas raízes reais dos polinômios abaixo:
Se forem detectadas raízes imaginárias apenas apontar a existência das mesmas.

$2x^5 + 3x^4 – x^3 + 2x^2 – 4 = 0$

$x^3 + 2x^2 – x + 1 = 0$

$x^4 – 2x^3 + x^2 + 3x + 2 = 0$

Faz-se reagir num reator experimental de 10 litros, sob pressão e temperaturas adequadas ao processo 5moles de gás nitrogênio e 15 moles de gás hidrogênio segundo a equação química: $N_2 + 3H_2 \Leftrightarrow 2NH_3$
Sabe-se que para as condições adotadas a constante de equilíbrio é de 8,99 $(mol/l)^{-2}$. Quanto se pode esperar em produção de amônia ao se atingir o equilíbrio?
Fórmula da constate de equilíbrio:

$$K_{c(NH_3)} = \frac{[NH_3]^2}{[N_2]^1.[H_2]^3} = 8{,}99 \qquad (mol/l)^{-2}$$

## Sistemas de equações não lineares:

Sistemas de equações não lineares são compostos de equações que são funções de quaisquer variáveis $x_1$, $x_2$, .....$x_n$, podendo envolver potências, produtos de variáveis e funções transcendentes. O tratamento deste tópico será restrito aos sistemas onde m = n, ou seja, que têm a matriz dos coeficientes do tipo quadrada. Entretanto, ao contrário dos sistemas lineares, isto ***não garante*** a existência de soluções reais para o sistema. Para tais sistemas não existem algoritmos de eliminação, sendo necessário recorrer a sistemáticas de localização e posterior refinamento iterativo. Quando não há conhecimento do problema para sugerir estimativas iniciais das raízes, a simples localização destas pode ser um difícil problema, pois, exceto no caso de n = 2, não é possível usar esboços de simplificação de pesquisa como outros métodos permitem. Resumindo, ***não existe nenhum bom método geral para localizar raízes de sistemas de equações não lineares***.

*Método do refinamento sucessivo*. O método mais simples que pode ser aplicado consiste de, após ser obtido um conjunto de estimativas iniciais das raízes, tentar o refinamento usando um esquema semelhante ao método de Gauss-Seidel que existe para os sistemas de equações lineares. Em linhas gerais, o processo consiste em:

- Em cada equação do sistema se refina uma das variáveis, mantendo-se as demais como se fossem constantes e iguais à estima inicial feita para cada uma delas. Para a tarefa de refinamento se usa a técnica de Newton-Raphson considerando que a equação não pertença a um sistema. Para tal, é necessário conhecer a derivada parcial da equação com relação à variável sendo refinada;

- Continua-se este procedimento, até refinar todas as variáveis envolvidas no processo, até uma precisão desejada ou pré-estabelecida.

*Método de Newton-Raphson para sistemas (refinamento simultâneo)*. Uma maneira de se conseguir uma convergência mais rápida é usando o método de Newton-Raphson ampliado para sistemas de equações não lineares. Este método, embora tenha uma convergência mais rápida, envolve a resolução de um sistema de equações lineares a cada iteração, onde se determina de forma simultânea as variações particulares para cada uma das variáveis envolvidas, além das avaliações das derivadas, tornando-o um tanto complicado, de forma que, o mesmo só é utilizado quando o primeiro método descrito acima não funcionar, ou apresentar dificuldade de convergência.
*(para maiores detalhes, pesquisar junto à literatura especializada).*

*Método da continuidade*. No caso de existirem muitas equações, o trabalho pode demandar um esforço considerável. Nestes casos, o método da continuidade é uma forma elegante e prática para ser usada em computadores, na elaboração de programas eficientes. O método consiste basicamente na introdução de uma variável extra nas ***n*** equações $f(x_1, x_2,....x_n) = 0$ de forma a ter $f(x_1, x_2,.....x_n, \lambda) = 0$, onde quando $\lambda = 0$, o sistema se reduz a um sistema de equações lineares, e para $\lambda = 1$, o sistema manterá as equações originais. Um sistema de equações diferenciais ordinárias com relação à variável $\lambda$ pode, então, ser construído usando o princípio matemático abaixo:

$$\sum_{j=1}^{n}\left(\frac{\partial f_i}{\partial x_j}\times\frac{\partial x_j}{\partial \lambda}+\frac{\partial f}{\partial \lambda}\right)=0$$

Onde $x_i$,.....$x_n$, são considerados como funções de $\lambda$. Estas equações são integradas com os valores iniciais obtidos no sistema quando $\lambda = 0$, desde $\lambda = 0$ até $\lambda = 1$. Se a solução pode ser continuada até $\lambda = 1$, os valores de ***x*** obtidos para $\lambda = 1$, será a solução das equações originais. Se a integração resultar em valores infinitos, o parâmetro $\lambda$ deve ser introduzido de outra maneira. Para resolver sistemas de equações diferenciais de forma numérica pode-se recorrer aos métodos descritos para a solução numérica de equações diferenciais ordinárias, a ser visto ainda no decorrer do curso.

Exemplo: (Aplicação do primeiro método aqui citado)
Seja o sistema abaixo, formado pelas equações:

$x^3 - 4{,}2x^2 - 3xy^2 + 5{,}9x + 4{,}2y^2 - 4 = 0$

$3x^2 - 8{,}4x - y^2 + 5{,}9 = 0$

Determinar os valores para ***x*** e ***y***.

Na primeira equação refinamos o valor de y, usando a expressão:

$$y_{i+1} = y_i - \frac{f(y)}{f'(y)} = y_i - \frac{x_i^3 - 4{,}2x_i^2 - 3x_i y_i^2 + 5{,}9x_i + 4{,}2_i^2 - 4}{-6x_i y_i + 8{,}4y_i}$$

$$x_{i+1} = x_i - \frac{f(x)}{f'(x)} = x_i - \frac{3x_i^2 - 8{,}4x_i - y_{i+1}^2 + 5{,}9}{6x_i - 8{,}4}$$

Como valores iniciais, atribuir valor de 0,8 para ***x*** e 0,95 para ***y***

| k | y | x |
|---|---|---|
| 0 | 0,95 | 0,80 |
| 1 | 0,900731 | 0,880190 |
| 2 | 0,941190 | 0,862481 |
| 3 | 0,929854 | 0,869350 |
| 4 | 0,933770 | 0,867102 |
| 5 | 0,932446 | 0,867879 |
| 6 | 0,932899 | 0,867615 |
| 7 | 0,932744 | 0,867706 |
| 8 | 0,932797 | 0,867675 |
| 9 | 0,932779 | 0,867685 |
| 10 | 0,932785 | 0,867682 |

Percebe-se que a convergência é bastante lenta, pois, após dez iterações, somente quatro algarismos significativos estão coincidindo. É importante também, ressaltar o fato de que a ordem em que são colocadas as equações é importante, podendo o sistema convergir para uma ordem e divergir na outra.

Exemplo (Aplicação do método da continuidade)

Sejam as equações:

$F(x,y) = 2 + x + y - x^2 + 8xy + y^3 = 0$
$G(x,y) = 1 + 2x + 3y + x^2 + xy - y.e^x = 0$

Introduzindo $\lambda$ como:

$F(x,y,\lambda) = (2 + x + y) + \lambda(-x^2 + 8xy + y^3) = 0$
$G(x,y,\lambda) = (1 + 2x - 3y) + \lambda(x^2 + xy - y.e^x) = 0$

Para $\lambda = 1$, a equação se mantém original, mas para $\lambda = 0$, o sistema se converte no sistema de equações lineares abaixo:
$2 + x + y = 0$
$1 + 2x - 3y = 0$, o qual tem solução $x = -1{,}4$ e $y = -0{,}6$
As equações diferenciais em relação a $\lambda$, se reduzem a:

$$\frac{\partial f}{\partial x}\times\frac{\partial x}{\partial \lambda}+\frac{\partial f}{\partial y}\times\frac{\partial y}{\partial \lambda}=-\frac{\partial f}{\partial \lambda}$$

(eq´s. I)

$$\frac{\partial g}{\partial x}\times\frac{\partial x}{\partial \lambda}+\frac{\partial g}{\partial y}\times\frac{\partial y}{\partial \lambda}=-\frac{\partial g}{\partial \lambda}$$

Onde as derivadas parciais para o caso, são:

$$\frac{\partial f}{\partial x}=1-2\lambda x+8\lambda y$$
$$\frac{\partial f}{\partial y}=1+8\lambda x+3\lambda y^2$$

$$\frac{\partial g}{\partial x}=2+2\lambda x+\lambda y-\lambda y.e^x$$
$$\frac{\partial g}{\partial y}=-3+\lambda x-\lambda e^x$$

Fazendo as devidas substituições destas nas equações I e integrando com relação a λ, iniciando com x = -1,4 e y = -0,6 para λ = 0 até λ = 1. ***Os valores de x e y em λ = 1 constituem a solução.***

## Interpolação e diferenças finitas:

A prática da engenharia constantemente necessita de dados encontrados em tabelas, os quais são usados como fonte de informação. Uma tabela contém sempre um número limitado de valores. Admitindo-se que estes valores tenham sido obtidos com elevada precisão, resta o problema de obter valores da função correspondente a valores de x que não constem na tabela. Tais valores claramente não poderão ser obtidos exatamente, pois o comportamento da função entre os x´s dados é totalmente desconhecido. São necessários métodos de obtenção de boas estimativas destes valores. Para contornar esse pequeno problema existe a técnica de interpolação, a qual consiste em exigir que a aproximação coincida com a função f(x) que gerou a tabela. Dada uma tabela de valores provenientes desta função, com n pontos, deve-se escolher n funções $f_1(x)$, $f_2(x)$,....$f_n(x)$ cuja combinação linear seja usada como aproximação da função dada, conforme a equação abaixo:

$$f(x) \cong c_1.f_1(x) + c_2.f_2(x) + ..... + c_n.f_n(x)$$

Ou seja, uma função o que gera um sistema de ***n equações*** e ***n incógnitas*** que permite calcular os valores dos coeficientes $c_1$, $c_2$, ....$c_n$ convenientemente. A escolha do conjunto de funções $f_1$, $f_2$,....$f_n$, obviamente influencia de modo decisivo a aproximação obtida e deve ser feita baseando-se na forma que o gráfico dos pontos representa. A experiência com muitas diferentes escolhas é também muito importante. Caso não se tenha experiência, que é o que de fato quase sempre acontece, é muito comum escolher-se as funções abaixo:

$f_1(x) = 1$; $f_2(x) = x$; $f_3(x) = x^2$; .....; $f_n(x) = x^n$

Com o que a função aproximante se converte num polinômio, tal como:

$$f(x) \cong c_1 + c_2.x + c_3.x^2 + c_4.x^3 + ..... + c_n.x^{n-1}$$

Entre os métodos existentes, os mais usados são:

- A *interpolação linear*, que consiste em se passar segmentos de retas (polinômios do 1º grau) através de cada par de pontos da tabela. Em outros termos, encontra-se o valor intermediário através de uma simples regra de três entre estes pontos. Neste caso o polinômio de interpolação, para encontrar o valor da função entre os pontos $x_i$ e $x_{i+1}$, fica:

$$f(x) \cong p(x) = f(x_i) + \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} \times (x - x_i)$$

- *A interpolação quadrática*, a qual consiste em se passar polinômios do 2º grau através do conjunto de três pontos.

$$f(x) \cong p_2(x) = \frac{(x - x_1).(x - x_2)}{(x_0 - x_1).(x_0 - x_2)} \times f(x_0) + \frac{(x - x_0).(x - x_2)}{(x_1 - x_0).(x_1 - x_2)} \times f(x_1) + \frac{(x - x_0).(x - x_1)}{(x_2 - x_0).(x_2 - x_1)} \times f(x_2)$$

- *A interpolação do 3º grau*, o qual, da mesma maneira como foi feito no exemplo anterior, agora usa quatro pontos para se obter uma equação do 3º grau como uma aproximação da função da tabela.

**A fórmula de Lagrange:**

A fórmula de Lagrange fornece diretamente o polinômio interpolante para qualquer número de pontos. A mesma tem a forma genérica abaixo:

$$P_n(x) = \sum_{i=1}^{n} f(x_i).\prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)}$$

Quando n = 2 esta fórmula reproduz a fórmula da interpolação quadrática do item anterior. É possível escrever uma fórmula para estimar o erro cometido ao usar a fórmula de Lagrange, mas esta envolve a determinação da e-nésima derivada de f(x). Como quase sempre é dada uma tabela, a função normalmente é desconhecida e, assim, a fórmula do erro não ajuda muito. Na falta de uma estimativa de erro pode-se calcular aproximações pela fórmula

de Lagrange de graus sucessivamente mais elevados e comparar. Quanto maior o grau, menor o erro de aproximação. Como a fórmula tem muitas subtrações, entretanto, à medida que o grau aumenta os erros de arredondamento se propagam sempre mais, o que faz com que não se use aproximações de grau superior a cinco. (dez em computadores digitais quando trabalham com dupla precisão).

**A fórmula de Newton:**
Um outro método para interpolação pode ser usado empregando-se diferenças finitas entre os pontos constantes em uma tabela. Esta é a fórmula de Newton, apresentada abaixo:

$$f(x) \cong b_0 + b_1.(x - x_0) + b_2.(x - x_0).(x - x_1) + ..... + b_k.(x - x_0).(x - x_1).(x - x_{k-1}) + E_n(x)$$

Onde os termos $b_j$, (j = 1,2,..., k), são as diferenças obtidas a partir dos valores constantes na tabela e são relativos a $x_0$. O valor máximo que k pode assumir é sempre igual a n-1, onde n é igual à quantidade de pontos disponíveis na tabela. O termo $E_n(x)$ é o erro cometido e pode ser calculado através de uma fórmula complementar. O processo todo pode ser feito de forma iterativa, iniciando com o valor do polinômio inicial igual a $f(x_0)$. Quanto mais iterações forem feitas mais o valor se aproxima do valor verdadeiro.

$$f(x) \cong p_i(x) + E_i(x)$$
$$p_0(x) = f(x_0) \qquad \text{i = 1, 2, ...., n-1}$$
$$p_i(x) = p_{i-1}(x) + b_i(x - x_0).....(x - x_{(i-1)})$$

Obviamente, quanto maior n, maior será a aproximação do valor desejado e menor será $E_i(x)$. Normalmente se estende a precisão até atingir a precisão desejada, desprezando-se então, o termo $E_i(x)$. Os termos $b_j$ são definidos pelas fórmulas:

$$b_0 = f(x_0)$$
$$b_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = f(x_{0,} x_1)$$
$$b_2 = \frac{f(x_{1,} x_2) - f(x_{0,} x_1)}{x_2 - x_0} = f(x_{0,} x_1, x_2)$$
$$b_3 = \frac{f(x_{1,} x_2, x_3) - f(x_{0,} x_1, x_2)}{x_3 - x_0} = f(x_{0,} x_1, x_{2,} x_3)$$
$$b_k = \frac{f(x_{1,} x_2, x_k) - f(x_{0,} x_1, x_{(k-1)})}{x_k - x_0}$$

É fácil perceber ao calcular $b_2$, que são necessários calcular dois valores $b_1$. O primeiro deles obtido pela diferença entre $f(x_1)$ e $f(x_0)$ e o outro pela diferença entre $f(x_2)$ e $f(x_1)$. Uma generalização do processo indica que no final resultarão n-1 valores $b_1$, n-2 valores $b_2$, n-3 valores $b_3$, e assim sucessivamente até obter-se apenas 1 valor $b_k$. Assim, ao optar-se por esse método, é necessário se montar a tabela dos valores $b_j$ e estes podem ser obtidos pela expressão genérica de cada $b_j$, apresentadas abaixo:

$$b_{1i} = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} = f(x_{i,} x_{i+1}) \qquad \text{i = 0, 1, 2, ...., k (k = n-1)}$$
$$b_{2i} = \frac{f(x_{i+1,} x_{i+2}) - f(x_{i,} x_{i+1})}{x_{i+2} - x_i} = f(x_{i,} x_{i+1}, x_{i+2}) \qquad \text{i = 0, 1, 2, ...., k-1}$$

$$b_{3i} = \frac{f(x_{i+1,} x_{i+2}, x_{i+3}) - f(x_{i,} x_{i+1}, x_{i+2})}{x_{i+3} - x_0} = f(x_{i,} x_{i+1}, x_{i+2,} x_{i+3}) \qquad \text{i = 0, 1, 2, ...., k-2}$$

$$b_{ki} = \frac{f(x_{i+1,}x_{i+2},...,x_k) - f(x_{0,}x_1,...,x_{(k-1)})}{x_k - x_0} \qquad i = 0$$

Para evitar todo este trabalho, ou para facilitar o processo de programação, existe a generalização abaixo que permite calcular-se somente os valores $b_{j0}$ necessários ao cálculo em questão.

$$b_{ji} = f_{(l=0,j)}(x_{l,}x_{l+1},x_{l+2},....,x_j) = \sum_{l=0}^{j} \frac{f(x_l)}{*(x_l - x_0).(x_l - x_1).(x_l - x_2).....(x_l - x_j)} \quad j = 1, 2,.., k; \quad i = 0, 1, 2, ..,k$$

*O asterisco no denominador indica que o termo* $(x_l\text{-}x_l)$ *não existe. Ou seja, forçar todo* $x_l\text{-}x_l = 1$.

Exemplo:
Construir uma tabela de diferenças para os dados apresentados, abaixo, extraídos de uma tabela de cosh (x).

| i | x | f(x) | $b_1$ | $b_2$ | $b_3$ |
|---|---|---|---|---|---|
| 0 | cosh (0,60) = | 1,18547 | 0,75980 | 0,65600 | 0,15834 |
| 1 | cosh (0,80) = | 1,33743 | 0,95660 | 0,73517 | |
| 2 | cosh (0,90) = | 1,43309 | 1,17715 | | |
| 3 | cosh (1,10) = | 1,66852 | | | |

Usando as fórmulas acima calculamos as diferenças entre os vários pontos existentes na tabela.

$$f|x_0, x_1| = \frac{1{,}33743 - 1{,}18547}{0{,}8 - 0{,}6} = 0{,}75980 \qquad f|x_0, x_1, x_2| = \frac{0{,}95660 - 0{,}75980}{0{,}9 - 0{,}6} = 0{,}65600$$

$$f|x_1, x_2| = \frac{1{,}43309 - 1{,}33743}{0{,}9 - 0{,}8} = 0{,}95660 \qquad f|x_1, x_2, x_3| = \frac{1{,}17715 - 0{,}95660}{1{,}1 - 0{,}8} = 0{,}73517$$

$$f|x_2, x_3| = \frac{1{,}66852 - 1{,}43309}{1{,}1 - 0{,}9} = 1{,}17715 \qquad f|x_0, x_1, x_2, x_4| = \frac{0{,}73517 - 0{,}65600}{1{,}1 - 0{,}6} = 0{,}15834$$

Com a tabela acima, calcular através de uma interpolação de 3ª ordem, o valor do cosh (0,83)

$$\cosh(0.83) \cong 1{,}18547 + 0{,}23 \times 0{,}7598 + 0{,}23 \times 0{,}03 \times 0{,}656 + 0{,}23 \times 0{,}03 \times (-0{,}07) \times 0{,}15834 = 1{,}36467$$

O valor correto para cosh (0,83) = 1,36468, portanto um erro de 0,00001.

Para efeito de comparação, pode-se calcular o valor de cosh (0,83) variando a ordem da interpolação:

1ª ordem: $\cosh(0.83) \cong 1{,}18547 + 0{,}23 \times 0{,}7598 = 1{,}36022$

2ª ordem: $\cosh(0.83) \cong 1{,}18547 + 0{,}23 \times 0{,}7598 + 0{,}23 \times 0{,}03 \times 0{,}656 = 1{,}36475$

3ª ordem: $\cosh(0.83) \cong 1{,}36467$

Pode-se usar outro ponto qualquer como base para a interpolação. Assim, ao tomar-se como base o valor correspondente a x = 0,8, com a tabela acima, pode-se no máximo se atingir uma interpolação de 2ª ordem, ou seja:

$$\cosh(0.83) \cong 1{,}33743 + 0{,}03 \times 0{,}9566 + 0{,}03 \times (-0{,}07) \times 0{,}73517 = 1{,}36458$$

Fica claro que quanto mais pontos estiverem disponíveis, mais precisa será a interpolação. Para evitar o trabalho de montagem da tabela, pode-se usar a generalização apresentada acima. Pela fórmula apresentada tem-se:

$$b_1 = f_0(x_{0,}x_1) = \frac{f(x_0)}{(x_0 - x_1)} + \frac{f(x_1)}{(x_1 - x_0)} = \frac{1{,}18547}{0{,}6 - 0{,}8} + \frac{1{,}33743}{0{,}8 - 0{,}6} = 0{,}75980$$

$$b_2 = f_0(x_0, x_1, x_2) = \frac{f(x_0)}{(x_0 - x_1).(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0).(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0).(x_2 - x_1)} =$$

$$b_2 = \frac{1{,}18547}{(0{,}6 - 0{,}8).(0{,}6 - 0{,}9)} + \frac{1{,}33743}{(0{,}8 - 0{,}6).(0{,}8 - 0{,}9)} + \frac{1{,}43309}{(0{,}9 - 0{,}6).(0{,}9 - 0{,}8)} =$$

$$b_2 = 19{,}75783 - 66{,}87150 + 47{,}76967 = 0{,}65600$$

$$b_3 = f_0(x_0, x_1, x_2, x_4) = \frac{f(x_0)}{(x_0 - x_1).(x_0 - x_2).(x_0 - x_3)} + \frac{f(x_1)}{(x_1 - x_0).(x_1 - x_2).(x_1 - x_3)} + \frac{f(x_2)}{(x_2 - x_0).(x_2 - x_1).(x_2 - x_3)} + \frac{f(x_3)}{(x_3 - x_0).(x_3 - x_1).(x_3 - x_2)} =$$

$$b_3 = \frac{1{,}18547}{(0{,}6 - 0{,}8).(0{,}6 - 0{,}9).(0{,}6 - 1{,}1)} + \frac{1{,}33743}{(0{,}8 - 0{,}6).(0{,}8 - 0{,}9).(0{,}8 - 1{,}1)} + \frac{1{,}43309}{(0{,}9 - 0{,}6).(0{,}9 - 0{,}8).(0{,}9 - 1{,}1)} + \frac{1{,}66852}{(1{,}1 - 0{,}6).(1{,}1 - 0{,}8).(1{,}1 - 0{,}9)} =$$

$$b_3 = -39{,}51567 + 222{,}90500 - 238{,}84833 + 55{,}61733 = 0{,}15834$$

Observações importantes:

Embora manualmente este cálculo (acima) é mais complicado, mas em um programa, resultará simples e direto, pois dispensa a montagem da tabela.

Se os valores de x numa tabela forem igualmente espaçados, é possível usar uma fórmula de interpolação mais simples. A tabela das diferenças é construída tal que:

$$x_{i+1} - x_i = h \qquad \text{i = 1, 2, ...., n-1 (h = constante)}$$

Dada uma tabela com valores válidos, pode-se formar uma nova coluna com os valores das diferenças entre os valores da função denotados como $\Delta f$ e calculados pela equação:

$$\Delta f(x_i) = f(x_{i+1}) - f(x_i) \qquad \text{i = 1, 2,...., n-1}$$

Estes valores recebem o nome de ***diferenças de primeira ordem***. Com estes valores se obtém agora as diferenças de segunda ordem, denotadas por $\Delta^2 f$, calculadas pela fórmula:

$$\Delta^2 f(x_i) = \Delta f(x_{i+1}) - \Delta f(x_i) \quad \text{i = 1, 2,...., n-2}$$

*O processo continua até se obter uma única diferença. Uma análise da variação das diferenças conforme abaixo permite a generalização da fórmula, conforme mostrado abaixo:*

$$b_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{\Delta' f_0}{h}$$

$$b_2 = \frac{\Delta' f_1 - \Delta' f_0}{2h} = \frac{\dfrac{f(x_2) - f(x_1) - (f(x_1) - f(x_0)}{h}}{2h} = \frac{\Delta'' f_0}{2h^2}$$

$$b3 = \frac{\Delta'' f_1 - \Delta'' f_0}{3h} = \frac{\Delta''' f_0}{6h^3}$$

$$b_i = \frac{\Delta^i f_0}{i!.h^i}$$

Substituindo os valores $b_j$, pela equivalente obtida acima se constrói a fórmula de interpolação de Newton para tabelas que tenham intervalos igualmente espaçados. O erro cometido na aproximação é da ordem do primeiro termo de correção abandonado e as únicas exigências são que os valores de ***x*** devem ser igualmente espaçados e se necessita do cálculo prévio da tabela de diferenças, mas tem a vantagem de não se perder cálculos e ser fácil avaliar a perfeição da aproximação obtida.

$$f(x) \cong p_i(x) + E_i(x)$$

$$p_0(x) = f(x_0)$$

$$p_i(x) = p_{i-1}(x) + \frac{\Delta^i f(x_0)}{i!.h^i}(x - x_0).(x - x_1).....(x - x_{i-1}) \qquad \text{i = 1, 2, ...., n-1}$$

*Convém observar que a tabela a ser construída agora consiste de simples diferenças entre os valores constantes na tabela, as quais podem ser obtidas até a ordem n-1.*

Exemplos:

Dada a tabela abaixo, calcular o valor de sen(π/8).

O valor real de sen(π/8) é igual a 0,38268

| x | f(x) |
|---|---|
| -π/4 | -0,707 |
| -π/6 | -0,500 |
| 0 | 0, |
| π/6 | 0,5 |
| π/4 | 0,707 |
| π/3 | 0,866 |
| π/2 | 1 |
| 2π/3 | 0,866 |
| 3π/4 | 0,707 |
| 5π/6 | 0,5 |
| π | 0 |

1. Usando a interpolação linear. Como o ponto π/8 está entre 0 e π/6, deve ser usado o segmento de reta que passa por estes dois pontos, o que dá:

$$\text{sen } x \cong p_1(x) = 0 + \frac{0,5 - 0}{(\pi/6) - 0} \times (x - 0)$$

$$\text{sen } x \cong p_1(x) \cong \frac{3}{\pi}x = \frac{3}{\pi} \times \frac{\pi}{8} = \frac{3}{8} \cong 0,375$$

2. Usando a interpolação quadrática.
   Usando os pontos 0, π/6 e π/4, resulta em:

$$\text{sen } x \cong p_2(x) = \frac{(x - \frac{\pi}{6}).(x - \frac{\pi}{4})}{(0 - \frac{\pi}{6}).(0 - \frac{\pi}{4})} \times 0 + \frac{(x - 0).(x - \frac{\pi}{4})}{(\frac{\pi}{6} - 0).(\frac{\pi}{6} - \frac{\pi}{4})} \times 0,5 + \frac{(x - 0).(x - \frac{\pi}{6})}{(\frac{\pi}{4} - 0).(\frac{\pi}{4} - \frac{\pi}{6})} \times 0,707 =$$

$$\text{sen}(\frac{\pi}{8}) \cong \frac{\frac{\pi}{8} \times (\frac{\pi}{8} - \frac{\pi}{4})}{\frac{\pi}{6} \times (\frac{\pi}{6} - \frac{\pi}{4})} \times 0,5 + \frac{\frac{\pi}{8} \times (\frac{\pi}{8} - \frac{\pi}{6})}{\frac{\pi}{4} \times (\frac{\pi}{4} - \frac{\pi}{6})} \times 0,707 \cong 0,386$$

3. Usando uma interpolação polinomial do 3º grau, usando a fórmula de Lagrange, vem:

$$p_3(x)=f(x_0)\times\frac{x-x_1}{x_0-x_1}\times\frac{x-x_2}{x_0-x_2}\times\frac{x-x_3}{x_0-x_3}+f(x_1)\times\frac{x-x_0}{x_1-x_0}\times\frac{x-x_2}{x_1-x_2}\times\frac{x-x_3}{x_1-x_3}+f(x_2)\times\frac{x-x_0}{x_2-x_0}\times\frac{x-x_1}{x_2-x_1}\times\frac{x-x_3}{x_2-x_3}+f(x_3)\times\frac{x-x_0}{x_3-x_0}\times\frac{x-x_1}{x_3-x_1}\times\frac{x-x_2}{x_3-x_2}=$$

Adotando os pontos -π/6, 0, π/6 e π/4, vem:

$$\text{sen}(\frac{\pi}{8})\cong\frac{\frac{\pi}{8}-0}{-\frac{\pi}{6}-0}\times\frac{\frac{\pi}{8}-\frac{\pi}{6}}{-\frac{\pi}{6}-\frac{\pi}{6}}\times\frac{\frac{\pi}{8}-\frac{\pi}{4}}{-\frac{\pi}{6}-\frac{\pi}{4}}\times(-0,5)+0+\frac{\frac{\pi}{8}-\frac{\pi}{6}}{\frac{\pi}{6}+\frac{\pi}{6}}\times\frac{\frac{\pi}{8}-0}{\frac{\pi}{6}-0}\times\frac{\frac{\pi}{8}-\frac{\pi}{4}}{\frac{\pi}{6}-\frac{\pi}{4}}\times0,5+\frac{\frac{\pi}{8}+\frac{\pi}{6}}{\frac{\pi}{4}+\frac{\pi}{6}}\times\frac{\frac{\pi}{8}-0}{\frac{\pi}{4}-0}\times\frac{\frac{\pi}{8}-\frac{\pi}{4}}{\frac{\pi}{4}-\frac{\pi}{6}}\times0,707\cong$$

$$\text{sen}(\frac{\pi}{8})\cong\frac{3}{4}\times\frac{1}{8}\times\frac{3}{10}\times(-0,5)+\frac{7}{8}\times\frac{3}{4}\times\frac{3}{2}\times0,5-\frac{7}{10}\times\frac{1}{2}\times\frac{1}{2}\times0,707=0,3825$$

Como se pode observar, a precisão da aproximação está relacionada ao grau do polinômio adotado. Para n = 1, obteve-se 0,375, para n = 2, obteve-se 0,386 e para n = 3, obteve-se 0,3825. Entretanto, para cada novo n, os cálculos todos devem ser refeitos, tornando este processo muito pouco prático.

4. Usando a fórmula de Newton.

   Primeiro deve ser construído a tabela das diferenças, lembrando que os pontos ***x*** devem ser igualmente espaçados, e, portanto, retiramos da tabela os pontos não igualmente espaçados, conforme mostrada abaixo:

| i | X | f(x) | Δ´f | Δ´´f | Δ´´´f | $\Delta^{IV}$f | $\Delta^{V}$f |
|---|---|---|---|---|---|---|---|
| 0 | -π/6 | -0,500 | 0,5 | 0 | -0,134 | 0,036 | |
| 1 | 0 | 0, | 0,5 | -0,134 | -0,098 | 0,062 | |
| 2 | π/6 | 0,5 | 0,366 | -0,232 | -0,036 | 0,072 | |
| 3 | π/3 | 0,866 | 0,134 | -0,268 | 0,036 | 0,062 | |
| 4 | π/2 | 1 | -0,134 | -0,232 | 0,098 | | |
| 5 | 2π/3 | 0,866 | -0,366 | -0,134 | | | |
| 6 | 5π/6 | 0,5 | -0,5 | | | | |
| 7 | π | 0 | | | | | |

   (Os pontos marcados com asteriscos estão igualmente espaçados)

Usando os pontos -π/6, 0, π/6 e π/3, vem:

$$p_0(\frac{\pi}{8})=f(-\frac{\pi}{6})=-0,500$$

$$p_1(\frac{\pi}{8})=p_0(\frac{\pi}{8})+\frac{\Delta^1 f(x_0)}{1!.\frac{\pi}{6}}\times(\frac{\pi}{8}-(-\frac{\pi}{6}))=(-0,5)+\frac{0,5}{1\times\frac{\pi}{6}}\times0,91630=0,375$$

$$p_2(\frac{\pi}{8})=p_1(\frac{\pi}{8})+\frac{\Delta^2 f(x_0)}{2!.(\frac{\pi}{6})^2}\times(\frac{\pi}{8}-(-\frac{\pi}{6})).(\frac{\pi}{8}-0)=0,375+\frac{0\times36}{2\times\pi^2}\times0,35983=0,375$$

$$p_2(\frac{\pi}{8})=p_1(\frac{\pi}{8})+\frac{\Delta^3 f(x_0)}{3!.(\frac{\pi}{6})^3}\times(\frac{\pi}{8}-(-\frac{\pi}{6})).(\frac{\pi}{8}-0).(\frac{\pi}{8}-\frac{\pi}{6})=0,375+\frac{-0,134\times6^3}{6\times\pi^3}\times(-0,04710)=0,3823$$

Pode-se continuar o processo dependendo da precisão desejada no resultado. Uma nova iteração trará como resultado o valor 0,38292. Compare este valor com o valor verdadeiro fixando em sua calculadora 6 casas decimais e compare as primeiras cinco casas.

**Aproximações de funções:**

- Regressão linear
- Regressão polinomial
- Suavização de curvas
- Extrapolação

*Regressão linear e polinomial*. Os métodos de interpolação a rigor, extraem o valor intermediário de uma função aproximada obtida com os pontos disponíveis usados para a interpolação, quando estes pontos são obtidos de tabelas confiáveis e cujos valores foram obtidos com extrema precisão. Existe um outro caso, quando os dados são obtidos de experimentos, e cujos valores não são tão confiáveis assim, tanto pelo método de medida quanto mesmo pelo método experimental, às vezes, realizados de forma rudimentar ou por deficiências de ordem técnica. Neste caso costuma-se aproximar os pontos a uma função simples conhecida, sendo a mais simples delas a equação de uma reta. Se a correlação dos pontos apresentarem um afastamento muito grande, tenta-se uma aproximação a função do segundo grau e assim sucessivamente até se obter um bom enquadramento dos pontos disponíveis. Normalmente ao se dispor de tais dados em forma gráfica, pode-se escolher rapidamente a função que mais se aproxima da forma disposta no gráfico. ***O processo consiste em determinar os parâmetros da função de maneira que a soma dos quadrados das distâncias das ordenadas dos pontos dados às ordenadas dos pontos da curva na mesma abscissa seja mínima***. Quando o ajuste aos pontos é feito através de uma reta, diz-se que se fez uma ***regressão linear***. Quando a função que representa os pontos é diferente do primeiro grau, então se fez uma ***regressão polinomial***. Seja um conjunto de pontos $p(x_i)$, com i = 1, n. Para uma regressão linear, usa-se como função uma reta y = ax + b onde as distâncias do ponto $p(x_i)$ ao ponto $y(x_i)$ serão dadas pelas equações:

$$d_{(i)} = p_{(i)} - y_{(i)} = p_{(i)} - ax_{(i)} - b$$

De maneira que a soma dos valores $d_{(i)}$ elevados ao quadrado, seja a mínima possível. Logo, as derivadas parciais de S com relação a ***a*** e a ***b*** devem ser nulas, o que resulta em:

$$S = \sum_{i=1}^{n} (p_{(i)} - ax_{(i)} - b)^2 \qquad \left\{ \begin{array}{l} \dfrac{\partial S}{\partial b} = \displaystyle\sum_{i=1}^{n} -2.(p_{(i)} - ax_{(i)} - b) = 0 \\ \dfrac{\partial S}{\partial a} = \displaystyle\sum_{i=1}^{n} -2x_{(i)}.(p_{(i)} - ax_{(i)} - b) = 0 \end{array} \right\}$$

*Resolvendo o sistema via algébrica obtém-se as fórmulas básicas, que permitem obter os coeficientes da reta, também chamada de reta de regressão diretamente. Estas fórmulas estão apresentadas abaixo:*

$$a = \frac{n.\sum_{i=1}^{n} x_{(i)}.p_{(i)} - (\sum_{i=1}^{n} x_{(i)}.\sum_{i=1}^{n} p_{(i)})}{n.\sum_{i=1}^{n} x_{(i)}^2 - (\sum_{i=1}^{n} x_{(i)})^2} = 0$$

$$b = \frac{\sum_{i=1}^{n} p_{(i)} - a.\sum_{i=1}^{n} x_{(i)}}{n} = \frac{\sum p}{n} - a.\frac{\sum x}{n}$$

Ou também apresentada na forma de sistema escrito de maneira simplificada, tal como:

$$a\sum x + b.n = \sum p$$

$$a\sum x^2 + b\sum x = \sum x.p$$

Para conhecer o erro de enquadramento basta se obter o valor do fator de correlação ***r***. Para se obter este valor em função dos dados iniciais pode se usar a expressão:

$$r^2 = \frac{b.\sum p_{(i)} + a.\sum x_{(i)}.p_{(i)} - \frac{1}{n}.(\sum p_{(i)})^2}{\sum p_{(i)}^2 - \frac{1}{n}.(\sum p_{(i)})^2} = \frac{\left[n\sum xp - \sum x.\sum p\right]^2}{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum p^2 - (\sum p)^2\right]}$$

Se o valor obtido de ***r***, não for satisfatório, ou seja, apresentar um desvio muito grande nos valores, é bem provável que a função que pode representar os pontos em questão seja diferente da equação de uma reta. Pode-se tentar outros tipos de regressão, entre as quais existem as trigonométricas, logarítmicas, exponenciais e polinomiais entre outras. A ***regressão polinomial*** consiste em se buscar uma aproximação da forma:

$$y = a_0 + a_1 x + a_2 x^2 + ..... + a_n.x^n$$

Onde as diferenças das ordenadas aos pontos do polinômio correspondem a

$$d_{(i)} = p_{(i)} - y_{(i)}$$
$$d_{(i)} = p_{(i)} - (a_0 + a_1 x + a_2 x^2 + ..... + a_n.x^n)$$

E a expressão da soma dos pontos fica

$$S = \sum_{i=1}^{n} p_{(i)} - y_{(i)}$$

Da mesma maneira que na regressão linear, as derivadas parciais devem ser iguais a 0. Assim, as expressões destas assumem a forma

$$\frac{\partial S}{\partial a_k} = \sum_{i=1}^{n} -2.x_i^k.(p_{(i)} - a_0 + a_1 x_i + a_2 x_i^2 + ..... + a_n.x_i^n) = 0 \quad \text{k = 1, 2, ...., n}$$

As quais irão gerar um sistema de n equações a n incógnitas, as quais depois de resolvido e rearrumado, pode ser expresso pelas seguintes expressões escritas de maneira simplificadas:

$$\begin{Bmatrix} a_0\sum 1 & + & a_1\sum x & + & a_2\sum x^2 & + & ..... & + & a_n\sum x^n & = & \sum p \\ a_0\sum x & + & a_1\sum x^2 & + & a_2\sum x^3 & + & ..... & + & a_n\sum x^{n+1} & = & \sum x\ .p \\ a_0\sum x^2 & + & a_1\sum x^3 & + & a_2\sum x^4 & + & ..... & + & a_n\sum x^{n+2} & = & \sum x^2.p \\ ..... & + & ..... & + & ..... & + & ..... & + & ..... & = & ..... \\ a_0\sum x^n & + & a_1\sum x^{n+1} & + & a_2\sum x^{n+2} & + & ..... & + & a_n\sum x^{2n} & = & \sum x^n.p \end{Bmatrix}$$

Onde a solução deste sistema fornece os coeficientes procurados da aproximação. Sua solução, entretanto, deve ser feita com cuidado, pois, pela sua formação, o sistema é mal condicionado e, para aproximações polinomiais de grau superior a 10 o problema já fica muito delicado. Até aproximadamente grau 6, entretanto, o processo pode ser usado com segurança. Observar que este sistema é uma generalização da regressão linear, estando o mesmo aqui também contido, bastando fazer b = $a_0$ e a = $a_1$. Observar também que o primeiro termo da matriz é igual a n.

Exemplo 1 (regressão linear):
Num experimento de calibração de um aparelho de dosagem de oxigênio dissolvido realizou-se 7 medidas em função de soluções padrões contendo respectivamente 2, 4, 6, 8, 10, 12 e 14 ppm de $O_2$ dissolvido, obtendo-se leituras sucessivas de 4, 6, 9, 10,15, 19 e 20. Ao se dispor destes valores em gráfico observa-se claramente uma tendência linear. Assim, ao ajustar os pontos acima usando a regressão linear vem:

| i | $x_{(i)}$ | $p_{(i)}$ | $x_{(i)}.p_{(i)}$ | $x^2_{(i)}$ |
|---|---|---|---|---|
| 1 | 2 | 4 | 8 | 4 |
| 2 | 4 | 6 | 24 | 16 |
| 3 | 6 | 9 | 54 | 36 |
| 4 | 8 | 10 | 80 | 64 |
| 5 | 10 | 15 | 150 | 100 |
| 6 | 12 | 19 | 228 | 144 |
| 7 | 14 | 20 | 280 | 196 |
| S | 56 | 83 | 824 | 560 |

$$a_1 = \frac{7\times824 - 56\times83}{7\times560 - 56\times56} = \frac{1120}{784} = 1{,}4286 \qquad a_0 = \frac{83 - 1{,}4286\times56}{7} = \frac{3}{7} = 0{,}4286$$

$$f(x) = 1{,}4286.x + 0{,}4286$$

Exemplo 2 (regressão polinomial):
Obtidos valores num experimento, conforme mostrados na tabela abaixo, determinar os coeficientes do polinômio aproximante de 2º grau e, com este, calcular os valores de y e x inteiros de 1 a 7.

| ***i*** | $x_{(i)}$ | $p_{(i)}$ |
|---|---|---|
| 1 | 0,23 | 5,64 |
| 2 | 1,01 | 7,83 |
| 3 | 2,29 | 17,04 |
| 4 | 2,87 | 21,38 |
| 5 | 4,15 | 24,56 |
| 6 | 5,36 | 16,21 |
| 7 | 5,51 | 14,57 |
| 8 | 6,36 | 0,78 |
| 9 | 6,84 | -7,64 |
| 10 | 7,00 | -12,52 |

Neste caso a função aproximante será da forma de

$$f(x) = a_0 + a_1.x + a_2.x^2$$

Neste caso, montar o sistema de 3 equações, 3 incógnitas, a saber:

$$\left\{\begin{array}{lcl} a_0.n + a_1.\sum x_i + a_2.\sum x_i^2 = \sum p_i \\ a_0.\sum x_i + a_1.\sum x_i^2 + a_2.\sum x_i^3 = \sum p_i.x_i \\ a_{0.}\sum x_i^2 + a_1.\sum x_i^3 + a_3.\sum x_i^4 = \sum p_i.x_i^2 \end{array}\right\}$$

E a tabela abaixo:

| i | $x_{(i)}$ | $p_{(i)}$ | $x^2_{(i)}$ | $x^3_{(i)}$ | $x^4_{(i)}$ | $x_{(i)}.p_{(i)}$ | $x^2_{(i)}.p_{(i)}$ |
|---|---|---|---|---|---|---|---|
| 1 | 0,23 | 5,64 | 0,053 | 0,012 | 0,003 | 1,297 | 0,299 |
| 2 | 1,01 | 7,83 | 1,020 | 1,030 | 1,041 | 7,908 | 7,987 |
| 3 | 2,29 | 17,04 | 5,244 | 12,009 | 27,501 | 39,022 | 89,358 |
| 4 | 2,87 | 21,38 | 8,237 | 23,640 | 67,847 | 61,361 | 176,107 |
| 5 | 4,15 | 24,56 | 17,223 | 71,473 | 296,615 | 101,924 | 422,997 |
| 6 | 5,36 | 16,21 | 28,730 | 153,991 | 825,390 | 86,886 | 465,713 |
| 7 | 5,51 | 14,57 | 30,360 | 167,284 | 921,736 | 80,281 | 442,345 |
| 8 | 6,36 | 0,78 | 40,450 | 257,259 | 1636,170 | 5,335 | 31,551 |
| 9 | 6,84 | -7,64 | 46,786 | 320,014 | 2188,892 | -52,258 | -357,445 |
| 10 | 7,00 | -12,52 | 49,000 | 343,000 | 2401,000 | -87,640 | -613,480 |
| **Σ** | **41,62** | **87,85** | **227,103** | **1349,712** | **8366,195** | **244,116** | **665,432** |

Montando o sistema, vem:

$$\left\{\begin{array}{rcrcrcr} 10{,}000a_0 & + & 41{,}620a_1 & + & 227{,}103a_2 & = & 87{,}850 \\ 41{,}620a_0 & + & 227{,}103a_1 & + & 1349{,}712a_2 & = & 244{,}116 \\ 227{,}103a_0 & + & 1349{,}712a_1 & + & 8366{,}195a_2 & = & 665{,}432 \end{array}\right\}$$

Cuja solução obtida através do processo de eliminação de Gauss-Jordan fornece:
(a solução por relaxação exigiu aproximadamente 7000 iterações para atingir um resultado semelhante)

$a_0 = -2{,}571970$
$a_1 = 15{,}988840$
$a_2 = -2{,}430113$

e a forma final da solução fica $f(x) = -2{,}43x^2 + 15{,}99x - 2{,}57$

*Suavização de curvas*. A suavização de curva é outra técnica baseada no método dos mínimos quadrados que se aplica a ***tabelas com pontos igualmente espaçados*** e serve para eliminar erros grosseiros da tabela, ou seja ajustar os pontos que visivelmente se afastaram do comportamento dos demais. O processo todo consiste em dividir a tabela dada num conjunto de subtabelas com 2k+1 pontos cada uma (sempre um número ímpar de pontos) e em cada subtabela fazer um ajuste por mínimos quadrados e substituir o ponto na tabela original correspondente ao ponto central da subtabela por seu valor calculado conforme este ajuste. Para tanto, são necessárias tantas subtabelas quantos sejam os pontos da tabela dada. Para os pontos no início e no final da tabela dada não é possível fazer uma subtabela com eles no centro e então se usa a primeira e a última subtabelas para eles com ajustes ligeiramente modificados. A vantagem da suavização é que, como os pontos da tabela dada devem ser igualmente espaçados, é possível escrever fórmulas que permitem o cálculo do novo valor da ordenada em função dos valores de cada subtabela. Ao fazer a ***suavização linear*** usando 3 pontos que é o caso mais simples, pode-se usar as fórmulas:

Primeiro ponto: $y_{-1} = \frac{1}{6}(5p_{-1} + 2p_0 - p_1)$

Pontos intermediários: $y_0 = \frac{1}{3}(p_{-1} + p_0 + p_1)$ (média aritmética dos três pontos)

Último ponto: $y_1 = \frac{1}{6}(-p_{-1} + 2p_0 + 5p_1)$

Exemplo 1 (suavização de curvas):
Dada uma tabela de valores obtida em experimentação tal como a apresentada abaixo, suavizar a mesma em duas etapas.

| i | x(i) | p(i) |
|---|---|---|
| 1 | 0 | 2 |
| 2 | 1 | 3 |
| 3 | 2 | 5 |
| 4 | 3 | 5 |
| 5 | 4 | 9 |
| 6 | 5 | 8 |
| 7 | 6 | 10 |

Observar a grande disparidade dos dados obtidos. Se dispostos em gráfico, se percebe claramente a tendência de reta, no entanto, alguns valores estão completamente fora, tal como o ponto 4 e o ponto 6. Ao fazer a suavização tomamos os primeiros 3 valores como sendo a primeira subtabela, onde o primeiro termo corresponde ao termo –1

da subtabela, o segundo ao termo 0, e o terceiro como sendo o termo 1 da subtabela, representados pela letra ***j***. Aplica-se estes valores nas equações acima, ficando:

| i | x(i) | p(i) | j |
|---|---|---|---|
| 1 | 0 | 2 | −1 |
| 2 | 1 | 3 | 0 |
| 3 | 2 | 5 | 1 |

$$p_1^{(1)} = y_{-1} = \frac{1}{6} \times (5 \times 2 + 2 \times 3 - 5) = 1{,}8333$$

$$p_2^{(1)} = y_0 = \frac{1}{3} \times (2 + 3 + 5) = 3{,}3333$$

Para o terceiro termo da tabela, monta-se outra subtabela com os dados subseqüentes.

| i | x(i) | p(i) | j |
|---|---|---|---|
| 2 | 1 | 3 | −1 |
| 3 | 2 | 5 | 0 |
| 4 | 3 | 5 | 1 |

$$p_2^{(1)} = y_{-1} = \frac{1}{6} \times (5 \times 3 + 2 \times 5 - 5) = 3{,}3333$$

$$p_3^{(1)} = y_0 = \frac{1}{3} \times (3 + 5 + 5) = 4{,}3333$$

Foi refeito o cálculo do 2º termo apenas para mostrar a validade das equações, as quais fornecem o mesmo valor. Para o quarto termo o processo continua

| i | x(i) | p(i) | j |
|---|---|---|---|
| 3 | 2 | 5 | −1 |
| 4 | 3 | 5 | 0 |
| 5 | 4 | 9 | 1 |

$$p_4^{(1)} = y_0 = \frac{1}{3} \times (5 + 5 + 9) = 6{,}3333$$

Para o quinto termo:

| i | x(i) | p(i) | j |
|---|---|---|---|
| 5 | 3 | 5 | −1 |
| 5 | 4 | 9 | 0 |
| 6 | 5 | 8 | 1 |

$$p_5^{(1)} = y_0 = \frac{1}{3} \times (5 + 8 + 8) = 7{,}3333$$

Para o sexto e último termos:

| i | x(i) | p(i) | j |
|---|---|---|---|
| 5 | 4 | 9 | −1 |
| 6 | 5 | 8 | 0 |
| 7 | 6 | 10 | 1 |

$$p_6^{(1)} = y_0 = \frac{1}{3} \times (9 + 8 + 10) = 9{,}0000$$

$$p_7^{(1)} = y_1 = \frac{1}{6} \times (-9 + 2 \times 8 + 5 \times 10) = 9{,}5000$$

O processo agora pode ser repetido n vezes, sempre se utilizando os novos valores obtidos pela suavização anterior, indicada pelo valor entre parêntesis colocado no lugar do expoente em p, A cada nova iteração no processo faz com que os dados obtidos se aproximem mais da reta obtida pela regressão linear. Para efeito de comparação observar a tabela abaixo com 2 iterações de suavização e o valor da regressão linear.

| i | x(i) | p(i) | $y^{(1)}$ | $y^{(2)}$ | $y_{RL}$ |
|---|---|---|---|---|---|
| 1 | 0 | 2 | 1,8333 | 1,9166 | 1,929 |
| 2 | 1 | 3 | 3,3333 | 3,1666 | 3,286 |
| 3 | 2 | 5 | 4,3333 | 4,6666 | 4,643 |
| 4 | 3 | 5 | 6,3333 | 6,0000 | 6,000 |

Exemplo 2. Ajuste de curva tração/deformação de um tipo de aço.
Feito um ensaio de tração em uma barra de um tipo de aço em uma máquina universal de Amsler, foram obtidos os valores constantes na tabela abaixo:

| i | tração (ton/cm$^2$) | d ( l-l$_0$/l ) | d(regressão) | d(suavizado**)** |
|---|---|---|---|---|
| 1 | 0,8 | 0,15 | -0,08 | 0,17 |
| 2 | 1,8 | 0,52 | 0,32 | 0,48 |
| 3 | 2,8 | 0,76 | 0,72 | 0,80 |
| 4 | 3,8 | 1,12 | 1,12 | 1,12 |
| 5 | 4,8 | 1,47 | 1,52 | 1,43 |
| 6 | 5,8 | 1,71 | 1,92 | 1,75 |
| 7 | 6,8 | 2,08 | 2,32 | 2,12 |
| 8 | 7,8 | 2,56 | 2,72 | 2,61 |
| 9 | 8,8 | 3,19 | 3,12 | 3,37 |
| 10 | 9,8 | 4,35 | 3,52 | 4,26 |
| 11 | 10,0 | 4,55 | 4,06 | 4,84 |
| 12 | 10,2 | 5,64 | 5,59 | 5,65 |
| 13 | 10,4 | 6,76 | 7,18 | 6,86 |
| 14 | 10,6 | 8,17 | 8,82 | 8,34 |
| 15 | 10,8 | 10,10 | 10,50 | 10,32 |
| 16 | 11,0 | 12,70 | 12,30 | 13,00 |
| 17 | 11,2 | 16,20 | 14,10 | 16,40 |
| 18 | 11,4 | 20,30 | 24,80 | 22,17 |
| 19 | 11,6 | 30,00 | 39,00 | 36,77 |
| 20 | 11,8 | 60,00 | 53,10 | 56,62 |

Deseja-se obter a representação aproximada para a deformação em função da tração [d = f(t)]. Como os dados são obtidos de ensaio com prováveis erros, a aproximação será feita por regressão.

A regressão será dividida em 3 trechos, a saber:

- Entre os pontos 1 e 10, Regressão linear resultando em **d = -0,4 + 0,4.t**
- Entre os pontos 10 a 17, regressão quadrática resultando em **d = 1,06 – 6,9.t + 0,72 t$^2$**
- Entre os pontos 17 a 20, regressão linear resultando em **d = -780 + 70,6.t**

*Nas transições entre os trechos pode-se depois tirar a média aritmética entre as duas aproximações.*
*A 4$^a$ coluna da tabela indica os valores obtidos pela regressão. A 5$^a$ coluna indica os valores obtidos pela suavização.*

Refazer este exercício para conferir.

Observar no exercício acima a melhor concordância dos dados entre a regressão e a suavização.

*Extrapolação***.** Extrapolar significa usar uma aproximação funcional de qualquer tipo para calcular valores fora do intervalo em que se pode experimentar com a função. Assim, por exemplo, é o caso de se obter através de regressão linear uma reta e nela calcular valores fora dos valores experimentados. Os métodos de se obter uma aproximação podem ser os mesmos apresentados anteriormente. O problema é que não é possível se fazer uma boa estimativa do erro cometido e assim, a extrapolação deve ser usada com muito cuidado e o resultado somente pode ser usado acompanhado das ressalvas convenientes.

**Derivação e integração numéricas**

- Derivação numérica
- Integração numérica

Ao analisar uma função qualquer, necessita-se às vezes, conhecer a variação da mesma, e isto implicam determinar a função que representa a derivada desta função. Por outro lado, às vezes também é necessário se determinar outra função, aquela que representa a função primitiva da função dada f(x), ou seja, obter aquela função que quando derivada, fornece a função f(x). A função primitiva é sempre a função integral da função f(x). No entanto, tanto a derivação como a integração nem sempre são fáceis e dependem muito da complexidade da função dada. Para aplicações em problemas reais, normalmente não se deseja tanto a função, mas valores destas funções dentro de um intervalo válido para a aplicação. Em ambos os casos, os métodos numéricos são usados, porém somente quando não der para usar métodos analíticos ou quando seu uso for excessivamente trabalhoso, o que ocorre tipicamente quando f(x) é conhecida na forma de tabela ou tem forma analítica extremamente complexa.
Também em ambos os casos a solução é a mesma. Aproxima-se f(x) por um polinômio no intervalo de interesse e efetua-se a operação desejada analiticamente sobre o polinômio obtido, o que resulta em um valor da derivada ou da integral de f(x). Essa aproximação é muito melhor no caso da integração, que é dada pela área debaixo da curva, pois os erros de área entre as curvas ***tendem a se anular***, enquanto que no caso da derivação, que é dada pela inclinação da curva no ponto analisado, os erros tendem a se acentuar. Observar então, que na integração os erros se ***atenuam*** e na derivação os erros se ***intensificam***. Assim, apesar de a derivação numérica ser tão simples de executar quanto a integração, ela raramente é usada. Só mesmo quando os dados estão na forma de tabela.

*Derivação numérica.* Para se obter a expressão analítica da derivada, pode-se aproximar a função usando a fórmula de Newton, se os valores de x estiverem igualmente espaçados, ou então, a fórmula de Lagrange se estes estiverem dispersos. Uma vez obtido o polinômio, ***este deve ser derivado analiticamente.***
Para o caso quando se deseja calcular a derivada de uma função em um único ponto da função, a maneira mais simples de se obter uma aproximação consiste em fazer passar uma reta por dois pontos da função ($f_{(x0)}$ e $f_{(x1)}$). A expressão abaixo indicará a inclinação média da função entre os pontos $x_0$ e $x_1$. Se o valor $x_1$ escolhido for muito próximo de $x_0$, isto é, se $\Delta x \rightarrow 0$, então o valor obtido pela expressão será bastante aproximado ao valor real da derivada no ponto $x_0$.

Inclinação da curva no ponto $x_0$ = $$f'(x_0) \cong \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$ (eq. I)

Quando estiverem disponíveis valores da função em pontos igualmente espaçados de uma distancia h, para $h = x_1 - x_0$, incluindo, além do ponto $x_0$ onde se deseja a derivada, dois pontos, um anterior ao ponto desejado e outro posterior, pode-se fazer uma aproximação polinomial de segundo grau, que resulta nas equações abaixo, obtidas pela derivação da fórmula de Lagrange, usada somente para valores contidos em tabelas.

$$f'(x_{-1}) \cong \frac{-3.f(x_{-1}) + 4.f(x_0) - f(x_1)}{2 \times h}$$

$$f'(x_0) \cong \frac{-f(x_{-1}) + f(x_1)}{2 \times h}$$

$$f'(x_1) \cong \frac{f(x_{-1}) - 4.f(x_0) + 3.f(x_1)}{2 \times h}$$

A primeira e a última fórmula são usadas somente no inicio e no final da tabela, da mesma maneira como é feito na técnica de suavização de curvas. Ao analisar a fórmula central para três pontos, percebe-se que a mesma aplica o conceito da eq. I, apenas com a diferença de que esta obtém a inclinação média entre o primeiro e o último ponto, subentendo que o ponto do meio tenha então a inclinação média entre os dois pontos adjacentes, o que nem sempre é verdade. Atentar também que a precisão irá depender muito do intervalo entre os valores de ***x***. O erro será sempre maior quanto mais espaçados estiverem.
Para aumentar a precisão pode-se optar em se adotar 5 pontos consecutivos, resultando numa aproximação polinomial de quarto grau, um pouco mais preciso que aquele de segundo grau, resultando nas fórmulas:

$$f'(x_{-2}) \cong \frac{-25.f(x_{-2}) + 48.f(x_{-1}) - 36.f(x_0) + 16.f(x_1) - 3.f(x_2)}{12 \times h}$$

$$f'(x_{-1}) \cong \frac{-3.f(x_{-2}) - 10.f(x_{-1}) + 18.f(x_0) - 6.f(x_1) + f(x_2)}{12 \times h}$$

$$f'(x_0) \cong \frac{f(x_{-2}) - 8f(x_{-1}) + 8f(x_1) - f(x_2)}{12 \times h}$$

$$f'(x_1) \cong \frac{-3.f(x_{-2}) - 10.f(x_{-1}) + 18.f(x_0) - 6.f(x_1) + f(x_2)}{12 \times h}$$

$$f'(x_2) \cong \frac{-f(x_{-2}) + 6.f(x_{-1}) + -18.f(x_0) + 10.f(x_1) + 3.f(x_2)}{12 \times h}$$

*Integração numérica***.** O processo de integração numérica se aplica somente ao caso de integral definida entre dois pontos. Muitos métodos foram desenvolvidos para obter o valor de uma determinada função. Entre os métodos numéricos mais usados estão o método dos trapézios, o método de Simpson e o método de Romberg. Os métodos numéricos podem ser aplicados sucessivamente, permitindo a resolução de integrais duplas e triplas. O método básico na solução de uma integral consiste em dividir o intervalo $x_1$-$x_0$ em vários sub-intervalos e com eles se obter o produto $\Delta x.f(x)$ e ir acumulando a soma até o final. A maior fonte de erros neste processo está na obtenção do valor de f(x) (valor da função relativo a ***x***), o qual pode ser reduzido ao fazer $\Delta x \rightarrow 0$, ou seja, aumentar a quantidade de sub-intervalos de ***x***. Ao se processar os produtos $\Delta x.f(x)$ obtém-se áreas de pequenos retângulos que somados irão fornecer o valor da área total entre os pontos $x_0$ a $x_1$. O erro cometido nesta soma consiste em que as áreas sob os pontos considerados só serão verdadeiros retângulos quando a função for uma reta horizontal. Em todos os demais casos o erro pode ser muito grande. Existem casos onde o erro é atenuado. Quando a função sobre e desce dentro do intervalo de integração, ***o erro por falta*** cometido no trecho de subida pode ser compensado pelo ***erro por excesso*** cometido na descida. Para compensar este pequeno problema usa-se o ***método dos trapézios***, ou seja, considerar o valor médio da função entre os pontos x e $\Delta x$, o que equivale a somar a área de um triângulo à área do retângulo obtido pelo produto $\Delta x.f(x)$ sendo esta área o valor obtido pelo produto da diferença de valores entre $f_{(x+\Delta x)}$ e $f_{(x)}$ e $\Delta x$ dividido por 2. Representando em fórmula genérica, a integração pelo método dos trapézios adquire a forma:

$$\int_{x_0}^{x_1} f(x)dx \cong \sum_{i=0}^{n} \frac{f(x_i) + f(x_{i+1})}{2} \times (x_{i+1} - x_i) \cong \frac{h}{2}[f(x_0) + 2.f(x_1) + 2.f(x_2) + .... + 2.f(x_{n-1}) + f(x_n)]$$

Para um único intervalo a expressão fica:

$$T_{(1)} \cong \frac{h}{2}[f(x_0) + f(x_1)]$$

(média aritmética entre os dois valores multiplicados pelo valor do intervalo de ***x***)

É possível estimar o erro absoluto cometido ao se integrar uma função qualquer pelo método dos trapézios usando-se a fórmula, onde $M_2(x)$ é um máximo da segunda derivada da função, no intervalo de integração.

$$\varepsilon = \frac{h^2}{12} \times (x_n - x_0) \times M_2(x)$$

Como se vê, o erro é proporcional a $h^2$. É possível calcular o número de intervalos em que uma função deve ser subdividida para que a integral por trapézios tenha uma precisão predefinida.

Um método mais preciso, consiste em usar a ***interpolação quadrática*** em cada conjunto de três pontos da função e em seguida integrar os polinômios de segundo grau resultantes. Este método recebe o nome de ***método de Simpson***. Um desenvolvimento semelhante ao usado no caso do método dos trapézios resulta, para o caso de dois intervalos de mesma largura h, a fórmula abaixo:

Para um intervalo de x subdividido em 2 segmentos**:**

$$\int_{x_0}^{x_1} f(x)dx \cong \frac{h}{3}[f(x_0) + 4.f(x_1) + f(x_2)]$$

Para um intervalo de ***x,*** subdividido em 4 ou mais segmentos:

$$\int_{x_0}^{x_1} f(x)dx \cong \frac{h}{3}[f(x_0) + 4.f(x_1) + 2.f(x_2) + 4.f(x_3) + 2.f(x_4) + .... + 2.f(x_{n-2}) + 4.f(x_{n-1}) + f(x_n)]$$

O método de Simpson exige sempre que o intervalo de integração seja subdividido em um número par de subintervalos, ou seja, ***n deve ser sempre par***. Para esta fórmula é também possível estimar uma cota superior de erro absoluto cometido através da expressão, onde $M_2(x)$ é um máximo da quarta derivada de f(x) no intervalo de integração.

$$\varepsilon = \frac{h^4}{180} \times (x_n - x_0).M_2(x)$$

Observações:
Se f(x) for dada na forma de tabela ou se for difícil obter os máximos das derivadas, pode-se usar tanto a ***fórmula dos trapézios*** como a ***fórmula de Simpson*** de uma maneira recursiva, ou seja, Estabelece-se um número inicial (par) de subintervalos e aplica-se uma das fórmulas. Em seguida, dobra-se o número de subintervalos e aplica-se novamente a mesma fórmula e comparam-se os resultados. Este segundo resultado deve ser melhor que o primeiro, pois o erro é proporcional a uma potência de ***h*** e este foi reduzido à metade, logo o erro deve ser reduzido à raiz quadrada se aplicado o método dos trapézios ou a raiz quarta, se for aplicado o método de Simpson. O processo é continuado até que a diferença entre dois resultados sucessivos esteja abaixo de uma tolerância desejada.

*Observar que enquanto o método dos trapézios usa a interpolação linear, o método de Simpson usa a interpolação quadrática.*

Um método que permite um uso mais eficiente dos sucessivos valores obtidos pela sistemática recursiva descrita acima é o ***método de Romberg***, o qual é um melhoramento aplicado ao método dos trapézios reduzindo o número de cálculos para se chegar à precisão desejada.

Neste método, usa-se em primeira mão o processo dos trapézios e chama-se o resultado obtido ($T_{(h)}$) de $R_0(h)$ *Resultado Romberg de ordem zero*. Ao calcular a integral pelo método dos trapézios comete-se um erro proporcional a $h^2$, ou seja, pode-se representar o valor da integral como sendo:

$$I = \int_a^b f(x)dx = T_{(h)} + \varepsilon(h^2)$$

Onde o erro $\varepsilon$ tem a forma matemática igual ao um polinômio da forma...

$$\varepsilon = c_2h^2 + c_4h^4 + c_6h^6 + ..... + c_nh^n + .....$$

Resultando em

$$I = T_{(h)} + \varepsilon(h^2) = T_{(h)} + c_2h^2 + c_4h^4 + c_6h^6 + ..... + c_nh^n + .....$$

Para eliminar este erro, pode-se calcular novamente pelo método dos trapézios reduzindo ***h*** à metade, ou seja, dobrando-se a quantidade de subintervalos. Com este resultado ($T_{(h/2)}$) e o resultado $R_0(h)$ pode-se então calcular o $R_1(h)$ *Resultado Romberg de primeira ordem*, através da relação:

$$R_{1(h)} = \frac{4T_{1(h/2)} - T_{1(h)}}{3}$$

Ao se obter o valor $R_1(h)$, elimina-se a parcela do erro relativa ao termo $h^2$, restando as parcelas subseqüentes.
Para continuar o processo, é necessário calcular pelo método dos trapézios o valor T(h/4), o que permite calcular agora o valor $R_1(h/2)$ e com este valor obter o valor de $R_2(h)$ através da relação:

$$R_{1(h/2)} = \frac{4T_{1(h/4)} - T_{1(h/2)}}{3} \qquad R_{2(h)} = \frac{4R_{1(h/2)} - R_{1(h)}}{3}$$

Este novo valor eliminou a parcela de erro relativa ao termo $h^4$. Para continuar, elimina-se progressivamente as parcelas referentes a $h^6$, $h^8$, etc., sempre se aproximando cada vez mais do verdadeiro valor da integral.

A fórmula geral para a fórmula de Romberg é:

$$R_{n(h/2^k)} = \frac{4^n.R_{n-1(h/2^{k+1})} - R_{n-1(h/2^k)}}{4^n - 1}$$

Exemplo:
Calcular a integral definida abaixo dentro do intervalo apresentado, com uma precisão mínima de casas decimais ($\varepsilon < 0{,}001$):

$$Y = \int_2^3 (\ln x.)^2 dx \qquad y = (\ln x)^2$$

Para efeito de comparação entre os métodos, usaremos 6 casas decimais. Inicialmente fazendo os cálculos usando um cálculo simples de áreas. Dividindo-se o intervalo de x entre 2 e 3 em dois subintervalos, teremos que o primeiro subintervalo vai de 2 a 2,5 e o segundo vai de 2,5 a 3. Calculando-se os valores desta função para estes valores obtém-se a tabela:

| x | y |
|---|---|
| 2,0 | 0,480453 |
| 2,5 | 0,839589 |
| 3,0 | 1,206949 |

Fazendo-se os cálculos de uma área aproximada, podemos tomar como base os valores iniciais da função em cada subintervalo, quando então se obtém a área inferior ou pode-se optar pelos valores finais de cada subintervalo, quando se obtém a área superior. Uma erra por falta enquanto a outra erra por excesso. Para diminuir o erro cometido pode-se então se fazer a média entre os dois valores que é equivalente a usar a integração pelo método dos trapézios, como pode ser verificado abaixo.

$$A_i = 0,480453 \times 0,5 + 0,839589 \times 0,5 = 0,660021$$

$$A_s = 0,839589 \times 0,5 + 1,206949 \times 0,5 = 1,023269$$

$$A_m = \frac{A_i + A_s}{2} = \frac{0,660021 + 1,023269}{2} = 0,841645$$

Fazendo a integração pelo método dos trapézios com h = 0,5 (dois subintervalos entre 2 e 3), como foi feito acima, tem-se:

$$T_{(h/2)} = \frac{0,5}{2}[0,480453 + 2 \times 0,839589 + 1,206949] = 0,841645$$

Para aumentar a precisão dobramos o número de subintervalos esse refaz o cálculo, com isto obtém-se a tabela.

| x | y |
|---|---|
| 2,00 | 0,480453 |
| 2,25 | 0,657608 |
| 2,50 | 0,839589 |
| 2,75 | 1,023336 |
| 3,00 | 1,206949 |

$$T_{(h/4)} = \frac{0,25}{2}[0,480453 + 2 \times 0,657608 + 2 \times 0,839589 + 2 \times 1,023336 + 1,206949] = 0,841077$$

Com 8 intervalos tem-se:

| x | y | | |
|---|---|---|---|
| 2,000 | 0,480453 | x 1 | 0,480453 |
| 2,125 | 0,568172 | x 2 | 1,136344 |
| 2,250 | 0,657608 | x 2 | 1,315216 |
| 2,375 | 0,748221 | x 2 | 1,496441 |
| 2,500 | 0,839589 | x 2 | 1,679177 |
| 2,625 | 0,931381 | x 2 | 1,862762 |
| 2,750 | 1,023336 | x 2 | 2,046673 |
| 2,875 | 1,115247 | x 2 | 2,230495 |
| 3,00 | 1,206949 | x 1 | 1,206949 |
| Total | | | **13,454510** |

$$T_{(h/8)} = \frac{0,125}{2}[13,454510] = 0,840907$$

Comparando estes valores com o valor obtido com a integral definida obtida analiticamente tem-se:

$$Y = \int_2^3 (\ln x)^2 dx = x(\ln x)^2 - 2x \ln x + 2x =$$

$$Y = [3 \times (\ln 3)^2 - 2 \times 3 \times \ln 3 + 2 \times 3] - [2(\ln 2)^2 - 2 \times 2 \times \ln 2 + 2 \times 2] =$$

$$Y = 3 \times 1,206949 - 6 \times 1,098612 + 6 - 2 \times 0,480453 + 4 \times 0,693147 - 4 =$$

$$Y = 0,840857$$

Aplicando o refinamento de Romberg, tem-se:
Para aplicar o refinamento de Romberg deve-se proceder ao cálculo conforme o esquema mostrado abaixo, onde cada fila deve ser ***toda*** calculada antes de se passar a fila seguinte.

| | | | |
|---|---|---|---|
| $R_0(1)$ | | | |
| $R_0(0,5)$ | $R_1(1)$ | | |
| $R_0(0,25)$ | $R_1(0,5)$ | $R_2(1)$ | |
| $R_0(0,125)$ | $R_1(0,25)$ | $R_2(0,5)$ | $R_3(1)$ |
| ..... | ..... | ..... | ..... |

| | | | | | |
|---|---|---|---|---|---|
| 0 | $R_0(1) = T_0(1) =$ | 0,843701 | | | |
| 2 | $R_0(0,5) = T_0(0,5) =$ | 0,841645 | a | | |
| 4 | $R_0(0,25) = T_0(0,25) =$ | 0,841077 | b | c | |
| 8 | $R_0(0,125) = T_0(0,125) =$ | 0,840907 | d | e | f |

$$a = \frac{4 \times 0,841645 - 0,843701}{3} = 0,840960$$

$$b = \frac{4 \times 0,841077 - 0,841645}{3} = 0,840888$$

$$d = \frac{4 \times 0,840907 - 0,841077}{3} = 0,840850$$

$$c = \frac{16b - a}{15} = \frac{15 \times 0,840888 - 0,840960}{15} = 0,840883$$

$$e = \frac{16d - b}{15} = \frac{16 \times 0,840850 - 0,840888}{15} = 0,840847$$

$$f = \frac{64e - c}{63} = \frac{64 \times 0,840847 - 0,840883}{63} = 0,840846$$

Para finalizar a comparação, obter o resultado da integral acima usando o método de Simpson. Aplicando 4 subintervalos e substituído os valores na fórmula vem:

$$S_{(0,\,5)} \cong \frac{0,50}{3}[0,480453 + 4 \times 0,839589 + 1,206949] = 0,841639$$

$$S_{(0,\,25)} \cong \frac{0,25}{3}[0,480453 + 4 \times 0,657608 + 2 \times 0,839589 + 4 \times 1,023336 + 1,206949] = 0,840863$$

$$S_{(0,\,125)} \cong \frac{0,125}{3}[20,180552] = 0,840856$$

Comparando-se os resultados nos três processos e o valor obtida via analítica, 0,840855844.... :

| N | h | T(h) | S(h) | $R_1(h)$ | $R_2(h)$ | $R_3(h)$ |
|---|---|---|---|---|---|---|
| 1 | 1 | 0,843701 | - | 0,840960 | 0,840883 | 0,840847 |
| 2 | 0,5 | 0,841645 | 0,841639 | 0,840888 | 0,840847 | |
| 4 | 0,25 | 0,841077 | 0,840863 | 0,840850 | | |
| 8 | 0,125 | 0,840907 | 0,840856 | | | |

*Percebe-se que o método de Simpson também fornece bons resultados e de todos os métodos, é de longe o mais usado, porém quando se deseja maior precisão usa-se o refinamento de Romberg***.**

Percebe-se também que estes métodos exigem bem poucos passos para a solução de uma integral numérica. São bastante úteis quando se precisa de uma integração rapidamente quando não se dispõem de recursos melhores e que os cálculos devam ser feitos manualmente. Atualmente, entretanto, com o advento da informática pode-se usar de programação para se obter o valor de uma integração com altíssimo grau de precisão de uma maneira muito simples, ou seja:

1. Dividir o intervalo de x num número desejado de subintervalos, pode ser um número muito grande, como por exemplo, 1000 ou mais. O valor obtido é chamado de h. $h = (x_1 - x_0) / n$
2. Uma vez obtido o subintervalo, obter o ponto médio de h, dividindo o valor h por dois.
3. Fazer $z = x_0$
4. Somar a z o valor h/2, obtendo-se assim o primeiro valor de z com o qual se usa na obtenção de $f(z_0)$. Obter o valor da função f(x+h/2) e multiplicar este valor por h. Chama-se este valor de Σ
5. Em seguida, soma-se ao valor de $z_0$, o valor h, obtendo-se $z_1$
6. Com este valor obter o valor da função $f(z_1)$. Multiplicar este valor por h e somar a Σ
7. Somar a $z_1$, h novamente, conseguindo-se $z_2$. Repetir o passo 6, sempre incrementando o valor h em z, enquanto $z < x_1$

Resumindo tudo se tem:

$$\int_{x_0}^{x_1} f(x)dx = \left[ \left( f(x_0 + \Delta x/2) \right) + \sum_{i=1}^{n-1} \left( f(z_i) \right) \right] \times h \qquad h = \frac{x_1 - x_0}{n} \qquad z_i = x_0 + h/2 + i \times h$$

**Exercícios:**
Resolver as seguintes integrações numéricas. Tentar todos os métodos apresentados.

$$\int_2^3 \sqrt{\ln x}\,dx =$$

$$\int_2^4 \frac{(\ln x)^3 dx}{x^2} =$$

$$\int_0^{10} \frac{x^2 dx}{\ln x} =$$

$$\int_1^5 \frac{dx}{x.\ln x} =$$

$$\int_0^{\pi/2} \ln(\tan x)dx =$$

**Solução numérica de equações diferenciais**

- Método de Euler
- Métodos de Runge-Kutta

*Método de Euler*. Este capítulo trata da solução de equações diferencial, que em sua forma mais simples se apresentam como mostrado abaixo:

$$\frac{dy}{dx} = f(x, y)$$

Como toda derivada ao ser integrada, gera outra função, a qual recebe o nome de primitiva, esta difere da primitiva real apenas por uma constante que desapareceu no ato da derivação. Para recompor esta constante, toda equação diferencial necessita de uma condição inicial ou também chamada de *condição de contorno*. Como condição inicial pode-se ter uma função da forma como mostrada abaixo:

$$y(x_0) = y_0$$

Como f(x,y) é dada e sendo a derivada de ***y*** em relação a ***x***, esta representa a inclinação da função em cada ponto $\boldsymbol{x}_i$ , sendo razoável admitir esta inclinação constante dentro de um espaço suficientemente pequeno de ***x***.
Assim, partindo do valor inicial, $y_0$ obtém $y_1$. Neste ponto calcula-se nova inclinação $f(x_1,y_1,)$ e com esta obtém $y_2$ e assim sucessivamente até $y_n$, obtendo-se desta forma os pontos relativos à função ***y***. A equação diferencial desta função pode ser escrita na forma:

$$dy = f(x, y)dx$$

A qual integrada no intervalo genérico de $x_i$ a $x_{i+1}$, adquire a forma:

$$\int_{y(x_i)}^{y(x_{i+1})} dy = \int_{x_i}^{x_{i+1}} f(x, y)dx \Rightarrow [y]_{y(x_i)}^{y(x_{i+1})} = \int_{x_i}^{x_{i+1}} f(x, y)dx \Rightarrow y(x_{i+1}) = y(x_i) + \int_{x_i}^{x_{i+1}} f(x, y)dx$$

Entretanto, na solução numérica, o valor $y(x_i)$ não é conhecido. Usa-se então a aproximação yi e obtém-se outra aproximação $y_{i+1}$, A equação acima mostrada fica então:

$$y_{i+1} = y_i + \int_{x_i}^{x_{i+1}} f(x, y)dx$$

Como se admite f(x,y) constante no intervalo e igual ao seu valor no início deste, então f(x,y).h é igual a variação de y, a qual somada a $y_i$ resulta em $y_{i+1}$, daí vem que para $x_i$, tem-se $f(x_i,y_i)$ e...

$$y_{i+1} = y_i + f(x_i, y_i) \int_{x_i}^{x_{i+1}} dx \Rightarrow y_i + f(x_i, y_i).(x_{i+1} - x_i) =$$

$$y_{i+1} = y_i + f(x_i, y_i).h$$

Onde ***h*** é o valor do subintervalo de x. Para um h suficientemente pequeno ($h \rightarrow 0$), o erro cometido pelo método de Euler é proporcional ao próprio ***h***, de maneira que quanto menor ***h***, melhor deveria ser a solução. Na prática tal não ocorre porque a redução de ***h*** implica em um aumento do número de cálculos e, a partir de um certo ponto a propagação do erro de arredondamento anula a vantagem do ***h*** pequeno.

Exemplo:
Seja a equação abaixo, cuja condição inicial y(0) =0

$$5\frac{dy}{dx}+2y=10 \qquad \text{e cuja solução analítica é} \quad y=\frac{10}{2}(1-e^{-\frac{2}{5}x})$$

Rearranjando a equação acima, vem:

$$\frac{dy}{dx}=\frac{10-2y}{5}=0{,}4(5-y)$$

Aplicando o método de Euler, pode-se escrever...

$$y_{i+1}=y_i+0{,}4.(5-y_i).h$$

Observar que 0,4(5-y) é o valor da derivada da função primitiva que não se conhece. Atribui-se um valor para h, e se começa pelo valor da condição inicial, obtendo-se... Para h=1, lembrando que $h = x_1-x_0$, no entanto não se tem intervalo um intervalo fixo, apenas a condição inicial mostra que um dos pontos é 0, e neste ponto o valor de y é igual a zero. No ponto seguinte, vem...

$$y_1=0+0{,}4\times(5-0)\times 1=2{,}0000$$

E no ponto seguinte, vem...

$$y_2=2+0{,}4\times(5-2)\times 1=3{,}2000$$

$$y_3=3{,}2+0{,}4\times(5-2)\times 1=3{,}9200$$

Continua...

Colocando os valores em tabela e comparando-os com os valores exatos obtidos pela solução analítica, tem-se:

| x | y(h=1) | y(h=0,1) | y(h=0,01) | y(h=0,001) | $y_{exato}$ |
|---|---|---|---|---|---|
| 0 | 0,00000 | 0,00000 | 0,00000 | 0,00000 | 0,00000 |
| 1 | 2,00000 | 1,67584 | 1,65109 | 1,64867 | 1,64840 |
| 2 | 3,20000 | 2,78999 | 2,75696 | 2,75371 | 2,75336 |
| 3 | 3,92000 | 3,53071 | 3,49765 | 3,49439 | 3,49403 |
| 4 | 4,35200 | 4,02317 | 3,99375 | 3,99084 | 3,99052 |
| 5 | 4,61120 | 4,35057 | 4,32603 | 4,32359 | 4,32332 |
| 6 | 4,76672 | 4,56824 | 4,54859 | 4,54663 | 4,54641 |
| 7 | 4,86003 | 4,71295 | 4,69765 | 4,69612 | 4,69595 |
| 8 | 4,91602 | 4,80916 | 4,79749 | 4,79632 | 4,79619 |
| 9 | 4,94961 | 4,87312 | 4,86436 | 4,86349 | 4,86338 |
| 10 | 4,96977 | 4,91565 | 4,90915 | 4,90849 | 4,90842 |

*Atentar ao fato de que para passar de x=1 para x=2 na primeira coluna foi necessário apenas um cálculo porque h=1. Na segunda coluna foram necessários 10 cálculos consecutivos, cujos resultados não foram anotados e na terceira coluna foram necessários 100 cálculos dos quais anotou-se somente aqueles que passaram pelos valores inteiros de x. Ou seja, quanto menor o número de intervalos, mais pontos devem ser calculados dentro de um mesmo intervalo.*

*Métodos de Runge-Kutta***.** O método de Euler não é tão preciso, pois baseia-se na derivada no ponto inicial e um sub-intervalo sempre admite pelo menos duas derivadas diferentes, ou seja, a do ponto inicial do subintervalo e a do ponto final, desprezando-se qualquer variação da derivada nos pontos intermediários a este intervalo. Os métodos de Runge-Kutta baseiam-se em tomar uma média entre os dois pontos do subintervalo, entretanto, o ponto y seguinte à condição inicial é inicialmente ***desconhecido*** exigindo então uma ***estimativa*** para se calcular a

inclinação $f(x_{i+1},y^*_{i+1})$, onde $y^*_{i+1}$ é o valor da função estimado. Os métodos de Runge-Kutta são denominados de métodos de *previsão-correção*, nos quais, no caso mais simples, se faz:

- Inicialmente se obtém uma estimativa $y^*_{i+1}$ do próximo valor da função primitiva (solução da equação);
- Usando-se este valor, calcula-se a inclinação $f(x_{i+1}, y^*_{i+1})$;
- Calcula-se uma inclinação média usando os valores de $f(x_i,y_i)$ e $f(x_{i+1},y^*_{i+1})$;
- Calcula-se então o valor definitivo de $y_{i+1}$.

A maneira de se obter a primeira estimativa e a maneira de obter a média definem os vários métodos de Runge-Kutta. Alguns têm erros proporcionais a h2 quando $h \rightarrow 0$, outros têm erros proporcionais a h4, enquanto outros tem erros proporcionais a h6, sendo então denominados de Runge-kutta de 2ª ordem, 4ª ordem, 6ª ordem, etc.

Um método de Runge-Kutta de 2ª ordem é o método de Euler modificado e aperfeiçoado onde se faz a estimativa do próximo ponto usado o método de Euler e com ele se faz a média aritmética das inclinações em $y_i$ e em $y^*_{i+1}$, e com esta média usar novamente o método de Euler para corrigir o valor $y_{i+1}$. Este método é conhecido pelo nome de *método de Heun*.

Exemplo:
Resolver a mesma equação diferencial do exemplo anterior, agora pelo método de Heun.

$y_0 = 0{,}0000$

$$y^*_1 = 0 + 0{,}4 \times (5-0) \times 1 = 2{,}0000$$
$$f(x_1, y_1) = 0{,}4 \times (5-0) \times 1 = 2{,}000$$
$$f(x_1, y^*_1) = 0{,}4 \times (5-2) \times 1 = 1{,}200$$
$$y_1 = 0 + \frac{2+1{,}2}{2} \times 1 = 1{,}60000$$

$$y^*_2 = 1{,}6 + 0{,}4 \times (5-1{,}6) \times 1 = 2{,}9600$$
$$f(x_2, y_2) = 0{,}4 \times (5-1{,}6) \times 1 = 1{,}36$$
$$f(x_2, y^*_2) = 0{,}4 \times (5-2{,}96) \times 1 = 0{,}816$$
$$y_2 = 1{,}60 + \frac{1{,}36+0{,}816}{2} \times 1 = 2{,}688$$

$$y^*_3 = 2{,}688 + 0{,}4 \times (5-2{,}688) \times 1 = 3{,}6128$$
$$f(x_3, y_3) = 0{,}4 \times (5-2{,}688) \times 1 = 0{,}9248$$
$$f(x_3, y^*_3) = 0{,}4 \times (5-3{,}6128) \times 1 = 0{,}55488$$
$$y_2 = 2{,}688 + \frac{0{,}9248+0{,}55488}{2} \times 1 = 3{,}42784$$

| x | y(h=1) | y(h=0,1) | $y_{exato}$ |
|---|---|---|---|
| 0 | 0,00000 | 0,00000 | 0,00000 |
| 11 | 1,60000 | 1,64803 | 1,64840 |
| 12 | 2,68800 | 2,75286 | 2,75336 |
| 13 | 3,42784 | 3,49353 | 3,49403 |
| 14 | 3,93093 | 3,99007 | 3,99052 |
| 15 | 4,27303 | 4,32295 | 4,32332 |
| 16 | 4,50566 | 4,54611 | 4,54641 |
| 17 | 4,66385 | 4,69572 | 4,69595 |
| 18 | 4,77142 | 4,79601 | 4,79619 |
| 19 | 4,84456 | 4,86325 | 4,86338 |
| 20 | 4,89430 | 4,90832 | 4,90842 |

*Como se pode perceber, a solução é muito melhor do que aquela obtida pelo método anterior. Mesmo com h=1,0 já se obtém dois algarismos significativos corretos.*

A título de apresentação segue a fórmula para um método Runge-Kutta de 4ª ordem. O mesmo pode ser usado quando se deseja alta precisão, mas exige cálculos um pouco mais trabalhosos, os quais podem ser muito bem executado através de programação em computadores.

$$y_{i+1} = y_i + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$ Onde:

$$k_1 = f(x_i, y_i)$$

$$k_2 = f(x_i + \frac{h}{2}, y_i + \frac{hk_1}{2})$$

$$k_3 = f(x_i + \frac{h}{2}, y_i + \frac{hk_2}{2})$$

$$k_4 = f(x_i + h, y_i + hk_3)$$

Considerações finais:
O método de Euler e os métodos de Runge-Kutta têm a característica comum de que os valores a obter na etapa i+1 dependem unicamente dos valores obtidos na etapa i. São, por isso, chamados métodos de uma etapa. Em essência, todos eles consistem em se fazer uma extrapolação linear, sendo a reta de extrapolação definida por um ponto (o ponto $x_i$,$y_i$ onde já se chegou, e uma direção). As diferenças entre eles consistem do modo de se obter esta direção. Sua grande vantagem é que, dadas às condições iniciais das equações diferenciais, pode-se iniciar a solução imediatamente.
Houve, entretanto, várias tentativas de se definir métodos mais eficientes através do uso da extrapolação com polinômios de grau mais elevado. Em todos estes novos métodos, partindo-se de n valores da função solução, interpola-se por eles um polinômio de grau n-1 e usa-se este para extrapolar para o próximo ponto. Em vários métodos também se usa o conceito de prever-corrigir, aperfeiçoando o valor obtido por interpolação através de uma fórmula corretora. Entre estes métodos estão os métodos de ADAMS-BASHFORD que usa interpolação através de um polinômio de 3º grau entre os pontos usados; O método de ADAMS-MOULTON que melhora o anterior, por este não ser muito preciso; O método de MILNE, que usa interpolação de 2º grau associado à técnica de prever-corrigir, exigindo menor número de cálculos e o método de HAMMING que é muito semelhante ao de Milne sendo este um dos mais populares, ao lado do Runge-Kutta de 4º ordem, para a solução de equações diferenciais.
A desvantagem destes últimos métodos é que eles não podem ser aplicados diretamente à condição inicial, exigindo uma série de outros valores para calcular o próximo, exigindo, portanto um número maior de operações e tem precisão comparável ao método de Runge-Kutta de 4ª ordem, sendo, portanto este último o mais popular entre todos.

**Exercícios:**

1. $\frac{dy}{dx} = 3$ y(0) = 3

2. $\frac{dy}{dx} = 2y$ y(1)=0

3. $\frac{dy}{dx} = 5x + 3$ y(1)=8

4. $y^3 + 4.\frac{dy}{dx} = 8$ y(0)=0

5. $\frac{dy}{dx} = y^2 + 1$ y(0)=0

## Desenvolvimento de modelos matemáticos

A matemática em si é uma linguagem como qualquer outra (português, inglês, música, etc.) e é composta de símbolos com a finalidade de nos dizer o que se deve fazer seqüencialmente para ordenar um raciocínio, com o objetivo de chegar a uma conclusão. Assim, a matemática é um poderoso instrumento intelectual que através da abstração e formalização, sintetiza idéias as quais poderiam ser demoradamente explicadas usando o bom português, exigindo quilômetros de leitura. Embora semelhantes muitas dessas abstrações surgem em situações as mais diversas e por isto mesmo camufladas em sua existência. O objetivo é então extrair essa essência. O desenvolvimento de modelos matemáticos para a representação teórica de um fenômeno físico, químico, biológico, de economia ou de engenharia é um processo que compreende as seguintes etapas:

- Experimentação e coleta de dados;
- Abstração;
- Resolução;
- Validação;
- Modificação;
- Aplicação.

*Obtenção de dados experimentais ou empíricos*. Esta etapa se preocupa da coleta de dados relevantes que ajudam na compreensão do problema, na modificação do modelo e na decisão de sua validade. É um processo essencialmente laboratorial e/ou estatístico;

*Abstração*. Processo de seleção das variáveis essenciais e formulação em linguagem natural do problema ou da situação real;

*Resolução*. O modelo matemático é montado quando se substitui a linguagem natural pela linguagem matemática. O estudo do modelo depende de sua complexidade e pode ser um processo numérico. Quando os argumentos conhecidos não são eficientes novos métodos podem ser criados ou então o modelo deve ser modificado;

*Validação*. É a comparação entre a solução obtida via resolução do modelo matemático e os dados reais. É um processo de decisão de aceitação ou não do modelo inicial. O grau de aproximação desejado será o fator preponderante na decisão;

*Modificação*. Caso o grau de aproximação entre os dados reais e a solução do modelo não seja aceito, deve0-se modificar as variáveis ou a lei de formação, e com isso o próprio modelo original é modificado e o processo se inicia novamente;

*Aplicação*. A modelagem eficiente permite fazer previsões, tomar decisões, explicar e entender e enfim, participar do mundo real com capacidade de influenciar em suas mudanças.

**Sistemas de equações diferenciais e equações diferenciais de ordens superiores**

- Sistemas de equações diferenciais de 1ª ordem
- Equações diferenciais de ordens superiores

*Solução numérica de sistemas equações diferenciais de 1ª ordem.* Um sistema de equações diferenciais de 1ª ordem pode ser representado como mostrado abaixo, juntamente com as condições iniciais:

$$\left\{\begin{array}{l} \dfrac{dy}{dx} = f_1(x, y_1, y_2, y_3, ....., y_n) \\ \dfrac{dy}{dx} = f_2(x, y_1, y_2, y_3, ....., y_n) \\ ..... \\ \dfrac{dy}{dx} = f_n(x, y_1, y_2, y_3, ....., y_n) \end{array}\right\} \qquad \left\{\begin{array}{l} y_1(0) = y_{10} \\ y_2(0) = y_{20} \\ ..... \\ y_n(0) = y_{n0} \end{array}\right\}$$

Sistemas de equações diferenciais de 1ª ordem são sistemas do tipo em que aparecem somente derivadas de 1ª ordem, associadas as suas condições iniciais. Para se obter a solução simultânea destas equações, pode-se aplicar qualquer um dos métodos já vistos, simplesmente usando os dados obtidos em cada equação seqüencialmente até que uma etapa esteja completada em todas as equações. Aí, repete-se para a próxima etapa, e assim, sucessivamente.

*Solução numérica de sistemas de equações diferenciais de ordens superiores.* Equações diferenciais de ordens superiores são aquelas em que aparecem derivadas de segunda ordem ou ordens maiores. É importante observar que todas elas podem ser transformadas em sistemas de equações de 1ª ordem do tipo acima por simples mudança de variáveis. Veja o exemplo sobre uma equação genérica abaixo:

$$\frac{d^n y}{dx^n} + a_1 . \frac{d^{n-1} y}{dx^{n-1}} + a_2 . \frac{d^{n-2} y}{dx^{n-2}} + ..... + a_{n-1} . \frac{dy}{dx} + a_n . y = F(x, y)$$

Fazendo as mudanças nas variáveis substituindo-as pelas equivalentes abaixo:

$$\left\{\begin{array}{c} y_0 = y \\ y_1 = \dfrac{dy}{dx} \\ y_2 = \dfrac{d^2 y}{dx^2} \\ ..... \\ y_{n-1} = \dfrac{d^{n-1} y}{dx^{n-1}} \end{array}\right\} \qquad \text{vem:} \qquad \left\{\begin{array}{l} \dfrac{dy_0}{dx} = y_1 \\ \dfrac{dy_1}{dx} = y_2 \\ \dfrac{dy_2}{dx} = y_3 \\ ..... \\ \dfrac{dy_{n-2}}{dx} = y_{n-1} \\ \dfrac{dy1}{dx} = F(x, y_0) - a_1 y_{n-1} - a_2 y_{n-2} - ..... - a_{n-1} y_1 - a_n y_0 \end{array}\right\}$$

Exemplo:
Qualquer equação de segunda ordem ou maior pode ser reduzida a um sistema de equações simultâneas de primeira ordem pela introdução de variáveis auxiliares. Como exemplo considerar as equações abaixo:

$$\frac{d^2x}{dt^2} + xy.\frac{dx}{dt} + z = e^z$$

$$\frac{d^2y}{dt^2} + zy.\frac{dy}{dt} = 7 + t^2$$

$$\frac{d^2z}{dt^2} + xz.\frac{dz}{dt} + x = e^x$$

Ao introduzir novas variáveis de substituição, mostradas abaixo:

$$x_1 = x \qquad x_4 = \frac{dx_1}{dt}$$

$$x_2 = y \qquad x_5 = \frac{dx_2}{dt}$$

$$x_3 = z \qquad x_6 = \frac{dx_3}{dt}$$

Após a devida substituição vem:

$$\frac{dx}{dt} + x.x_2.x_4 + x_3 = e^{x_3} \Rightarrow \frac{dx}{dt} = -x.x_2.x_4 - x_3 + e^{x_3}$$

$$\frac{dx_5}{dt} + x_3.x_2.x_5 = 7 + t^2 \Rightarrow \frac{dx_5}{dt} = -x_3.x_2.x_5 + 7 + t^2$$

$$\frac{dx_6}{dt} + x_1x_3.x_6 - x = e^{x_1} \Rightarrow \frac{dx_6}{dt} - x_1x_3.x_6 - x + e^{x_1}$$

As quais têm a forma geral igual a:

$$\frac{dy}{dt} = f_i(t, x_1, x_2, x_3, ....., x_n) \quad \text{i= 1, 2, 3, .....n}$$

Tal sistema pode ser resolvido pela aplicação simultânea de qualquer um dos métodos para resolução de equações diferenciais de primeira ordem.

Exemplo:
Seja a equação abaixo:

$$\frac{d^2x}{dt^2} - 2.(1-x^2).\frac{dx}{dt} + x = 0$$ com as condições iniciais: $$x(0) = 0{,}02 \qquad \frac{dx}{dt}(0) = 0$$

Fazendo as substituições de variáveis, vem:

$$\left\{ \begin{array}{l} \frac{dx}{dt} = y \\ \frac{dy}{dt} = 2.(1-x^2).y - x \end{array} \right\}$$ e as condições iniciais passam para: $$x(0) = 0{,}02 \qquad y(0) = 0$$

Resolvendo pelo método de Euler, com h=0,1 e 0,01 para efeito de comparação obtém os seguintes valores os quais poderão ser confirmados pelo aluno.

| t | x(h=0,1) | x(h=0,01) | dx/dt(h=0,1) | dx/dt(h=0,01) |
|---|---|---|---|---|
| 0.0000 | 0.0200 | 0.0200 | 0.0000 | 0.0000 |
| 0.5000 | 0.0176 | 0.0166 | -0.0146 | -0.0163 |
| 1.0000 | -0.0047 | 0.0005 | -0.0471 | -0.0535 |
| 1.5000 | -0.0304 | -0.0431 | -0.1139 | -0.1320 |
| 2.0000 | -0.1100 | -0.1430 | -0.2437 | -0.2871 |
| 2.5000 | -0.2737 | -0.3475 | -0.4773 | -0.5544 |
| 3.0000 | -0.5757 | -0.6951 | -0.7928 | -0.8029 |

O sistema é do tipo não linear e, por isso, é necessária maior precisão na solução. Por isso são mostradas soluções com h=0,1 e h=0,01 onde se pode muito bem ver que ***ainda não coincidência de nenhuma casa***. Na solução com h=0,01 foram anotados somente os valores a cada cinqüenta etapas. Como pose claramente se verificar este tipo de cálculo exige naturalmente um meio computacional para ser realizado.

Este mesmo problema resolvido pelo método de Heun fica:

| t | x(h=0,01) | dx/dt(h=0,01) | *x(h=0,01) | *dx/dt(h=0,01) |
|---|---|---|---|---|
| 0.0000 | 0.0200 | 0.0000 | 0.0200 | 0.0000 |
| 0.5000 | 0.0165 | -0.0165 | 0.0165 | -0.0165 |
| 1.0000 | 0.0001 | -0.0542 | -0.0443 | -0.1343 |
| 1.5000 | -0.0445 | -0.1336 | -0.1473 | -0.2926 |
| 2.0000 | -0.1463 | -0.2901 | 0.3568 | -0.5630 |
| 2.5000 | -0.3537 | -0.5565 | -0.7079 | -0.7980 |
| 3.0000 | -0.7003 | -0.7875 | -1.0594 | -0.4966 |

As colunas marcadas com * referem-se a solução do mesmo problema usando o método Runge-Kutta de 4ª ordem. Os resultados pelos três métodos não concordam bem em valores numéricos, mas representam um comportamento semelhante. As diferenças se devem à maior propagação de erros causada pela não linearidade da equação.

*As soluções numéricas para **equações diferenciais parciais** envolvem técnicas de diferenças finitas e são um pouco mais complexas. Se for do interesse, deve-se consultar literatura específica sobre o assunto.*

**Solução numérica de equações integrais**

Para a solução de equações integrais, as quais se apresentam sob a forma semelhantes àquelas mostradas abaixo:

$$u(x) = f(x) + \lambda.\int_a^b k(x,t)u(t).dt$$

Como a integral pode ser obtida numericamente através de qualquer um dos métodos existentes para solução de integrais definidas, pode-se então, se escrever a mesma equação assim:

$$u(x) = f(x) + \lambda(b-a).\left[\sum_{i=1}^{n} c_i.k(x,t_i).u(t_i)\right]$$

Onde $t_1$, $t_2$, $t_3$,....., $t_n$, são os pontos da subdivisão do eixo t, onde $a \le t \le b$ e os valores $c_i$ são coeficientes que dependem do tipo de fórmula de integração usada. A função u(x) deve levar em conta todos os valores possíveis, os quais inicialmente se atribuem como sendo os valores de t e a fazer isto sucessivamente para cada valor de t e fazendo $u(t_i)=u_i$ e $f(t_i)=f_i$, recai-se num sistema de n equações a n incógnitas, e assim, tem-se:

$$u_i = f_i + (b-a)[c_1.k(t_i,t_1)u_1 + c_2.k(t_i,t_2)u_2 + ..... + c_n.k(t_i,t_n)u_n] \qquad i=1,2,...., n$$

Exemplo:
Resolver numericamente a equação:

$$u(x) = x + \frac{1}{3}.\int_0^1 (t+x)u(t).dt$$

Esta equação diz que u é uma função que só depende de x e que por sua vez é igual a soma da própria função com a integral de um produto entre outra função dependente de x e de t e a mesma função u, mas dependente de t ao invés de x. a solução desta equação integral irá encontrar a função u dependente apenas de x.

Neste exemplo, a=0, b=1. Fazendo n=3, isto implica em fazer t1=0, t2=0,5 e t3=1. Usando a fórmula de Simpson para fazer a integração (interpolação parabólica para a integração), vem:

$$u(x) = x + \frac{1}{3}\left[\frac{0,5}{3}(t_1+x)u(t_1) + 4(t_2+x)u(t_2) + (t_3+x)u(t_3)\right]$$

$$u(x) = x + \frac{1}{18}\left[(t_1+x)u(t_1) + 4(t_2+x)u(t_2) + (t_3+x)u(t_3)\right]$$

A integração da função u(t) é obtida usando a equação acima u(x) onde x é substituído por t em todos os subintervalos, tal que $0 \le x \le 1$. Fazendo uma integração para $x=t_1=0$, outra para $x=t_2=0,5$ e outra para $x=t_3=1$, resulta em 3 equações:

$$u(t_1) = t_1 + \frac{1}{18}\left[(2t_1).u(t_1) + 4(t_2+t_1)u(t_2) + (t_3+t_1)u(t_3)\right]$$

$$u(t_2) = t_2 + \frac{1}{18}\left[(t_1+t_2).u(t_1) + 4(2.t_2).u(t_2) + (t_3+t_2)u(t_3)\right]$$

$$u(t_3) = t_3 + \frac{1}{18}\left[(t_1+t_3).u(t_1) + 4(t_2+t_3)u(t_2) + (2.t_3)u(t_3)\right]$$

Colocando os valores para $t_1$, $t_2$ e $t_3$ e fazendo $u(t_i)= u_i$ as equações assumem as formas:

$$u_1 = 0 + \frac{1}{18}\left[4(0,5)u_2 + 1u_3\right]$$

$$u_2 = 0,5 + \frac{1}{18}\left[0,5.u_1 + 4(2\times 0,5).u_2 + (1+0,5)u_3\right]$$

$$u_3 = 1 + \frac{1}{18}\left[(0+1).u_1 + 4(0,5+1)u_2 + (2\times 1)u_3\right]$$

Resultando no seguinte sistema de 3 equações, três incógnitas

$$18u_1 - 2u_2 - 1u_3 = 0$$

$$-u_1 + 28u_2 - 3u_3 = 18$$

$$-u_1 - 6u_2 + 16u_3 = 18$$

Cuja solução obtida por um método de solução de equações lineares qualquer, fornece:

$$u_1 = \frac{12}{71} \qquad u_2 = \frac{57}{71} \qquad u_3 = \frac{102}{71}$$

Assim a solução para a equação integral fica:

$$u(x) = x + \frac{1}{18}\left[(0+x)\frac{12}{71} + 4(\frac{1}{2}+x)\frac{57}{71} + (1+x)\frac{102}{71}\right]$$

$$u(x) = \frac{90}{71}x + \frac{12}{71}$$

Observar que as funções u(t) que eram desconhecidas foram calculadas de uma forma iterativa através da própria fórmula da equação integral e depois resolvida simultaneamente.

**Exercícios:**

***Resolver o sistema de equações*** *diferenciais* ***de 1ª ordem:***

Um sistema hidráulico composto por dois reservatórios de água onde um reservatório através de um orifício D1 despeja água no reservatório 2 e este por sua vez tem também um orifício dando vazão para o meio exterior. Considerando a altura no primeiro reservatório como sendo $h_1$ e no segundo $h_2$ e sendo a área do dois igual a A, ambos simultaneamente têm as seguintes equações diferenciais que regem o comportamento deste conjunto:

$$A\frac{dh_1}{dt} + \frac{h_1}{D_1} = 0$$

$$A\frac{dh_2}{dt} + \frac{h_2}{D_2} = \frac{h_1}{D_1}$$

Onde as condições iniciais $h_1(0)= 1,0$ m e $h_2(0)=0,3$ m. Considerar como dados complementares os valores: A= $2m^2$, $D_1=0,9s/m^2$ e $D_2=0,85s/m^2$

*A primeira equação diz que o produto da área do reservatório multiplicada por dh e dividida este produto por dt, fornece a vazão que sai do reservatório num tempo infinitesimal e esta vazão é igual à razão entre a altura atual e a resistência que o orifício opõe à saída da água. Esta razão utiliza o modelo matemático empregado na eletricidade onde E=R.i, sendo E equivalente à pressão, que neste problema é expressa em metros de coluna de água. i é equivalente à vazão, que neste problema a mesma usa a unidade $m^3/s$. Fazendo h=D.Q, que por analogia, h≡E, D≡R e Q≡i, vem que D é então dado em $s/m^2$, sendo esta a unidade da resistência que o orifício exerce sobre o fluxo de água para fora do reservatório.*

Resolver os sistemas abaixo:

2. $\frac{dy}{dx} = z$
$\frac{dz}{dx} = 3z - 2y$ — y(0) = -1 e z(0)= 0

3. $\frac{dy}{dx} = z$
$\frac{dz}{dx} = w$
$\frac{dw}{dx} = x^2 y - 4z + 2xw + 1$ — y(0)=1, z(0)=2 e w(0)=3

Resolver as equações de ordens superiores abaixo:

4. $\frac{d^2y}{dx^2} - y = x$ — onde y(0) = 0 e y'(0) = 1 (use o método de Euler com h = 0,1)

5. $\frac{d^2y}{dx^2} - 3\frac{dy}{dx} + 2y = 0$ — onde y (0) =-1 e y'(0) =0 (use o método de Heun com h = 0,1)

6. $\frac{d^3y}{dx^3} + 2.\frac{dy}{dx} - y = 10$ — onde y(0) = 1 e y'(0) = 0,5

Resolver as equações integrais:

7. $S(x) = x^2 + \int_0^1 2t^2 dt$

8. $S(x) = x^2 + \int_0^1 2(t+x)t^2 dt$

9. $S(x) = x^2 + \int_0^5 (x-z)^2.(\frac{3z}{1+z})dz$

10. $F(t) = e^t + 2\int_1^2 (ty + y^2)dy$

Resolver as equações acima usando o método de integração de Simpson, subdividindo os intervalos de ***t*** em pelo menos **2** subintervalos nos exercícios 7 e 8. No exercício 9, subdividir **z** em pelo menos **8** intervalos. Isto irá resultar num sistema de 8 equações, 8 incógnitas. Para resolver este sistema fazer uso dos programas Gauss-Jordan ou relaxação, já disponíveis. O exercício 10 fica com a escolha ao encargo do aluno.

**Otimização 1**

- Otimização analítica
- Otimização numérica envolvendo uma única variável
- Otimização numérica envolvendo 2 ou mais variáveis

Por otimização se entende encontrar a melhor maneira de fazer as coisas. Em termos matemáticos, o termo ***ótimo*** corresponde sempre a um ponto de ***máximo***.
A otimização pode ser ***estática*** ou ***dinâmica***. É estática quando procura otimizar operações e processos às condições mais freqüentes de operação, sem considerar as variações. Em engenharia, isto consiste em encontrar tamanhos ótimos de equipamento e volume de produção e composição de produtos em função de variáveis como temperatura, pressão, vazão, custo, etc. É dinâmica quando se deseja incluir as flutuações que uma primeira otimização provoca no processo. Este tipo de otimização requer o conhecimento das características dinâmicas do equipamento e do processo e também necessita de predição de como a mudança de condições pode ser corrigida. Na realidade a otimização dinâmica é uma extensão à análise do controle automático do processo.

*Otimização analítica***.** Quando um processo pode ser descrito com um número pequeno de variáveis, é possível encontrar um ponto de ***máximo*** ou de ***mínimo*** através da diferenciação, igualando a derivada a zero e resolvendo a equação resultante. Entretanto, quando o número de variáveis independentes é grande, este método se torna proibitivo devido a sua complexidade e dificuldade. Para pequenas quantidades de variáveis observar as regras abaixo:
Se a função matemática que relaciona uma variável independente é conhecida, então num ponto ***P*** onde a derivada primeira é nula existe um ponto de máximo se a derivada segunda for menor que zero; um ponto de mínimo se esta for maior que zero ou um ponto de inflexão se a derivada segunda for igual a zero e a derivada terceira for diferente de zero. Neste último caso, é necessário verificar a derivada terceira, pois se a mesma for igual a zero, este ponto é na verdade um ponto de máximo. Em termos matemáticos resumindo tudo, pode-se dizer:

1. Para funções dependentes de uma única variável independente.

   Seja a função: $y = f(x)$ e um ponto ***P*** onde

   $$\frac{dy}{dx} = 0 \text{ e } \frac{d^2y}{dx^2} < 0 \quad \Rightarrow \quad \text{ponto de máximo}$$

   $$\frac{dy}{dx} = 0 \text{ e } \frac{d^2y}{dx^2} > 0 \quad \Rightarrow \quad \text{ponto de mínimo}$$

   $$\frac{dy}{dx} = 0 \text{ e } \frac{d^2y}{dx^2} = 0 \text{ e } \frac{d^3y}{dx^3} \neq 0 \quad \Rightarrow \quad \text{ponto de inflexão}$$

   $$\frac{dy}{dx} = 0 \text{ e } \frac{d^2y}{dx^2} = 0 \text{ e } \frac{d^3y}{dx^3} = 0 \text{ e } \frac{d^4y}{dx^4} \neq 0 \quad \Rightarrow \quad \text{ponto de máximo}$$

2. Se a função depender de mais de uma variável, se z = f(x,y) e um ponto ***P*** onde

$$\frac{\partial z}{\partial x} = 0 \qquad \text{e} \qquad \frac{\partial z}{\partial y} = 0 \qquad \text{então:}$$

***P*** é um ponto de máximo se $\frac{\partial^2 z}{\partial x^2} < 0$ e $\frac{\partial^2 z}{\partial x^2} \times \frac{\partial^2 z}{\partial y^2} - \left(\frac{\partial^2 z}{\partial x dy}\right)^2 > 0$

***P*** é um ponto de mínimo se $\frac{\partial^2 z}{\partial x^2} > 0$ e $\frac{\partial^2 z}{\partial x^2} \times \frac{\partial^2 z}{\partial y^2} - \left(\frac{\partial^2 z}{\partial x dy}\right)^2 > 0$

***P*** é ponto de sela se $\frac{\partial^2 z}{\partial x^2} \forall real$ e $\frac{\partial^2 z}{\partial x^2} \times \frac{\partial^2 z}{\partial y^2} - \left(\frac{\partial^2 z}{\partial x dy}\right)^2 < 0$

e ***P*** deve ser melhor investigado se $\frac{\partial^2 z}{\partial x^2} \times \frac{\partial^2 z}{\partial y^2} - \left(\frac{\partial^2 z}{\partial x dy}\right)^2 = 0$

Exemplo: (para z = f (x,y) )
Encontrar os pontos críticos na função: $z = x^3 - 6xy + y^2 + 5y$ e classificar cada ponto como ponto de máximo, ponto de mínimo ou ponto de sela.

Derivando a função com relação a ***x*** vem: $\frac{\partial z}{\partial x} = 3x^2 - 6y$

Derivando a função com relação a ***y*** vem: $\frac{\partial z}{\partial y} = -6x + 2y + 5$

Igualando ambas a zero, encontram-se os pontos x, y que satisfazem o sistema sendo: (1, 0,5) e (5 e 12,5). São dois os pontos porque uma das equações do sistema é de segunda ordem**.**

Derivando mais uma vez as derivadas parciais primeiras vem: $\frac{\partial^2 z}{\partial x^2} = 6x$ e $\frac{\partial^2 z}{\partial y^2} = 2$

E derivando z em função de x e y simultaneamente vem: $\frac{\partial^2 z}{\partial x \partial y} = -6$

Executando o teste $\frac{\partial^2 z}{\partial x^2} \times \frac{\partial^2 z}{\partial y^2} - \left(\frac{\partial^2 z}{\partial x \partial y}\right)^2$

No ponto 1 (1, 0,5) vem $D_1 = 12 - 36 = -24$ < 0 (sela)
No ponto 2 (5, 12,5) vem $D_2 = 60 - 36 = 24$ > 0 (min)

3. Se ***z*** for uma função dependente de mais de duas variáveis independentes, ou seja, $z = f(x_1, x_2, .....x_n)$, e....para simplificar a notação adota-se a forma abaixo

$$\frac{\partial z}{\partial x_i} = z_i' \quad \text{e} \quad \frac{\partial^2 z}{\partial x_i \partial x_j} = z_{ij}''$$

Então, se $z_i' = 0$ calcular $H = \begin{bmatrix} z_{11}'' & z_{12}'' & .... & z_{1n}'' \\ .... & .... & .... & .... \\ z_{n1}'' & z_{n2}'' & .... & z_{nn}'' \end{bmatrix}$

Conforme o resultado vem:
***P*** é um ponto de máximo se H é positivo definido
***P*** é um ponto de mínimo se H é negativo definido
***P*** é um vale ou uma crista se H é semidefinido
***P*** é uma sela se H é indefinido

Condições restritivas:
As restrições aplicadas às variáveis independentes restringem o domínio da função, e são consideradas então como condições restritivas. As restrições podem ser de várias maneiras, como:

- $a < x < b$ A variável independente x fica restrito entre dois pontos;
- $x + y = 5$ A soma de duas ou mais variáveis independentes devem atender um valor específico;
- $x + y < 10$ A soma entre várias variáveis independentes devem ficar restritas a um valor máximo;

*Ao resolver um problema de otimização, as condições restritivas devem ser todas atendidas. É importante observar também que as restrições facilitam a resolução do problema, pois diminuem ou restringem a área de pesquisa.*

*Otimização numérica envolvendo uma única variável (método numérico).* Encontrar um ponto ótimo numa função envolve pesquisa ao longo de um intervalo preestabelecido da variável independente. Se a função for conhecida e bem determinada, uma simples derivação da função e fazendo-a igual a zero incidirá numa equação que pode ser resolvida através de métodos já vistos (*obtenção dos zeros de função, usando os métodos para funções não lineares e transcendentes*). Se a função não for conhecida como é o caso quando se está se obtendo resultados através de experimentação ou se for complexa demais para ser derivada, pode-se fazer a pesquisa diretamente. Existem métodos simultâneos e métodos seqüenciais, sendo estes últimos mais eficientes. O princípio da pesquisa se baseia no fato de ao se realizar dois experimentos estabelecendo-se uma pequena diferença ε entre os dois pontos a serem experimentados, ***o ponto que estiver mais próximo do ponto ótimo será em termos absoluto, o maior entre os dois.*** Assim, uma vez estabelecido o intervalo de pesquisa, pode-se realizar dois experimentos a cada vez, de forma seqüencial, selecionando sempre aquele que apresentar o maior valor absoluto, ou fazer distinção entre o maior e o menor no caso de se desejar um ponto máximo ou um ponto mínimo. (escolher sempre o maior se for desejado encontrar um ponto de máximo e o menor se for desejado encontrar um ponto de mínimo). O problema que surge é escolher por onde começar a pesquisa. Entre os métodos banais encontram-se a pesquisa aleatória e as seqüenciais puras, ambas ineficientes. A primeira depende muito da sorte e a segunda envolve uma série interminável de experimentos ou cálculos porque parte do princípio de iniciar num extremo do intervalo e ir incrementando o mesmo e ir testando até encontrar o ponto desejado. Dois outros métodos são mais interessantes. O ***método dicotômico*** e o ***método de Fibonacci***. Ambos partem do princípio de *dividir-e-conquistar*. O primeiro também é conhecido como método binário, pois envolve sempre a divisão por dois. Este método consiste em escolher dois pontos bastante próximos do centro, ou seja, dividir o intervalo de pesquisa em duas metades e escolher um ponto aquém e outro além do centro longe um do outro num valor ε e escolher o valor desejado, limitando agora o novo domínio de pesquisa imposto por este ponto e o extremo do intervalo oposto ao ponto encontrado (para o caso de um máximo). Fazendo isto, verifica-se que o intervalo de pesquisa fica reduzido à metade do intervalo inicial mais o valor ε/2. Repete-se agora, mais duas experimentações, da mesma maneira que anteriormente escolhendo-se o centro deste novo intervalo. ***A cada duas experimentações o intervalo de pesquisa se reduz à metade chegando rapidamente a valor desejado***. O *método de Fibonacci* é um pouco mais eficiente. Enquanto o método dicotômico reduz o intervalo de pesquisa na proporção de $2^n$, onde n é o número de tentativas (cada uma com dois experimentos), o que implica dizer que com 10 tentativas o intervalo fica reduzido a 1/1024 do valor original, o método de Fibonacci reduz este mesmo intervalo a 1/10946. A série de Fibonacci é caracterizada como sendo 0, 1, 1, 2, 3, 5, 8, 13, 21, 34,.....e como se pode observar o próximo valor é sempre obtido somando-se os dois imediatamente anteriores. Esta seqüência tem enumeras aplicações na ciência da computação, em matemática e na teoria dos jogos. O processo todo consiste em calcular o valor da função no valor $x_{inicial} = L_0$, onde $L_0$ é obtido pela expressão $L_0 = F_n^{-1}[ F_{n-1} (-1)^n].\varepsilon$.

Exemplo. (*Resolvido pelo método dicotômico*):
Dada a expressão:

$$y = 2x^3 - 7{,}5x^2 - 9x + 5$$

Pesquisar a existência de um máximo entre os pontos –5 < x ≤ 10, com ε = 0,2, sem restrições. O ponto de partida da pesquisa é obtido pela media do intervalo:

$$x_{n=1} = a + \frac{b-a}{2} = -5 + \frac{10-(-5)}{2} = -5 + 7{,}5 = 2{,}5$$

O valor x=2,5±0,1, correspondente ao centro do intervalo. Lembrar que ε é a distancia entre $x_{k+1}$ e $x_k$.

| | | | |
|---|---|---|---|
| n =1 | para | $x_1$ = 2,4 | y = -32,152 |
| | | $x_2$ = 2,6 | y = -33,948 |

Como se pode observar, $x_1$ é maior que $x_2$, portanto se deve eliminar o intervalo que vai de $\boldsymbol{x_2}$ a ***b***, restando o intervalo de ***a*** a $\boldsymbol{x_1}$. Assim o novo ponto a ser testado fica em:

$$x_{n=2} = a + \frac{x_1 - a}{2} = -5 + \frac{2{,}4-(-5)}{2} = -5 + 3{,}7 = -1{,}30$$

| | | | |
|---|---|---|---|
| n =2 | para | $x_3$ = -1,2 | y = 1,544 |
| | | $x_4$ = -1,4 | y = -2,588 |

Fazendo a verificação do maior valor percebe-se que x3 é o maior então o intervalo agora fica reduzido ao subintervalo entre $x_3$ e $x_1$, assim:

$$x_{n=3} = x_3 + \frac{x_1 - x_3}{2} = -1{,}2 + \frac{2{,}4-(-1{,}2)}{2} = -1{,}2 + 1{,}8 = 0{,}30$$

| | | | |
|---|---|---|---|
| n=3 | para | $x_5$ = 0,2 | y = 2,916 |
| | | $x_6$ = 0,4 | y = 0,328 |

***O intervalo de pesquisa agora ficará entre x3 e x5. Assim o novo ponto de pesquisa assume o valor –0,5.***

| | | | |
|---|---|---|---|
| n =4 | para | $x_7$ = -0,4 | y = 7,272 |
| | | $x_8$ = -0,6 | y = 7,268 |

Observar que tanto $x_7$ e $x_8$ são maiores que $x_5$ e $x_6$, assim o intervalo de pesquisa fica entre os valores de x = -0,4 e x = -0,6. Atentar para o fato que aqui agora se torna necessário diminuir o valor do intervalo para continuar, pois o valor do ponto máximo se encontra entre estes dois valores e esta diferença é justamente o valor de ε. Pode-se numa primeira aproximação admitir como sendo o ponto ótimo, o valor obtido pela média aritmética entre ambos. Efetuando este cálculo via álgebra encontra-se o ponto de máximo no ponto x = –0,5.

Exercícios**:**

1. Encontrar o ponto de mínimo na equação: $y = e^{4x^2}$
2. Pesquisar ponto de máximo ou mínimo na função abaixo $y = \ln x^2 - 6x$

*Otimização numérica envolvendo duas ou mais variáveis***.** A otimização de funções de duas variáveis ***sem restrições*** pode ser resolvida de forma semelhante aos métodos usados para funções de uma variável só. Apenas incrementam um pouco a complexidade porque tanto ***y*** como ***x*** devem ser resolvidos simultaneamente a cada passo. Ao se introduzir qualquer restrição estas devem ser atendidas em primeiro lugar. Quando as restrições são equações lineares se diz que o método é linear. *Para os problemas de otimização linear envolvendo mais de uma variável, e restrições do tipo inequação*, dois métodos são particularmente importantes e por si só constituem uma disciplina inteira, cada um deles. São os métodos de ***programação dinâmica*** e os métodos de ***programação linear***. São chamados de programação pois a quantidade de cálculos é demasiadamente grande para ser efetivada manualmente, mas facilmente programáveis, usando FORTRAN ou qualquer outro programa de simulação matemática e são de extrema importância, tanto em engenharia como em negócios. *Para as restrições não lineares e/ou do tipo igualdade*, pode-se usar o ***método direto*** se a função for simples e facilmente explicitada em termos de uma ou outra variável. Para as equações onde não é possível explicitar uma das variáveis em relação à outra, usa-se então o ***método dos multiplicadores de Lagrange***.

**Método direto:**
O método direto consiste em substituir na função F(x,y) a variável ***x*** ou a variável ***y*** pelo correspondente valor obtido da função originada pela restrição R(x,y) = 0. A função F assim modificada, fica dependente então somente de uma variável e portanto o problema fica reduzido ao estudo de máximos e mínimos de uma única variável. Percebe-se claramente que este método é de fácil aplicação se F for dependente apenas de duas variáveis. Nos demais casos, o problema assume uma complexidade de difícil solução.

Exemplo: Encontrar um ponto de máximo ou mínimo, dada pela função:

$$z = -(x-4)^2 - (y-9)^2 + 50$$

Cuja restrição é imposta pela expressão: $x + y = 20$

Da restrição se isola ***x*** ou ***y*** e substitui-se na primeira equação, ficando então:

$$z = -(x-4)^2 - (20-x-9)^2 + 50 = -(x-4)^2 - (11-x)^2 + 50$$

Como a função é facilmente derivável, obtém-se a derivada de z em relação a x, iguala-se a zero e verifica-se se a mesma é um ponto de máximo ou mínimo. Deve-se tomar cuidado porque às vezes se incorre em pontos que não são nem máximos nem mínimos, mas também conhecidos como ***pontos de sela***, e são equivalentes aos pontos de inflexão existente nas equações de uma única variável. A derivada da equação acima fica:

$$\frac{dz}{dx} = -2(x-4) - 2(11-x)(-1) = -4x + 30$$

fazendo dz/dx = 0, vem: $\frac{dz}{dx} = -4x + 30 = 0 \Rightarrow x = 7{,}5$

Encontrando o valor para y, na restrição, obtém o valor y = 12,5, substituindo estes valores na função obtém-se o valor máximo desta função. Assim:

$$z = -(7{,}5-4)^2 - (12{,}5-9)^2 + 50 = 25{,}5$$

*Se a função não fosse facilmente derivável dever-se-ia usar um dos métodos de pesquisa apresentados para funções de uma única variável (**método dicotômico** ou **método de Fibonacci**) para se determinar o valor da variável **x**. Uma vez determinado este, a outra variável, y, no caso, seria obtida diretamente através da equação da restrição.*

**Exercícios:**
Calcular seguindo o modelo acima, sobre a seguinte função: $z = 16x + 26y - x^2 - y^2$, sabendo que $3x + 4y = 26$;
Calcular o valor máximo de $z = xy$, sabendo que $x + y = 40$ com $x>0$ e $y > 0$;
Calcular o valor mínimo de $z = y^2 - x - 5$, sabendo que $x - \ln(y) = 0$;

**Método dos multiplicadores de Lagrange:**
Este método é útil quando o problema a ser pesquisado envolve variáveis independentes, dependentes entre si, e estão ligadas por uma equação da forma R(x,y) = 0, da qual não se pode explicitar x como função de y, nem y como função de x. Este método já foi aplicado nos métodos de interpolação (*método da continuidade*) e segue basicamente o mesmo procedimento.A técnica de Lagrange fica mais bem ilustrada ao analisar um problema de tornar máximo ou mínima uma função F(x,y,z) sujeita a duas condições de forma g(x,y,z) = 0 e h(x,y,z) = 0 onde g e h não são funcionalmente dependentes, de modo que os vínculos nem são equivalentes nem incompatíveis. Com a introdução dos multiplicadores de Lagrange arbitrários $\lambda_1$ e $\lambda_2$ obtém-se três equações, a saber:

$$\psi_x = \frac{\partial F}{\partial x} + \lambda_1 . g(x) + \lambda_2 h(x) = 0$$

$$\psi_x = \frac{\partial F}{\partial x} + \lambda_1 . g(x) + \lambda_2 h(x) = 0$$

$$\psi_x = \frac{\partial F}{\partial x} + \lambda_1 . g(x) + \lambda_2 h(x) = 0$$

Que juntamente com as condições g = 0 e h = 0 constituem um sistema de 5 equações , 5 incógnitas (x, y, z, $\lambda_1$ e $\lambda_2$). Generalizando vem que, se F for uma função de ***n*** variáveis e F tiver que ser otimizado, porém sujeita a m restrições independentes, tal que ***m<n***, se escreve tantas funções auxiliares $\Psi$, tal que

$$\psi_x = \frac{\partial F}{\partial x} + \sum_{i=1}^{m} \lambda_i . g_i = 0$$

De forma a ter um sistema de m+n equações, m+n incógnitas.

Exemplo 1: *(Para melhor entender o procedimento acompanhe o exemplo resolvido abaixo)*
Determinar, caso existam, os possíveis pontos de máximo ou de mínimo da função dada por z = xy, sabendo que $x^2 + y^2 = 9$

Neste caso tem-se: F(x,y) = xy e $R(x,y) = x^2 + y^2 - 9$
E, portanto,

$$\frac{\partial F}{\partial x} - \lambda . \frac{\partial R}{\partial x} = 0 \qquad\qquad y - 2\lambda x = 0$$

$$\frac{\partial F}{\partial y} - \lambda . \frac{\partial R}{\partial y} = 0 \qquad \Rightarrow \qquad x - 2\lambda y = 0$$

$$R(x, y) = 0 \qquad\qquad x^2 + y^2 - 9 = 0$$

Resolvendo o sistema de 3 equações, três incógnitas, vem:

ponto 1: $\left(\frac{3}{\sqrt{2}} \quad \frac{3}{\sqrt{2}} \quad \frac{1}{2}\right)$

ponto 2: $\left(\frac{3}{\sqrt{2}} \quad -\frac{3}{\sqrt{2}} \quad -\frac{1}{2}\right)$

ponto 3: $\left(-\frac{3}{\sqrt{2}} \quad \frac{3}{\sqrt{2}} \quad -\frac{1}{2}\right)$

ponto 4: $(-\frac{3}{\sqrt{2}} \quad -\frac{3}{\sqrt{2}} \quad \frac{1}{2})$

Logo, substituindo o valor de λ nas equações acima, os 4 possíveis pontos de máximo ou de mínimo, são:

$P_1 = (\frac{3}{\sqrt{2}} \quad \frac{3}{\sqrt{2}})$ $P_2 = (\frac{3}{\sqrt{2}} \quad -\frac{3}{\sqrt{2}})$ $P_3 = (-\frac{3}{\sqrt{2}} \quad \frac{3}{\sqrt{2}})$ $P_4 = (-\frac{3}{\sqrt{2}} \quad -\frac{3}{\sqrt{2}})$

Exemplo 2.

Seja : $F = x^2y^2z^2$ e cuja restrição é dada por $R = x^2 + y^2 + z^2 - c^2$

Fazendo $\psi = x^2y^2z^2 + \lambda(x^2 + y^2 + z^2 - c^2)$

E calculando as derivadas, vem...

$$\psi_x^{'} = 2xy^2z^2 + 2\lambda x \quad \Rightarrow \quad xy^2z^2 + \lambda x = 0$$

$$\psi_y^{'} = 2x^2yz^2 + 2\lambda y \quad \Rightarrow \quad x^2yz^2 + \lambda y = 0$$

$$\psi_z^{'} = 2x^2y^2z + 2\lambda z \quad \Rightarrow \quad x^2y^2z + \lambda z = 0$$

$$R = x^2 + y^2 + z^2 - c^2 \quad \Rightarrow \quad x^2 + y^2 + z^2 = c^2$$

onde as respostas possíveis são: $x^2 = y^2 = z^2$ e $\lambda = -x^4$

resultando em $x = \pm\frac{c}{\sqrt{3}}$ $y = \pm\frac{c}{\sqrt{3}}$ $z = \pm\frac{c}{\sqrt{3}}$ $\lambda = -\frac{c^4}{9}$

os quais substituídos na função $\psi$ acima fornece um ponto de máximo ou de mínimo em $\psi_{\max} = \frac{c^6}{27}$

**Exercícios:**
Seguindo um dos exemplos acima, para resolver os seguintes exercícios:

3. Resolver da mesma maneira a função $P = 0{,}8T^{0,2}.\ C^{0,8}$ tendo como restrição a função 2T+3C=6000.
   Este problema tem apenas um ponto possível de máximo ou de mínimo, com a restrição dada, sendo o ponto T = 2142,8 e C = 571,4.

4. $U = (x+2)^{2/3}.(y+1)^{1/3}$, com restrição igual a 2x+y = 7

5. Z = 4x + 5y, restrição imposta por $(x-2)^2+(y-4)^2=16$

6. F = 4xy, com restrição imposta por $g = x^2+y^2 =1$

7. Encontrar os possíveis pontos de máximo e mínimo da função $F = 4x^2 – 5y^2 + 2xy + 4x – 8y + 10$, sem nenhuma restrição. [solução: (–2/7, –6/7), ponto de sela]

**Otimização 2 - Programação:**

- Programação linear

A programação linear é um tipo de otimização estática. É linear porque é um caso típico da aplicação da álgebra linear e das técnicas de cálculo numérico para a resolução de problemas de engenharia. A *programação linear* visa fundamentalmente encontrar a melhor solução para problemas que tenham seus modelos representados por expressões lineares. A grande simplicidade e aplicabilidade que a caracterizam devem-se a linearidade do modelo. O objetivo da programação linear consiste na maximização ou minimização de uma função linear, chamada ***função objetivo***, respeitando um sistema linear de igualdades ou desigualdades que recebem o nome de ***condições restritivas***. Essas condições representam normalmente limitações de recursos disponíveis ou exigências técnicas que devem ser cumpridas e determinam uma região à qual se dá o nome de ***conjunto das soluções viáveis***. A melhor solução, aquela que maximiza ou minimiza a função objetivo denomina-se ***solução ótima***.
A resolução de problemas de programação linear consiste basicamente em dois passos:

- A modelagem, ou também chamado de equacionamento do problema. Não existem técnicas precisas capazes de permitir o estabelecimento de um modelo. Dependerá exclusivamente da natureza do problema e da capacidade de análise e síntese da pessoa envolvida.
- Solução do problema. A solução do problema passa pelo uso das técnicas de álgebra linear, os quais podem ser resolvidos numericamente.

Exemplos de problemas de programação linear:

1. Um nutricionista precisa estabelecer uma dieta contendo, pelo menos 10 unidades de vitamina A, 30 unidades de vitaminas B e 18 unidades de vitamina C. Essas vitaminas estão contidas em quantidades variáveis em cinco tipos de alimentos que recebem aqui o nome de $s_1$, $s_2$, $s_3$, $s_4$ e $s_5$. O quadro abaixo dá as características composicional e o custo de cada alimento.

| Alimento | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| Vit A | 0 | 1 | 5 | 4 | 3 |
| Vit B | 2 | 1 | 0 | 3 | 2 |
| Vit C | 3 | 1 | 0 | 9 | 0 |
| Custo | 4 | 2 | 1 | 10 | 5 |

*O objetivo da otimização consiste em determinar as quantidades dos cinco alimentos que devem ser incluídas na dieta diária, que supre a dieta prescrita acima e que tenha o menor custo.*

*Modelagem*. Fazendo $x_i$ as quantidades relativas de cada alimento e observando as quantidades de vit A nos cinco alimentos pode-se expressar a quantidade da mesma como sendo:

$$x_2 + 5x_3 + 4x_4 + 3x_5 \geq 10$$

Analogamente, indicamos os outros teores mínimos, respectivamente, da seguinte forma:

$$2x_1 + x_2 + 3x_4 + 2x_5 \geq 30$$

$$3x_1 + x_2 + 9x_4 \geq 18$$

Como não podemos consumir uma quantidade negativa de unidades dos alimentos, tem-se também:

$$x_1 \geq 0 \qquad x_2 \geq 0 \qquad x_3 \geq 0 \qquad x_4 \geq 0 \qquad x_5 \geq 0$$

O custo da dieta, em unidades monetárias, fica expresso pela função linear:

$$Q(x) = 4x_1 + 2x_2 + x_3 + 10x_4 + 5x_5$$

O problema agora é determinar o ponto x = (x1, x2, x3, x4, x5) considerado ponto ótimo tal que satisfaça a todas as ***restrições*** *(inequações)* e minimize ao mesmo tempo, o valor da função Q(x)

2. Uma empresa para fabricar ***n*** produtos $P_j$ necessita de ***m*** recursos $F_i$. Para cada unidade de produto $P_j$ são necessários $a_{ij}$ unidades do recurso $F_i$. De cada recurso $F_i$ só existe a quantidade $b_i$ ($b_i \geq 0$)
Sabendo que cada unidade do produto $P_j$ dá um lucro $c_j$, qual a quantidade $x_j$ que deve ser produzida de cada produto $P_j$ para que o lucro seja o maior possível?

*Modelagem*. Para cada recurso $F_i$, existe uma restrição:

$$\begin{array}{rcrcrcrcl} a_{11}x_1 & + & a_{12}x_2 & + & ..... & + & a_{1n}x_n & \leq & b_1 \\ a_{12}x_1 & + & a_{22}x_2 & + & ..... & + & a_{2n}x_n & \leq & b_2 \\ ..... & + & ..... & + & ..... & + & ..... & \leq & .... \\ a_{m1}x_1 & + & a_{m2}x_2 & + & ..... & + & a_{mn}x_n & \leq & b_m \end{array}$$

Como não podemos fabricar uma quantidade negativa, temos também...

$$x_1 \geq 0 \qquad x_2 \geq 0 \qquad ..... \qquad x_n \geq 0$$

A função lucro Q(x) que é a função objetivo pode ser representada por:

$$Q(x) = c_1x_1 + c_2x_2 + ..... + c_nx_n$$

Portanto, esse problema consiste em determinar o ponto x = ($x_1$, $x_2$, ....., $x_n$) que satisfaça as ***m*** restrições dos recursos $F_i$ e as ***n*** restrições de não negatividade das quantidades produzidas, ao mesmo tempo que maximiza a função objetivo Q(x).

O desenvolvimento de um método que determine a solução de problemas deste tipo torna necessária a redução do problema a uma forma tal que permita a aplicação direta do método. Nos problemas de programação linear, o método mais usado é o método conhecido como ***SIMPLEX***. Para que este método seja aplicado a forma padrão deve ser como mostrada abaixo:

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \quad \text{Onde } b_i \geq 0 \qquad (i = 1, 2, .....m)$$

$$x_j \geq 0 \qquad (j = 1, 2,....., n)$$

$$\sum_{j=1}^{n} c_jx_j = Q(x) \rightarrow \min$$

*Os dois primeiros conjuntos de equações normalmente são denominados restrições do problema. O segundo denomina-se condição de não negatividade enquanto a última equação representa a função objetivo.*

Este mesmo modelo pode ser apresentado na forma matricial, como mostrado abaixo:

$$Ax = 0 \qquad \text{Onde } b \geq 0$$

$$x \geq 0$$

$$c^T x = Q(x) \rightarrow \min$$

Assim, todo e qualquer problema que não esteja na forma padrão deve primeiramente ser convertido para esta. Cabe lembrar que a forma padrão implica que o primeiro grupo de restrições envolva somente igualdades e que todas as variáveis do modelo sejam não negativas. Além disso, sempre deve ***minimizar*** a função objetivo.

Para efetivar essas mudanças, observar abaixo:

*Ocorrência de desigualdades.* Qualquer desigualdade ou inequação linear pode ser transformada em uma equação se subtrairmos ou adicionarmos variáveis positivas denominadas ***variáveis de folga***;

$$a_{k1}x_2 + a_{k2}x_3 + .... + a_{kn}x_n \leq b_k \Rightarrow a_{k1}x_2 + a_{k2}x_3 + .... + a_{kn}x_n + x_{n+1} + b_k \qquad x_{n+1} \geq 0$$

Observe que a variável $x_{n+1}$ foi introduzida convertendo a desigualdade em igualdade. ***Para inequações maiores que zero a variável de folga deve ser subtraída***.

*Ocorrência de $b_i < 0$*. Neste caso bastará multiplicar a restrição ***i*** por (-1), pois todos os coeficientes $a_{ij}$ podem ter qualquer sinal;

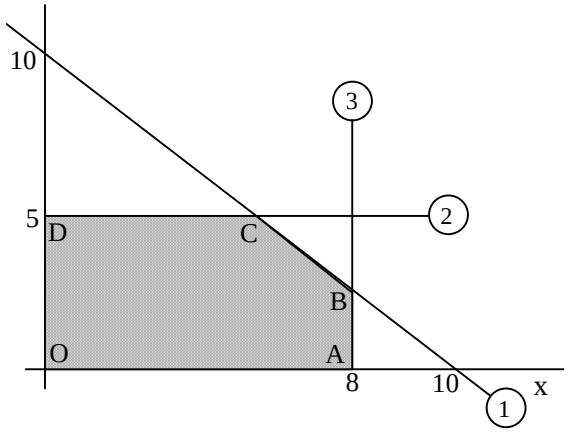
*Variáveis livres*. As variáveis livres que não tem qualquer restrição de sinal. Pode-se sempre substituir uma variável deste tipo ($x_k$) por outras duas ($x'_k$ e $x''_k$) de maneira que $x_k = x'_k - x''_k$ , sendo que $x'_k \geq 0$ e $x''_k \geq 0$, e assim, substitui-se uma variável livre por outras duas variáveis positivas;

*Variável não positiva*. Se o modelo for formulado com uma variável $x_k \leq 0$, basta substituí-la pela sua simétrica, isto é, basta fazer, $x'_k = -x_k$ e substituir $x_k$ por $x'_k$ nas equações do problema;

*A função objetivo é de maximização*. Substituir a função objetivo dada pela sua simétrica, passando a minimizar esta última, ou seja, MAX {Q(x)} = – MIN {– Q(x)}.

Solução gráfica:
Para compreender melhor o processo de solução de um problema de programação linear, pode-se usar o método gráfico. Embora facilmente visualizável este método só pode ser empregado quando o problema envolver apenas duas variáveis. Como exemplo numérico, deseja-se determinar os pontos de máximo e de mínimo da função z = 10x + 20 y.
As restrições impostas aos problemas são: $x + y \leq 10$, $x \leq 8$, $y \leq 5$, $x \geq 0$, $y \geq 0$.

- Passo 1. Fazer a representação gráfica , convertendo as inequações em igualdades.




1) $x + y = 10$

2 ) $y = 5$

3) $x = 8$

- Passo 2. Determinar os vértices do polígono OABCD
  Existe um teorema que demonstra que se existe um ponto de máximo ou mínimo, estes se encontrarão nos vértices do polígono, portanto busca-se a solução em um dos vértices possível.

  Vértice O – (0,0)
  Vértice A – (8,0)
  Vértice B – (intersecção entre x = 8 e x + y = 10 resultando em (8,2)).
  Vértice C – (intersecção entre y = 5 e x + y = 10 resultando em (5,5)).
  Vértice D – (0,5)

- Passo 3. Determinar e comparar os valores da função nos vértices

  No ponto O – z = 10 x 0 +20 x 0 = 0
  No ponto A – z = 10 x 8 + 20 x 0 = 80
  No ponto B – z = 10 x 8 + 20 x 2 = 120
  No ponto C – z = 10 x 5 + 20 x 5 = 150
  No ponto D – z = 10 x 0 + 20 x 5 = 100

  Portanto, o ponto de mínimo é o ***ponto O*** (0) e ponto de máximo é o ***ponto C*** (150)

## Otimização 3 - Programação Linear

- Método SIMPLEX para maximização

O método gráfico para a resolução de problemas de programação linear tem vários pontos negativos. Em primeiro lugar, o mesmo é tedioso de executar. Requer muitos cuidados e não é facilmente programável para execução em computadores e, além disso, tem aplicação restrita a problemas contendo no máximo três incógnitas independentes, sendo normalmente aplicado somente àqueles problemas de duas variáveis. Entretanto, como a solução, se existir, ocorre sempre nos vértices da figura, pode-se então se resolver esse tipo de problemas através de uma matriz e seus métodos de solução descritos pela álgebra linear. Este método é simplesmente chamado de ***método simplex***. O método em si, a priori, permite a solução de problemas de ***maximização***, entretanto, o método pode ser modificado para resolver problemas de minimização também.

Considere o problema abaixo, onde se deseja maximizar a expressão Z em função dos valores de x e y, e sujeita às restrições descritas logo abaixo:
Expressão:

$$Z = 5x + 4y$$

Restrições:

$$x + 2y \leq 12$$
$$6x + 4y \leq 48$$
$$-x + 2y \leq 8$$
$$x, y \geq 0$$

Este problema tem infinitas soluções possíveis, e entre elas, se deseja então, se determinar àquela que fornece o valor máximo a Z. Porque existe uma extensa teoria sobre sistemas de equações lineares, é vantajoso se converter este problema para um sistema de equações contendo o número de equações igual ao número de incógnitas. Assim, em primeiro lugar, se convertem as inequações em equações pela inclusão de uma ***variável de folga*** em cada restrição. A equação $\boldsymbol{ax + by \leq c}$ fica equivalente a $\boldsymbol{ax + by + s = c}$, com $\boldsymbol{s \geq 0}$. O mesmo se faz com as demais restrições, obtendo-se assim um sistema com 4 equações e 6 incógnitas (z, x, y, $s_1$, $s_2$ e $s_3$). Ao transformar uma inequação para uma equação pode-se depois escolher quaisquer valores de x e y de maneira que o resultado fique abaixo ou no máximo, sobre a reta imposta pela equação. Ao escolherem-se valores menores, a variável de folga assume valores diferentes de zero e quando os valores de x e y aplicados implicarem num valor sobre a reta, o valor da variável de folga torna-se zero. Devido ao fato de se encontrar um ponto máximo ou mínimo sempre nos vértices da figura geométrica, isto implica dizer que pelo menos duas variáveis de folga serão sempre zero e, portanto, das 6 incógnitas, 2 desaparecem, resultando num sistema de 4 equações a 4 incógnitas. O conjunto dos valores de ***x***, ***y*** e ***Z*** referentes aos vértices forma o que se chama de ***soluções básicas***. O conjunto de variáveis que *não são escolhidas* para serem zeros e que são determinadas pela resolução do problema é chamado de ***base*** e cada variável é chamada de ***variável básica***. As variáveis igualadas a zero são chamadas de não-básicas. Como exemplo, observar o exemplo dado acima. Se fizermos x = 0, y = 0 e Z = 0, $s_1$ assume o valor 12, $s_2$ assume 48 e $s_3$ assume o valor 8.

Resumindo**:**

*Soluções básicas*. Possíveis valores das variáveis da função objetivo. Uma das soluções terá o valor máximo que é o resultado desejado, a ser obtido no final do processo da resolução.

*Base*. Lugar onde entram as variáveis básicas, ou seja, aquelas diferentes de zero.

*Variáveis básicas*. São as variáveis diferentes de zero e ocupam sempre uma posição na base.

O argumento da matriz formada pelas equações é sempre escrito na forma de uma tabela que é conhecida como tableau (tábua). As variáveis são listadas no topo desta tábua e servem como rótulo para a coluna correspondente da tábua. Ao lado esquerdo da tábua, no topo entra Z ( ou sua equivalente) e as demais variáveis básicas são listadas na mesma coluna. Esta primeira coluna recebe o nome de *base*. À direita da tábua está a coluna do vetor resultado das equações.

Colocando-se as expressões do problema exemplo fica:

| | | | | | | | Vetor b¬ |
|---|---|---|---|---|---|---|---|
| | 1ª coluna ¬ | | | | | | |
| Coluna objetivo¬ | | | | | | | |
| Rótulo→ | *Z* | *x* | *y* | $s_1$ | $s_2$ | $s_3$ | *b* |
| Base↓ | | | | | | | |
| *Z* | 1 | −5 | −4 | 0 | 0 | 0 | **0** |
| $s_1$ | 0 | 1 | 2 | 1 | 0 | 0 | 12 |
| $s_2$ | 0 | 6 | 4 | 0 | 1 | 0 | 48 |
| $s_3$ | 0 | −1 | 2 | 0 | 0 | 1 | 8 |

Observar que desta matriz, os valores de todas as variáveis podem ser determinados imediatamente pois se referem à solução básica trivial que coincide com o ponto O da solução gráfica ( x = 0, y = 0, e Z = 0). Este será sempre o ponto de partida. Este método oferece uma sistemática para mover de um vértice a outro, sempre incrementando o valor de Z. Observar também que todos os valores que não estão na base são iguais a zero. Para mover de um vértice a outro, uma variável básica é permutada por outra não básica. A variável não básica entra na base e a variável básica do seu lugar muda para ***zero***.

Para fazer esta alteração são necessárias três decisões, a saber:

- Qual variável poderá entrar na base;
- Qual variável sairá da base;
- Qual valor deverá ser atribuído à variável que entra.

Examinando a primeira equação, pode-se perceber que Z aumentará se o conteúdo das colunas *x* ou da coluna *y* tornarem-se positivos. Das duas colunas, aumenta mais o valor da coluna x, pois a taxa de incremento da mesma é maior (a derivada parcial de Z em relação a x é maior que a derivada parcial de Z em relação a y). Assim, sempre se escolherá a variável que tiver o maior coeficiente negativo para entrar na primeira coluna da tábua. A variável que sai da base e o valor que entra são ambos determinados examinando as restrições e marcando a possível variável como a maior possível sem violar todas as restrições. Lembrando que neste caso, foi escolhido o x para entrar na base, e portanto ***y*** continuará sendo igual a zero, assim analisa-se a expressão resultante:

$$\begin{aligned} x + s_1 &= 12 \quad \Rightarrow x = 12 \\ 6x + s_2 &= 48 \quad \Rightarrow x = 8 \\ -x + s_3 &= 8 \quad \Rightarrow x = - \end{aligned}$$

porque $s_1$ deve ser sempre positivo e o menor valor positivo para s é zero, a primeira restrição requer que x não exceda o valor 12. Na segunda restrição este valor não deve exceder o valor 8. Na terceira restrição, uma vez que o sinal do coeficiente x é negativo, a variável de folga será sempre positiva independente da escolha de do valor para x. De posse destes dados, verifica-se que a condição mais restritiva é aquela que tem o menor valor para a variável x e neste caso escolhe-se então a 2ª restrição. ***Assim x entra no lugar de $s_2$***, e a linha escolhida é agora tratado como linha pivô. Para proceder esta permuta se completa as seguintes etapas:

- Divide-se toda a linha pivô pelo coeficiente de x de maneira a torná-lo unitário;

- Toma-se o coeficiente de x na primeira linha da tábua, troca-se o sinal, multiplica-se termo a termo, pela linha pivô e soma-se com os coeficientes da primeira linha;

- Toma-se o valor do coeficiente da 3ª linha, troca-se o sinal, multiplica-se termo a termo, com os coeficientes da linha pivô e soma-se com os coeficientes da terceira linha.

É praticamente o mesmo processo usado no processo de eliminação de Gauss. Procedendo assim todos os valores da coluna x, exceto da linha pivô, são convertidos para zero. No lugar de $s_2$, da coluna base, se escreve então o x.

Rescrevendo a tábua, vem:

| Rótulo→ / Base↓ | *Z* | *x* | *y* | $s_1$ | $s_2$ | $s_3$ | *b* |
|---|---|---|---|---|---|---|---|
| *Z* | 1 | 0 | –2/3 | 0 | 5/6 | 0 | **40** |
| $s_1$ | 0 | 0 | 4/3 | 1 | –1/6 | 0 | 4 |
| *x* | 0 | 1 | 2/3 | 0 | 1/6 | 0 | 8 |
| $s_3$ | 0 | 0 | 8/3 | 0 | 1/6 | 1 | 16 |

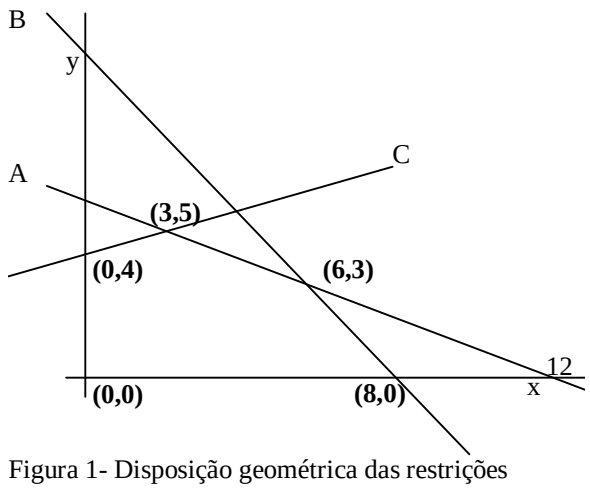
A tábua indica agora que as variáveis básicas são $s_1 = 4$, $x = 8$ e $s_3 = 16$ e as variáveis não básicas são $y = 0$ e $s_2 = 0$ e o valor de Z agora é 40. Geometricamente se passou do ponto (0,0) para o ponto (8,0).
Para decidir se existe a possibilidade de nova mudança de vértice, a qual poderá incrementar o valor de Z, deve-se pesquisar na primeira linha se existem ainda valores negativos. Deve-se sempre escolher o mais negativo entre eles, mas no caso, só existe mais um, o valor de ***y***. Assim, ainda é possível aumentar o valor de Z, fazendo mais uma permutação. Para determinar qual das variáveis de base que irá sair, novamente se toma a mais restritiva entre todas, ou seja, aquela restrição que tiver o menor valor possível para y. Fazendo a verificação vem:

$$\frac{4}{3}y + s_1 = 4 \quad \Rightarrow \quad y = 3$$
$$\frac{2}{3}y + s_2 = 8 \quad \Rightarrow \quad y = 12$$
$$\frac{2}{3}y + s_3 = 16 \quad \Rightarrow \quad y = 6$$

a linha pivô agora será a primeira das restrições, e a variável permutada será $s_1$. Procedendo a eliminação vem:

| Rótulo→ / Base↓ | *Z* | *x* | *y* | $s_1$ | $s_2$ | $s_3$ | *b* |
|---|---|---|---|---|---|---|---|
| *Z* | 1 | 0 | 0 | 1/2 | 3/4 | 0 | **42** |
| $s_1$ | 0 | 0 | 1 | 3/4 | –1/8 | 0 | 3 |
| *x* | 0 | 1 | 0 | –1/2 | 1/4 | 0 | 6 |
| $s_3$ | 0 | 0 | 0 | –2 | 1/2 | 1 | 12 |

Concluindo, percebe-se que agora x = 6, y = 3 e Z = 42. Geometricamente se passou para o vértice (6,3). Percebe-se também que não é mais possível se obter valor maio para Z pois não existe mais nenhum valor negativo na linha de Z.




A: x + 2y = 12

B: 6x + 4y = 48

C: -x + 2y = 8

Figura 1- Disposição geométrica das restrições

Resumo dos passos a serem seguidos no método SIMPLEX para maximização:

- Rescrever o problema de programação linear como um sistema de equações, inserindo em cada restrição uma variável de folga. Colocar este sistema de equações numa disposição de tábua como mostrada acima. Coloque todas as variáveis de folga na base juntamente com a variável a ser maximizada. (Existem outros modelos);

- Selecione como variável de entrada, a mais negativa entre todas existentes na linha da variável a ser maximizada (Z ou equivalente). A coluna onde se encontra esta variável é denominada então de coluna pivô;

- Para cada valor positivo da coluna pivô, divida o valor correspondente da coluna b por este valor. Entre todos os valores marque aquele que fornecer o menor quociente. Este valor definirá a linha pivô. Faça um circulo no valor que pertença simultaneamente à coluna e à linha pivôs. Este valor será denominado simplesmente como ***pivô***. A variável de base correspondente deixará a base e em seu lugar entrará a variável da coluna pivô;

- Divida todos os elementos da linha pivô pelo pivô de maneira a torná-lo unitário;

- Faça combinação linear de todas as demais linhas com a linha pivô de maneira a eliminar todos os valores diferentes de zero na coluna pivô com exceção do pivô. Para ter sucesso nesta operação basta tomar para cada linha , o valor existente na coluna pivô, inverter seu sinal e usar este valor para multiplicar termo a termo, todos os valores da linha pivô e somá-los aos termos da linha em questão;

- Troque a variável da base, na posição da linha pivô, pela variável da coluna pivô;

- Repita o processo desde o passo 2 até não haver nenhum outro valor negativo na linha da variável a ser maximizada.

Exercícios:

Maximize a expressão: $Q = 20x + 10y + 30z$

Sujeita as restrições:
$x + y + 4z \leq 60$
$3x + y + 2z \leq 120$
$x, y, z \geq 0$

Maximize a expressão: $Z = x_1 + 6x_2 + 3x_3 + 2x_4$

Sujeita as restrições:
$3x_1 - 5x_2 + 2x_3 + 3x_4 \leq 12$
$x_1 - 2x_2 + 3x_3 + x_4 \leq 2$
$5x_1 + 3x_2 + 2x_3 - x_4 \leq 3$
$x_1, x_2, x_3, x_4 \geq 0$