



MATEMÁTICA COMPUTACIONAL

Prof. Mauro Larrat
maurolarrat@ufpa.br
Primeiro semestre de 2016

Sumário

Introdução

Erros

Resolução de Sistemas Lineares

Sistemas Diretos (Gauss, Gauss-Jordan, Pivoteamento parcial e completo)

Para que serve a matemática computacional?

Modelar matematicamente problemas complexos e obter soluções para estes modelos através da aplicação de ferramentas computacionais que implementam métodos numéricos.

Para que serve a matemática computacional?

Modelar matematicamente problemas complexos e obter soluções para estes modelos através da aplicação de ferramentas computacionais que implementam métodos numéricos.

Problemas complexos: de difícil solução ou de solução aproximada.

Para que serve a matemática computacional?

Modelar matematicamente problemas complexos e obter soluções para estes modelos através da aplicação de ferramentas computacionais que implementam métodos numéricos.

Problemas complexos: de difícil solução ou de solução aproximada.

Métodos Numéricos: operações matemáticas elementares (lógica e aritmética).

Para que serve a matemática computacional?

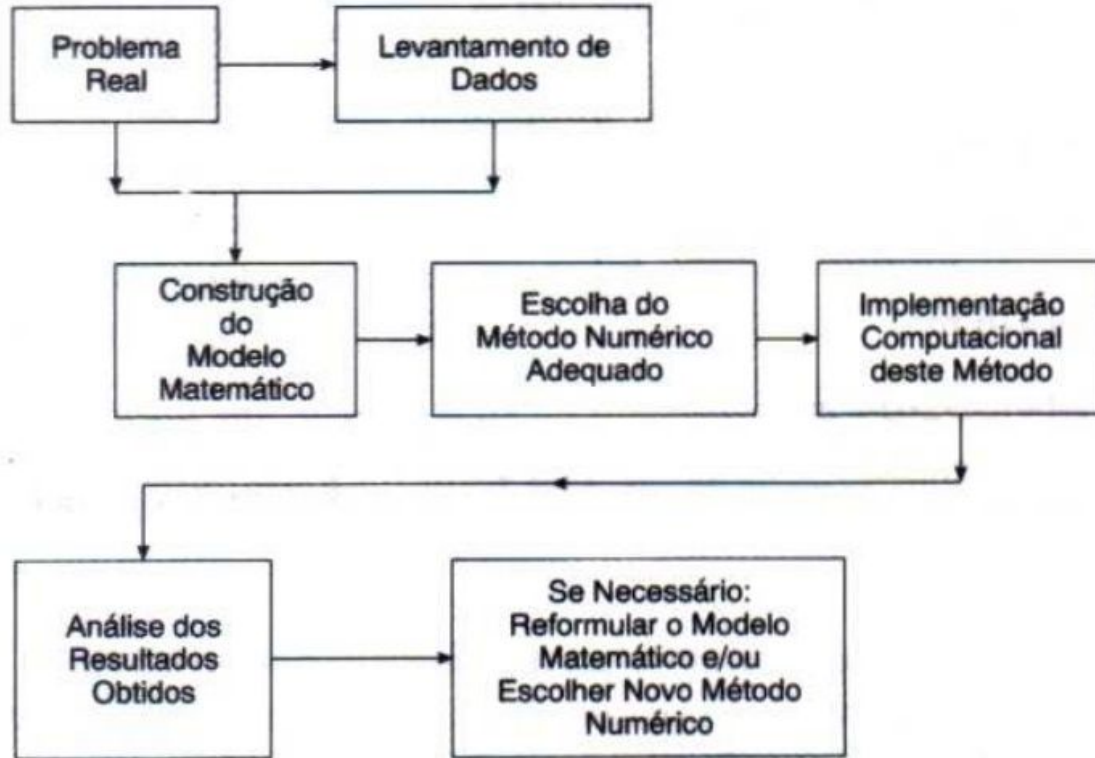
Modelar matematicamente problemas complexos e obter soluções para estes modelos através da aplicação de ferramentas computacionais que implementam métodos numéricos.

Problemas complexos: de difícil solução ou de solução aproximada.

Métodos Numéricos: operações matemáticas elementares (lógica e aritmética).

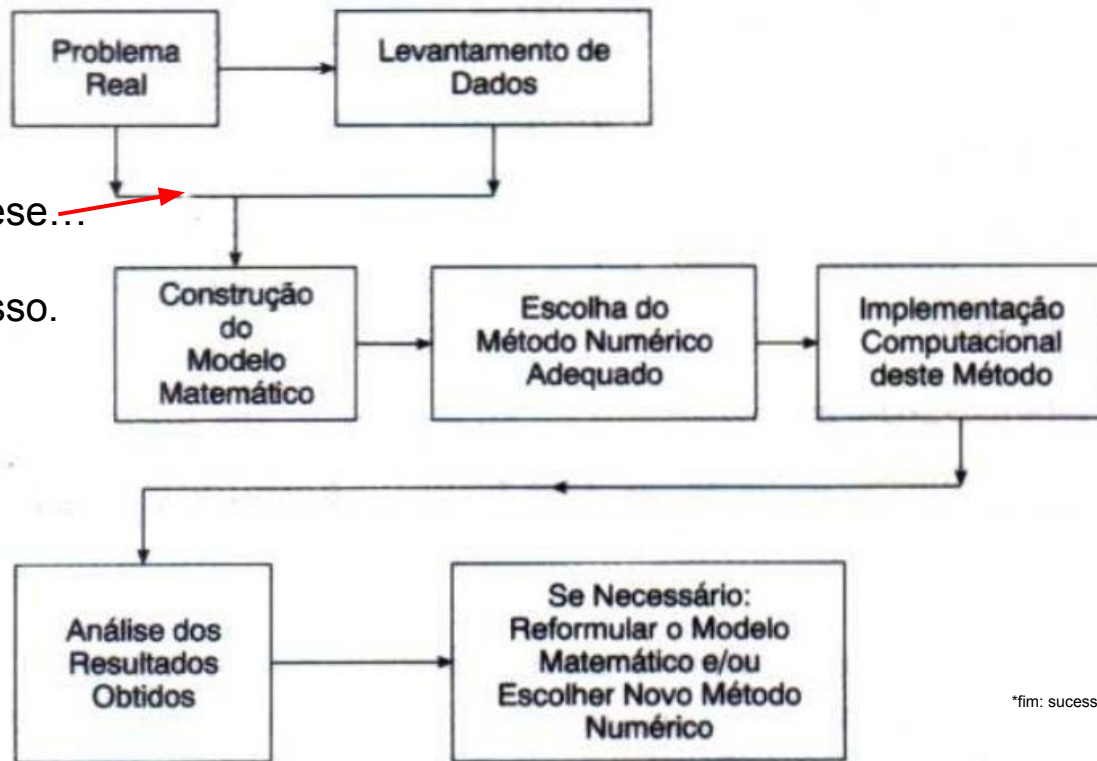
Ferramentas computacionais: **algoritmos!**

Onde o Cientista da Computação entra nisso tudo?



Onde o Cientista da Computação entra nisso tudo?

Ele formula a hipótese...
e segue adiante
até o fim* do processo.



*fim: sucesso ou fracasso.

Baseado na figura do livro Márcia Ruggiero, Vera Lopes: Cálculo Numérico: Aspectos Teóricos e Computacionais, página 1.

De onde surgem esses problemas complexos?

Desde calcular a velocidade de um saltador de bungee-jumping até todos os problemas que permeiem a criatividade de alguém que queira contratar você para resolvê-los!

Duvida!?

Um estudo de caso

Vamos supor que uma empresa de bungee-jumping contrate você para prever a velocidade de um saltador no tempo em que este está em queda livre (antes de a corda começar a esticar).

Problema real: Calcular o comprimento e a resistência da corda considerando a massa de cada pulador.

Um estudo de caso

Vamos supor que uma empresa de bungee-jumping contrate você para prever a velocidade de um saltador no tempo em que este está em queda livre (antes de a corda começar a esticar).

Problema real: Calcular o comprimento e a resistência da corda considerando a massa de cada pulador.

Levantamento de dados: Queda livre? Física? Química? Engenharia?

Pesquisar...

Uma saída através da Modelagem Matemática

Variáveis Dependentes = $f(\text{Variáveis Independentes, Parâmetros, Função})$

Variáveis Dependentes: características que refletem o comportamento ou estado do sistema (problema).

Variáveis Independentes: dimensões (espaço, tempo, etc.) onde as características do problema estão definidas.

Parâmetros: propriedades ou composições do problema.

Função Força: função (influência externa sobre o problema) dependente da dimensão do tempo.

Modelagem Matemática

Variáveis Dependentes = f(Variáveis Independentes, Parâmetros, Função)

Modelagem criada por Newton: $F = ma$ ou $a = F/m$

$a = ?$

$F = ?$

$m = ?$

Modelagem Matemática

Variáveis Dependentes = f(Variáveis Independentes, Parâmetros, Função)

Modelagem criada por Newton: $F = ma$ ou $a = F/m$

a = variável dependente (a aceleração do pulador de bungee-jumping).

F = função força (força resultante que atua no pulador de bungee-jumping).

m = parâmetro (massa do pulador de bungee-jumping).

Cadê as variáveis independentes?

Modelagem Matemática

Variáveis Dependentes = f(Variáveis Independentes, Parâmetros, Função)

Modelagem criada por Newton: $F = ma$ ou $a = F/m$

a = variável dependente (a aceleração do pulador de bungee-jumping).

F = função força (força resultante que atua no pulador de bungee-jumping).

m = parâmetro (massa do pulador de bungee-jumping).

Cadê as variáveis independentes? Ainda não estamos prevendo o comportamento/variação da aceleração conforme o tempo/espço...

Modelagem Matemática

Variáveis Dependentes = f (Variáveis Independentes, Parâmetros, Função)

Modelagem criada por Newton: $F = ma$ ou $a = F/m$

Observem que esta modelagem descreve um processo natural de forma simplificada e em termos matemáticos;

Contudo, exclui uma série de propriedades necessárias para descrever fenômenos mais complexos - no caso do nosso exemplo - a relatividade da queda livre quanto à superfície Terra e a resistência do ar.

Modelagem Matemática

Variáveis Dependentes = f(Variáveis Independentes, Parâmetros, Função)

Modelagem criada por Newton: $F = ma$ ou $a = F/m$

Observem que esta modelagem descreve um processo natural de forma simplificada e em termos matemáticos;

Contudo, exclui uma série de propriedades necessárias para descrever fenômenos mais complexos - no caso do nosso exemplo - a relatividade da queda livre quanto à superfície Terra e a resistência do ar.

Modelagem Matemática

Variáveis Dependentes = f (Variáveis Independentes, Parâmetros, Função)

Modelagem criada por Newton: $F = ma$ ou $a = F/m$

...descreve um processo natural de forma simplificada e em termos matemáticos;

Contudo, exclui uma série de propriedades necessárias para descrever fenômenos mais complexos - no caso do nosso exemplo - a relatividade da queda livre quanto à superfície Terra e a resistência do ar.

Vamos melhorar essa modelagem matemática para o nosso problema...

Modelagem Matemática

Variáveis Dependentes = f(Variáveis Independentes, Parâmetros, Função)

Modelagem criada por Newton: $a = F/m \Rightarrow dv/dt = F/m$

Agora temos a velocidade da queda livre **proporcional** à força resultante distribuída pela massa do pulador;

Ou seja, se a força resultante for positiva, o pulador vai aumentar a velocidade em queda livre; caso contrário, a velocidade de queda livre diminui;

E se a força resultante for igual a zero, o que acontece com a velocidade?

Modelagem Matemática

Variáveis Dependentes = f(Variáveis Independentes, Parâmetros, Função)

Modelagem criada por Newton: $dv/dt = F/m$

Vamos detalhar mais a **Força resultante** $F = F_d + F_u$ em termos de **variáveis independentes** e **parâmetros** mensuráveis:

$F_d = mg$ (aceleração da gravidade da Terra)

$F_u = -C_d v^2$ (resistência do ar).

Equação diferencial
resultante:

$$\frac{dv}{dt} = g - \frac{c_d}{m} v^2$$

Algebricamente não manipulável.

Soluções para o modelo matemático

Modelo $\rightarrow \frac{dv}{dt} = g - \frac{c_d}{m} v^2$

Solução analítica (integral) ... :

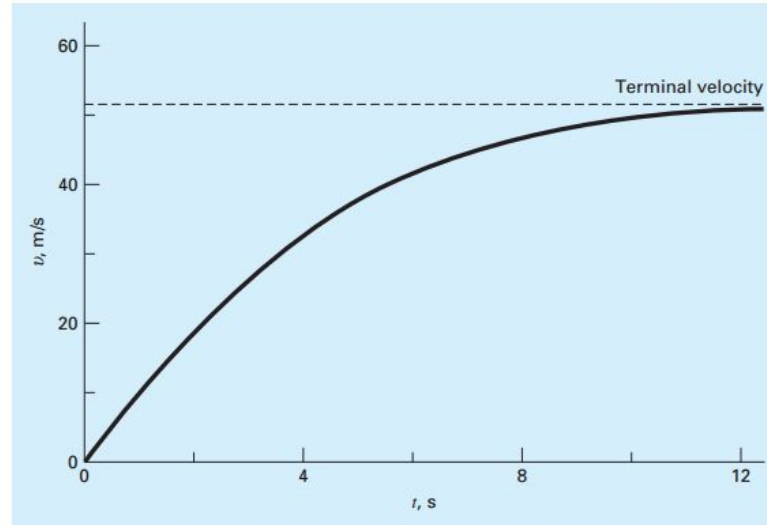
$$v(t) = \sqrt{\frac{gm}{c_d}} \tanh\left(\sqrt{\frac{gc_d}{m}} t\right)$$

...ou escolha de um método numérico adequado.

Solução Analítica: satisfaz exatamente a EDO original.

$$v(t) = \sqrt{\frac{9.81(68.1)}{0.25}} \tanh\left(\sqrt{\frac{9.81(0.25)}{68.1}}t\right) = 51.6938 \tanh(0.18977t)$$

$$v(t) = \sqrt{\frac{gm}{c_d}} \tanh\left(\sqrt{\frac{gc_d}{m}}t\right)$$



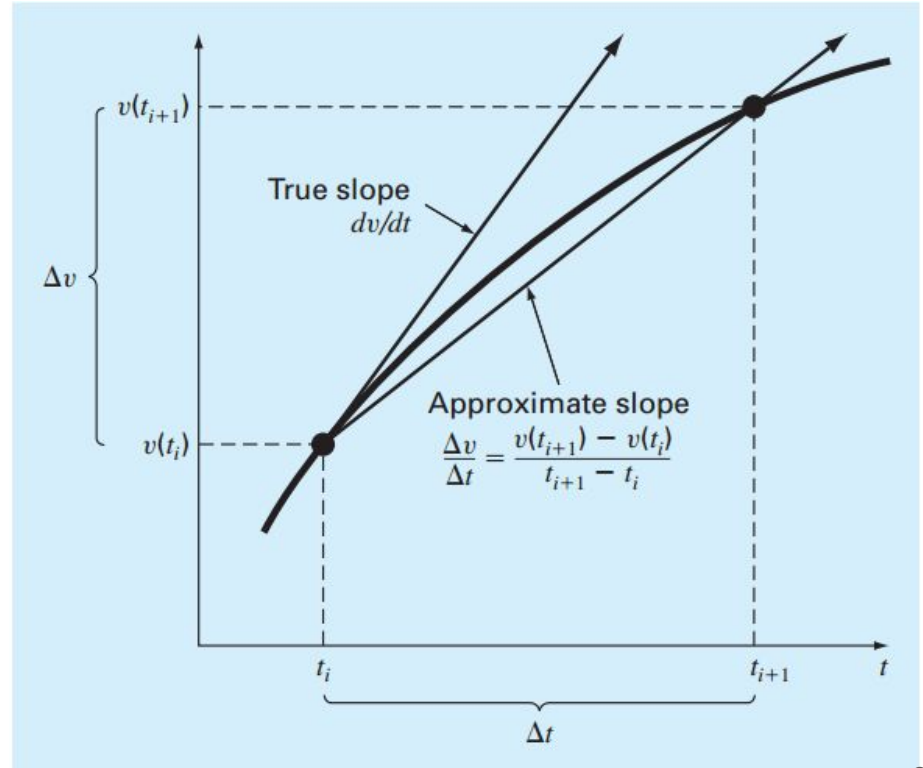
t, s	$v, m/s$
0	0
2	18.7292
4	33.1118
6	42.0762
8	46.9575
10	49.4214
12	50.6175
∞	51.6938

Solução Aproximada: método numérico

$$\frac{dv}{dt} \cong \frac{\Delta v}{\Delta t} = \frac{v(t_{i+1}) - v(t_i)}{t_{i+1} - t_i}$$

Aproximada, pois Δt é finito.

No gráfico ao lado, Δt é considerado “um instante” finito.



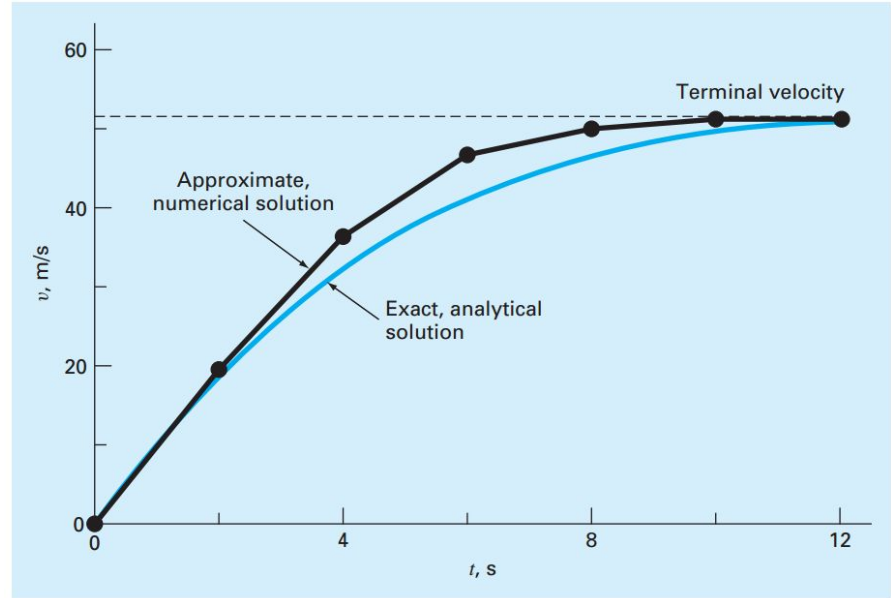
Solução Aproximada: método numérico

$$\frac{dv}{dt} \approx \frac{\Delta v}{\Delta t} = \frac{v(t_{i+1}) - v(t_i)}{t_{i+1} - t_i}$$

$$\frac{v(t_{i+1}) - v(t_i)}{t_{i+1} - t_i} = g - \frac{c_d}{m} v(t_i)^2$$

Método de Euler:

$$v_{i+1} = v_i + \frac{dv_i}{dt} \Delta t \quad \rightarrow \quad v(t_{i+1}) = v(t_i) + \left[g - \frac{c_d}{m} v(t_i)^2 \right] (t_{i+1} - t_i)$$



Questão 1.

A aplicação do método numérico capturou as principais características da solução analítica? Resolva e compare as duas formas para $T = \{1, 2, 3, 4, 5, 6\}$:

Usando a solução analítica:

$$v(t) = \sqrt{\frac{gm}{c_d}} \tanh\left(\sqrt{\frac{gc_d}{m}} t\right)$$

Usando o método numérico:

$$v(t_{i+1}) = v(t_i) + \left[g - \frac{c_d}{m} v(t_i)^2 \right] (t_{i+1} - t_i)$$

Questão 1.

A aplicação do método numérico capturou as principais características da solução analítica? Resolva e compare as duas formas para $T = \{1, 2, 3, 4, 5, 6\}$:

Usando a solução analítica:

$$v(t) = \sqrt{\frac{gm}{c_d}} \tanh\left(\sqrt{\frac{gc_d}{m}} t\right)$$

Usando o método numérico:

$$v(t_{i+1}) = v(t_i) + \left[g - \frac{c_d}{m} v(t_i)^2 \right] (t_{i+1} - t_i)$$

Encontraram alguma discrepância entre os resultados?

Discrepâncias entre a solução analítica e numérica

Pode-se considerar instantes cada vez menores para aproximar a curva.

Discrepâncias entre a solução analítica e numérica

Pode-se considerar instantes cada vez menores para aproximar a curva.

Utilizar manualmente instantes cada vez menores torna a solução numérica manualmente impraticável.

Discrepâncias entre a solução analítica e numérica

Pode-se considerar instantes cada vez menores para aproximar a curva.

Utilizar manualmente instantes cada vez menores torna a solução numérica manualmente impraticável.

Computadores podem realizar um grande número de cálculos facilmente.

Discrepâncias entre a solução analítica e numérica

Pode-se considerar instantes cada vez menores para aproximar a curva.

Utilizar manualmente instantes cada vez menores torna a solução numérica manualmente impraticável.

Computadores podem realizar um grande número de cálculos facilmente.

Contudo, reduzir à metade do instante equivale à dobrar a quantidade de operações computacionais. ***Trade-off* (custo-benefício) entre precisão e esforço computacional.**

Origem das discrepâncias nas soluções numéricas

Imprecisão inerente dos dados de entrada, devido às medidas utilizadas por equipamentos específicos;

Representação de dados numéricos no computador;

Operações numéricas efetuadas;

Representação de dados numéricos no computador

Problema: Calcular a área A de uma circunferência de raio 100 metros.

Alguns resultados obtidos em diferentes computadores:

- a) Se $\text{PI} = 3.14$, $A = 31400 \text{ m}^2$.
- b) Se $\text{PI} = 3.1416$, $A = 31416 \text{ m}^2$.
- c) Se $\text{PI} = 3.141592654$, $A = 31415.92654 \text{ m}^2$

Qual a relação das discrepâncias dos resultados com a representação dos dados numéricos?

Representação de dados numéricos no computador

Problema: Calcular a área A de uma circunferência de raio 100 metros.

Alguns resultados obtidos em diferentes computadores:

- a) Se $\pi = 3.14$, $A = 31400 \text{ m}^2$.
- b) Se $\pi = 3.1416$, $A = 31416 \text{ m}^2$.
- c) Se $\pi = 3.141592654$, $A = 31415.92654 \text{ m}^2$

A representação numérica depende da **base** escolhida ou disponível na máquina e do **número máximo de dígitos** usados na sua representação.

Representação de dados numéricos no computador

No Problema anterior ,o número irracional π não pode ser representado por um número finito de dígitos.

Na resposta em c) ($31415,92654 \text{ m}^2$) foi utilizada uma quantidade maior de dígitos na representação numérica e por isso a precisão foi maior.

Contudo, o resultado do cálculo da circunferência nunca será exato devido ao número π não poder ser representado por um número finito de dígitos.

Representação de dados numéricos no computador

Alem disso, alguns números possuem representação finita em determinada base e infinita em outra base.

Exemplos:

Base 10: representação finita.

Base 2: representação infinita.

$$\frac{1}{3}_{10}$$

$$0,1_{10}$$

Conversão binária para decimal

Considere os números $(347)_{10}$ e $(10111)_2$, os quais podem ser representados da seguinte forma:

$$(347)_{10} = 3 \times 10^2 + 4 \times 10^1 + 7 \times 10^0$$

$$(10111)_2 = 1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0$$

Genericamente, qualquer número na base β , $(a_j, a_{j-1}, \dots, a_1, a_0)_\beta$ pode ser representado pelo polinômio a seguir:

$$a_j \beta^j + a_{j-1} \beta^{j-1} + \dots + a_1 \beta^1 + a_0 \beta^0$$

Conversão binária para decimal

Análise: Quantos produtos existem no cálculo da conversão a seguir?

$$(10111)_2 = 1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = (23)_{10}$$

Conversão binária para decimal

Análise: Quantos produtos existem no cálculo da conversão a seguir?

$$(10111)_2 = 1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = (23)_{10}$$

Método de otimização: dividir e conquistar:

$$b_4 = a_4 = 1, (\underline{1}0111)_2$$

$$\begin{aligned} 1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 &= \\ 2 \times (1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0) + 1 &= \\ 2 \times (2 \times (1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0) + 1) + 1 &= \\ 2 \times (2 \times (2 \times (1 \times 2^1 + 0 \times 2^0) + 1) + 1) + 1 &= \\ 2 \times (2 \times (2 \times (2 \times (1) + 0) + 1) + 1) + 1 & \end{aligned}$$

Conversão binária para decimal

Análise: Quantos produtos existem no cálculo da conversão a seguir?

$$(10111)_2 = 1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = (23)_{10}$$

Método de otimização: dividir e conquistar:

$$b_4 = a_4 = 1, (\underline{1}0111)_2$$

$$\begin{aligned} 1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 &= \\ 2 \times (1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0) + 1 &= \\ 2 \times (2 \times (1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0) + 1) + 1 &= \\ 2 \times (2 \times (2 \times (1 \times 2^1 + 0 \times 2^0) + 1) + 1) + 1 &= \\ 2 \times (2 \times (2 \times (2 \times (1) + 0) + 1) + 1) + 1 & \end{aligned}$$

$$\begin{aligned} b_{j-1} &= 2 \times b_j + a_{j-1} \\ b_0 &= 2 \times b_1 + a_0 \\ b_1 &= 2 \times b_2 + a_1 \\ b_2 &= 2 \times b_3 + a_2 \\ b_3 &= 2 \times b_4 + a_3 \end{aligned}$$

Conversão binária para decimal

Análise: Quantos produtos existem no cálculo da conversão a seguir?

$$(10111)_2 = 1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = (23)_{10}$$

Método de otimização: dividir e conquistar:

$$b_4 = a_4 = 1, (\underline{1}0111)_2$$

$$\begin{aligned} 1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 &= \\ \mathbf{2 \times (1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0) + 1} &= \\ 2 \times (2 \times (1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0) + 1) + 1 &= \\ 2 \times (2 \times (2 \times (1 \times 2^1 + 0 \times 2^0) + 1) + 1) + 1 &= \\ 2 \times (2 \times (2 \times (2 \times (1) + 0) + 1) + 1) + 1 & \end{aligned}$$

$$\begin{aligned} b_{j-1} &= 2 \times b_j + a_{j-1} \\ \mathbf{b_0} &= \mathbf{2 \times b_1 + a_0 = (23)_{10}} \\ b_1 &= 2 \times b_2 + a_1 \\ b_2 &= 2 \times b_3 + a_2 \\ b_3 &= 2 \times b_4 + a_3 \end{aligned}$$

Conversão decimal para binária

Considere o número $(347)_{10}$ o qual queremos representar na forma binária $(a_j, a_{j-1}, \dots, a_2, a_1)_2$:

Temos que:

$$(347)_{10} = 2 \times (a_j \times 2^{j-1} + a_{j-1} \times 2^{j-2} + \dots + a_2 \times 2 + a_1) + a_0$$

Lembrem-se! $(a_j \times 2^{j-1} + a_{j-1} \times 2^{j-2} + \dots + a_2 \times 2 + a_1)$ é a função b_n , que na prática, representa a metade do valor decimal referente à b_{n-1} , ou seja:

$$(347)_{10} = 2 \times 173 + 1 \quad \Leftrightarrow \quad (347)_{10} = 2 \times 173 + a_0$$

...

Conversão decimal para binária

$$(347)_{10} = 2 \times 173 + 1 \quad \Leftrightarrow \quad (347)_{10} = 2 \times 173 + a_0$$

$$(173)_{10} = 2 \times 86 + 1 \quad \Leftrightarrow \quad (173)_{10} = 2 \times 86 + a_1$$

$$(86)_{10} = 2 \times 43 + 0 \quad \Leftrightarrow \quad (86)_{10} = 2 \times 43 + a_2$$

$$(43)_{10} = 2 \times 21 + 1 \quad \Leftrightarrow \quad (43)_{10} = 2 \times 21 + a_3$$

$$(21)_{10} = 2 \times 10 + 1 \quad \Leftrightarrow \quad (21)_{10} = 2 \times 10 + a_4$$

$$(10)_{10} = 2 \times 5 + 0 \quad \Leftrightarrow \quad (10)_{10} = 2 \times 5 + a_5$$

$$(5)_{10} = 2 \times 2 + 1 \quad \Leftrightarrow \quad (5)_{10} = 2 \times 2 + a_6$$

$$(2)_{10} = 2 \times 1 + 0 \quad \Leftrightarrow \quad (2)_{10} = 2 \times 1 + a_7$$

$$(2)_{10} = 2 \times 0 + 1 \quad \Leftrightarrow \quad (2)_{10} = 2 \times 0 + a_8$$

Conversão decimal para binária

$$(347)_{10} = (a_j, a_{j-1}, \dots, a_2, a_1)_2 = (1, 0, 1, 0, 1, 1, 0, 1, 1)_2$$

Conversão número fracionário base 10 para binária

Vamos considerar o primeiro caso, um número r com representação **finita** na base 10, e **finita** na base 2:

$$r = (0.125)_{10}$$

que vamos representar na forma binária por

$$(0.d_1d_2d_3\dots)_2$$

Assim,

$$0.125 = d_1 \times 2^{-1} + d_2 \times 2^{-2} + d_3 \times 2^{-3} + \dots$$

Conversão número fracionário base 10 para binária

Usando manipulação matemática, podemos multiplicar ambos os lados por 2:

$$2 \times 0.125 = 2 \times (d_1 \times 2^{-1} + d_2 \times 2^{-2} + d_3 \times 2^{-3} + \dots)$$

$$0.250 = d_1 + d_2 \times 2^{-1} + d_3 \times 2^{-2} + \dots$$

Observem que

$$0.250 = 0 + 0.250 = d_1 + d_2 \times 2^{-1} + d_3 \times 2^{-2} + \dots$$

Dessa forma, d_1 representa a parte inteira de 0.25 ($d_1=0$) e $d_2 \times 2^{-1} + d_3 \times 2^{-2}$ representa a parte fracionária (.250).

Conversão número fracionário base 10 para binária

Prosseguindo com a manipulação matemática...

$$2 \times 0.25 = 2 \times (0 + d_2 \times 2^{-1} + d_3 \times 2^{-2} + \dots)$$

$$0.5 = 0 + 0 + d_3 \times 2^{-1} + \dots$$

$$2 \times 0.5 = 2 \times (0 + 0 + d_3 \times 2^{-1} + \dots)$$

$$1.\mathbf{0} = 0 + 0 + 1 + \dots$$

Como a parte fracionária é igual à **zero**, o procedimento termina.

Observem que o resultado para $d_1 = 0$, $d_2 = 0$ e $d_3 = 1$.

Conversão número fracionário base 10 para binária

Vamos considerar o segundo caso, um número s com representação **finita** na base 10, mas **infinita** na base 2:

$$r = (0.1)_{10}$$

vamos á manipulação matemática, agora, de forma resumida:

$$2 \times 0.1 = 0.2, d_1 = 0,$$

$$2 \times 0.2 = 0.4, d_2 = 0,$$

$$2 \times 0.4 = 0.8, d_3 = 0,$$

$$2 \times 0.8 = \mathbf{1.6}, d_4 = \mathbf{1}, \text{ opa! Observe o resto} = 0.6.$$

$$2 \times 0.6 = 1.2, d_5 = 1, \text{ o resto voltou a ser } 0.2, \text{ o processo vai repetir...}$$

Conversão número fracionário base 10 para binária

$$r = (0.1)_{10}$$

$$2 \times 0.1 = 0.2, d_1 = 0,$$

$$2 \times 0.2 = 0.4, d_2 = 0,$$

$$2 \times 0.4 = 0.8, d_3 = 0,$$

$$2 \times 0.8 = \mathbf{1.6}, d_4 = \mathbf{1}, \text{ opa! Observe o resto} = 0.6.$$

$$2 \times 0.6 = 1.2, d_5 = 1, \text{ o resto voltou a ser } 0.2, \text{ o processo vai repetir...}$$

$$2 \times 0.2 = 0.4, d_6 = 0, \text{ ...a partir de } d_2, \text{ a saber } 0011...0011...$$

...

$$\text{logo, } r = (0.1)_{10} = 0.0001\overline{100110011}....$$

Conversão número fracionário base 10 para binária

$$\text{logo, } r = (0.1)_{10} = 0.0001100110011....$$

Como o computador possui um limite de dígitos para representação da parte fracionária, uma aproximação será utilizada para realizar os cálculos que envolvam o número $(0.1)_{10}$.

Dessa forma, não há como o resultado de operações que envolvam números com representação binária infinita fornecerem uma resposta exata.

Conversão binário para número fracionário base 10

Vamos considerar o a conversão de um número fracionário representado em base 2 para um número fracionário em base 10:

$$r = (0.000111)_2$$

a manipulação matemática é muito simples e similar á realizada anteriormente, claro, considerando multiplicar r por 10, que é a base a qual se deseja converter.

Contudo, estamos trabalhando com números binários. Logo, vamos representar 10 em binário nas operações.

$$(10)_{10} = (1010)_2$$

Conversão binário para número fracionário base 10

$$r = (0.000111)_2$$

vamos á manipulação matemática:

$$(1010)_2 \times (0.000111)_2 = (1.00011)_2, \quad d_1 = (1)_{10},$$

$$(1010)_2 \times (1.00011)_2 = (0.1111)_2, \quad d_2 = (0)_{10},$$

$$(1010)_2 \times (0.1111)_2 = (1001.011)_2, \quad d_3 = (9)_{10},$$

$$(1010)_2 \times (1001.011)_2 = (11.11)_2, \quad d_4 = (3)_{10},$$

$$(1010)_2 \times (11.11)_2 = (111.1)_2, \quad d_5 = (7)_{10},$$

$$(1010)_2 \times (111.11)_2 = (101)_2, \quad d_6 = (5)_{10},$$

A parte fracionária é igual a zero. Fim do procedimento.

$$r = (0.000111)_2 = (0.109375)_{10}$$

Aritmética de ponto flutuante

Ponto flutuante: uma representação computacional para números reais.

forma:

$$\pm(d_1d_2d_t\dots) \times \beta^e$$

Onde: β é a base em que a máquina opera;
e é o expoente no intervalo $[l, u]$;
t é o número de dígitos na mantissa;

Em computadores apenas um conjunto finito de números reais podem ser representados.

Aritmética de ponto flutuante

Em computadores apenas um conjunto finito de números reais podem ser representados.

Se o número real não possui representação exata em um dos elementos desse conjunto, o número é representado de forma truncada ou arredondada.

Por exemplo, o sistema: $t = 3$, $\beta = 10$, $e \in [-5, 5]$, $0 \leq d_j \leq \beta - 1$

representação no sistema: $0.d_1d_2d_3 \times 10^e$, $0 \leq d_j \leq 9$, $e \in [-5, 5]$

Menor representação (m)

$$m = 0.100 \times 10^{-5} = 10^{-6}$$

Maior representação (M)

$$M = 0.999 \times 10^{-5} = 99900$$

Aritmética de ponto flutuante

Considerando um conjunto $G = \{x \in \mathbb{R} \mid m \leq |x| \leq M\}$,

Se $x \in G$, por exemplo, $x = 235.89 = 0.23589 \times 10^3$.

Considerando $t = 3$, x poderá ser representado por 0.235 (truncamento) ou por 0.236 (arredondamento).

Se $x < m$, por exemplo, $x = 0.345 \times 10^{-7}$, este número não pode ser representado pelo sistema (slide anterior). O sistema acusa *underflow*.

Se $x > M$, por exemplo, $x = 0.875 \times 10^9$, o expoente é maior do que 5 e o sistema acusa *overflow*.

Aritmética de ponto flutuante

Alguns sistemas permitem uma representação do número real em **precisão dupla**, ou seja, a mantissa representada em ponto flutuante é armazenada com o dobro do número de dígitos.

Isso causa aumento do uso de memória e requer maior tempo de processamento.

O zero em ponto flutuante é representado com o menor número expoente possível no sistema, pois a adição de zero multiplicado pela base com qualquer expoente pode causar perdas de dígitos significativos nos resultados da soma deste zero com outro número.

Isto será observado no próximo tópico: Erros.

Erros

Erro absoluto (EA)

É a diferença entre o valor exato de um número x e de seu valor aproximado \bar{x} .

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

Em muitos casos apenas o valor aproximado é conhecido.

Devido a isso, obtemos um **limitante superior** para o módulo do EA.

Por exemplo, para $PI \in (3.14, 3.15)$ dizemos que PI é algum valor dentro deste intervalo.

Consequentemente, teremos $|EA_x| = |PI - \bar{PI}| < 0.01$

Erro absoluto (EA)

Agora seja $\bar{x} = 2112.9$ com $EA_x < 0.1$ para $x \in (2112.9, 2113)$. E seja $\bar{y} = 5.3$ com $EA_y < 0.1$ para $y \in (5.2, 5.4)$.

Sabendo que os limitantes superiores para x e y são os mesmos, pode-se afirmar que ambos estão representados com a mesma precisão?

Erro absoluto (EA)

Agora seja $\bar{x} = 2112.9$ com $EA_x < 0.1$ para $x \in (2112.9, 2113)$. E seja $\bar{y} = 5.3$ com $EA_y < 0.1$ para $y \in (5.2, 5.4)$.

Sabendo que os limitantes superiores para x e y são os mesmos, pode-se afirmar que ambos estão representados com a mesma precisão?

A resposta é não. Pois é preciso comparar a **ordem de grandeza** de x e y .

Para o mesmo valor limitante, o número que possui **maior** ordem de grandeza é representado com maior precisão.

Erro relativo (ER)

Devido ao EA não ser suficiente (depende da grandeza dos valores envolvidos) para descrever a precisão de um cálculo, utilizamos o erro relativo.

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

O erro relativo nos diz o quanto o erro absoluto (**incerteza**) tem relação com o valor aproximado.

Agora visualizamos o erro (ER) como uma incerteza sobre o valor aproximado.

Erro relativo (ER)

No exemplo anterior, x possui **maior** precisão (menor ER) do que y.

$$|ER_x| = \frac{|EA_x|}{|\bar{x}|} < \frac{0.1}{2112.9} \approx 4.7 \times 10^{-5}$$

$$|ER_y| = \frac{|EA_y|}{|\bar{y}|} < \frac{0.1}{5.3} \approx 0.02,$$

Erros de arredondamento e truncamento em um sistema de ponto flutuante

Considere um **sistema** ($\beta=10$, $e=3$, $t=4$) que represente números x em ponto flutuante através da seguinte **aritmética**:

$$x = f_x \times 10^e + g_x \times 10^{e-t} \quad \text{onde } 0.1 \leq f_x < 1 \quad \text{e} \quad 0 \leq g_x < 1$$

Se $x = 234.57$, então o sistema representaria x dessa forma:

$$x = 0.2345 \times 10^3 + 0.7 \times 10^{-1}, \quad \text{donde } f_x = 0.2345 \quad \text{e} \quad g_x = 0.7$$

Observem que nesse sistema $t=4$ e, por isso, $g_x \times 10^{e-t}$ não pode ser incorporado diretamente à mantissa.

Portanto, o sistema deve realizar o **truncamento** ou **arredondamento** de x .

Erros de arredondamento e truncamento em um sistema de ponto flutuante

Considerando o **truncamento** de x , $g_x \times 10^{e-t}$ é desprezado e x assume o **valor aproximado** $\bar{x} = f_x \times 10^e$.

O **erro absoluto** é dado por $|EA_x| = |x - \bar{x}| = g_x \times 10^{e-t} < 10^{e-t}$ (pois $|g_x| < 1$).

O erro relativo é dado por $|ER_x| = |EA_x| / |\bar{x}| = g_x \times 10^{e-t} / f_x \times 10^e < 10^{e-t} / 0.1 \times 10^e = 10^{-t+1}$ (note que 0.1 é o menor valor para f_x).

O menor valor para f_x garante que o ER considerado seja o maior possível!

Obs.: Observe o porquê de o menor valor para f_x não poder ser zero! (ER=EA/0)

Erros de arredondamento e truncamento em um sistema de ponto flutuante

Considerando o **arredondamento** (simétrico) de x , $f_x \times 10^e$ é modificada e x assume o valor aproximado \bar{x} conforme o valor de $g_x \times 10^{e-t}$.

$$\bar{x} = \begin{cases} f_x \times 10^e & \text{se } |g_x| < \frac{1}{2} \\ f_x \times 10^e + 10^{e-t} & \text{se } |g_x| \geq \frac{1}{2} \end{cases}$$

Em resumo, $|g_x|$ é desprezado se for menor que $\frac{1}{2}$, caso contrário, adicionamos 1 ao último dígito de f_x .

Erros de arredondamento e truncamento em um sistema de ponto flutuante

Então, se temos $|g_x| < \frac{1}{2}$:

$$|EA_x| = |x - \bar{x}| = |g_x| \times 10^{e-t} < \frac{1}{2} \times 10^{e-t}$$

$$|ER_x| = \frac{|EA_x|}{|\bar{x}|} = \frac{|g_x| \times 10^{e-t}}{|f_x| \times 10^e} < \frac{0.5 \times 10^{e-t}}{0.1 \times 10^e} = \frac{1}{2} \times 10^{-t+1}$$

Erros de arredondamento e truncamento em um sistema de ponto flutuante

Então, se
temos $|g_x| \geq \frac{1}{2}$:

$$\begin{aligned} |EA_x| &= |x - \bar{x}| = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e + 10^{e-t})| \\ &= |g_x \times 10^{e-t} - 10^{e-t}| = |(g_x - 1)| \times 10^{e-t} \leq \frac{1}{2} \times 10^{e-t} \end{aligned}$$

Observem que
consideramos o
maior EA e o menor
valor para $|\bar{x}|$.

$$\begin{aligned} |ER_x| &= \frac{|EA_x|}{|\bar{x}|} \leq \frac{\frac{1}{2} \times 10^{e-t}}{|f_x \times 10^e + 10^{e-t}|} < \frac{\frac{1}{2} \times 10^{e-t}}{|f_x| \times 10^e} < \\ &< \frac{\frac{1}{2} \times 10^{e-t}}{0.1 \times 10^e} = \frac{1}{2} \times 10^{-t+1} \end{aligned}$$

Erros de arredondamento e truncamento em um sistema de ponto flutuante

Ainda considerando o arredondamento de x , o EA_x e o ER_x são:

$$|EA_x| \leq \frac{1}{2} \times 10^{e-t} \quad \text{e} \quad |ER_x| < \frac{1}{2} \times 10^{-t+1}$$

Para qualquer caso em que $\frac{1}{2} \leq |g_x| < 1$;

(consulte o livro referência na página 15 para a prova matemática).

Apesar de incorrer em menores erros, o uso de arredondamento acarreta um tempo maior de execução, por isso o truncamento é mais utilizado.

Análise de erros em operações aritméticas de ponto flutuante

Dois casos serão estudados a seguir:

- 1) considerando que os valores usados nas operações aritméticas são exatos;
- 2) considerando que os valores usados nas operações aritméticas são aproximações;

No caso 1) o EA e ER vão incidir apenas no resultado da operação.

No caso 2) o EA e ER vão incidir nos valores e no resultados da operação.

Análise de erros em operações aritméticas de ponto flutuante

Observem a sequência de operações $u = [(x + y) - z - t] / w$.

Note:

O erro total em cada operação aritmética é composto erro da aproximação que representa o número real mais o erro do resultado da operação.

O erro da operação aritmética se propaga para as outras operações subsequentes.

Análise de erros em operações aritméticas de ponto flutuante

Para a ADIÇÃO:

- Requer o alinhamento dos pontos decimais ($t_x = t_y$).
- A mantissa do menor expoente deve ser deslocada para a direita.

Exemplo: adição: $x = 0.937 \times 10^4$ e $y = 0.1272 \times 10^2$.

$$x = 0.937 \times 10^4 \text{ e } y = 0.\textcolor{red}{00}1272 \times 10^{2+2}$$

$$\text{Logo, } x + y = (0.937 + 0.001272) \times 10^4 = 0.938272 \times 10^4.$$

Análise de erros em operações aritméticas de ponto flutuante

$$x + y = (0.937 + 0.00937) \times 10^4 = 0.938272 \times 10^4.$$

como neste sistema $t=4$, o resultado da operação de adição deve ser truncado ou arredondado.

Considerando truncamento: $\overline{x + y} = 0.9382 \times 10^4$

Considerando arredondamento: $\overline{x + y} = 0.9383 \times 10^4$

Observe que **o resultado** é agora considerado uma aproximação. 71

Análise de erros em operações aritméticas de ponto flutuante

Para a MULTIPLICAÇÃO não há a necessidade de $t_x = t_y$.

Exemplo: multiplicação: $x = 0.937 \times 10^4$ e $y = 0.1272 \times 10^2$.

Logo, $xy = (0.937 \times 0.1272) \times 10^{4+2} = 0.1191864 \times 10^6$.

Considerando truncamento: $\overline{xy} = 0.1191 \times 10^6$

Considerando arredondamento: $\overline{xy} = 0.1192 \times 10^6$

Análise de erros em operações aritméticas de ponto flutuante

Até aqui consideramos que x e y foram representados com seus valores exatos nas operações de adição e multiplicação. Assim, o EA e ER incidem apenas no resultado da operação.

Resumindo, para valores exatos, temos, para as operações:

$|ER_{op}| < 10^{-t+1}$, considerando truncamento.

$|ER_{op}| < \frac{1}{2} \times 10^{-t+1}$, considerando arredondamento.

Análise de erros em operações aritméticas de ponto flutuante

Agora vamos considerar que $x = \bar{x} + EA_x$ e $y = \bar{y} + EA_y$ são aproximações.

para a adição:

$$x + y = (\bar{x} + EA_x) + (\bar{y} + EA_y) = (\bar{x} + \bar{y}) + (EA_x + EA_y)$$

$$EA_{x+y} = (EA_x + EA_y)$$

$$\begin{aligned} ER_{x+y} &= \frac{EA_{x+y}}{\bar{x} + \bar{y}} = \frac{EA_x}{\bar{x}} \left(\frac{\bar{x}}{\bar{x} + \bar{y}} \right) + \frac{EA_y}{\bar{y}} \left(\frac{\bar{y}}{\bar{x} + \bar{y}} \right) = \\ &= ER_x \left(\frac{\bar{x}}{\bar{x} + \bar{y}} \right) + ER_y \left(\frac{\bar{y}}{\bar{x} + \bar{y}} \right). \end{aligned}$$

Deduzam para a subtração.

Análise de erros em operações aritméticas de ponto flutuante

Agora vamos considerar que $x = \bar{x} + EA_x$ e $y = \bar{y} + EA_y$ são aproximações.

para a multiplicação:

$$xy = (\bar{x} + EA_x)(\bar{y} + EA_y) = \bar{x}\bar{y} + \bar{x}EA_y + \bar{y}EA_x + EA_xEA_y$$

Note que EA_xEA_y é um número muito pequeno e pode ser desprezado.

$$EA_{xy} \approx \bar{x}EA_y + \bar{y}EA_x$$

$$ER_{xy} \approx \frac{\bar{x}EA_y + \bar{y}EA_x}{\bar{x}\bar{y}} = \frac{EA_x}{\bar{x}} + \frac{EA_y}{\bar{y}} = ER_x + ER_y$$

Resolução de Sistemas Lineares

Resolução de Sistemas Lineares

Muitos problemas do dia-a-dia podem ser modelados como sistemas lineares;

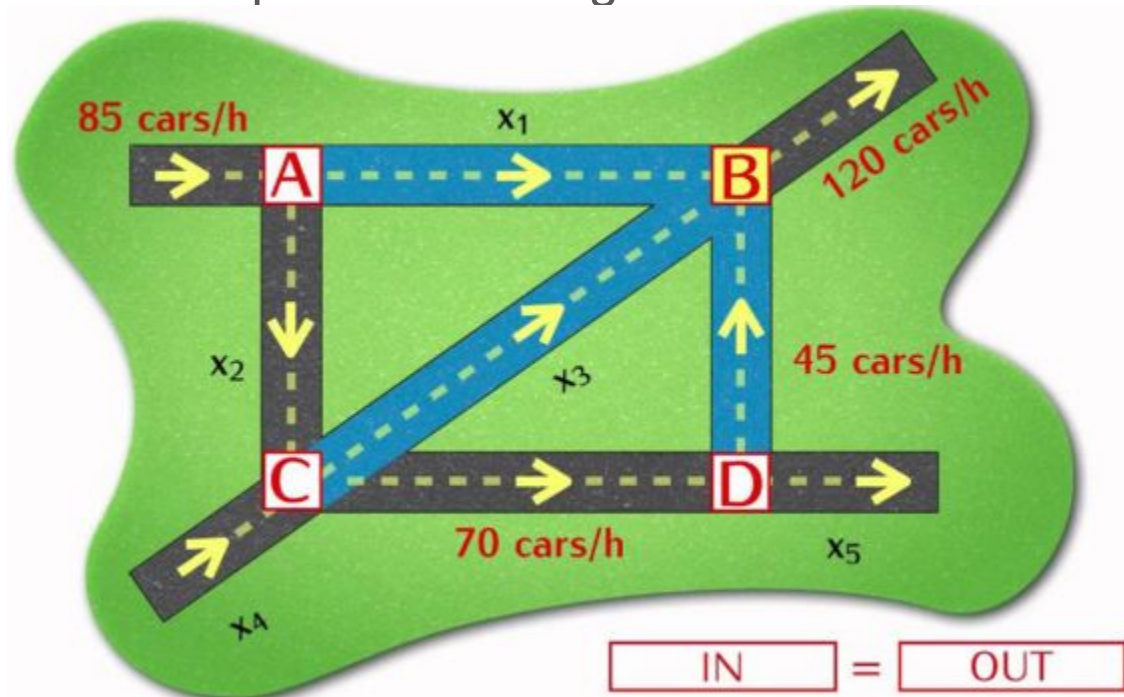
Exemplos:

- Tráfego de automóveis considerando várias vias intercruzadas;
- Calcular as forças que atuam em junções de um conjunto de treliças;

Ou seja, problemas que possuem várias variáveis inter relacionadas em subproblemas similares;

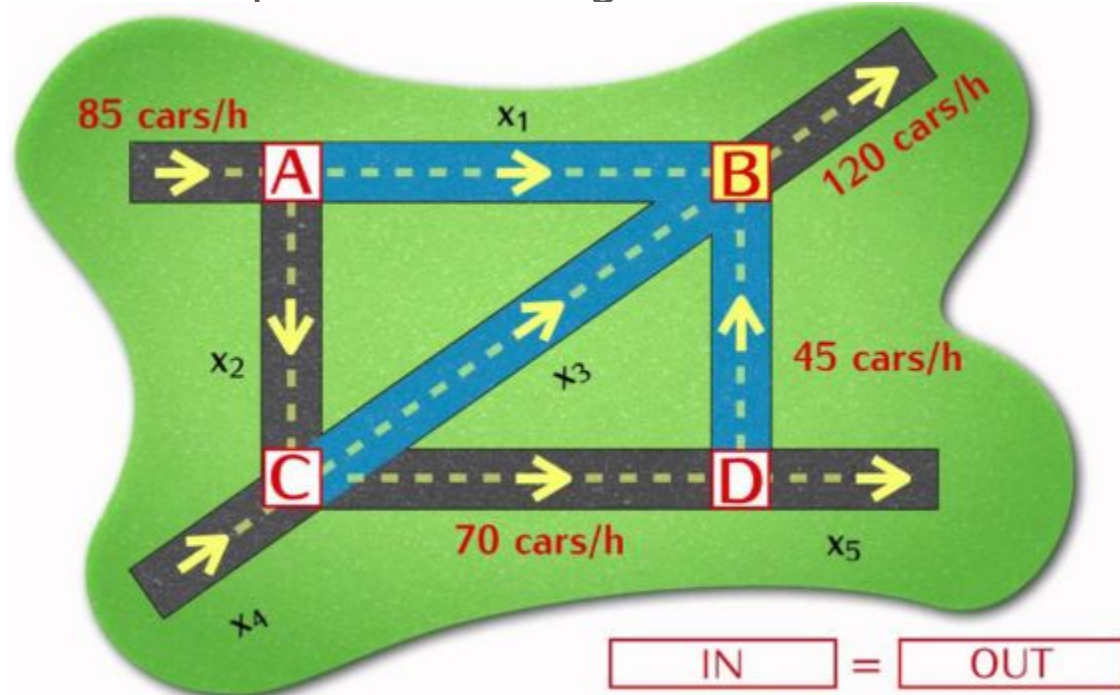
Resolução de Sistemas Lineares

Modelar o problema a seguir utilizando sistema linear:



Resolução de Sistemas Lineares

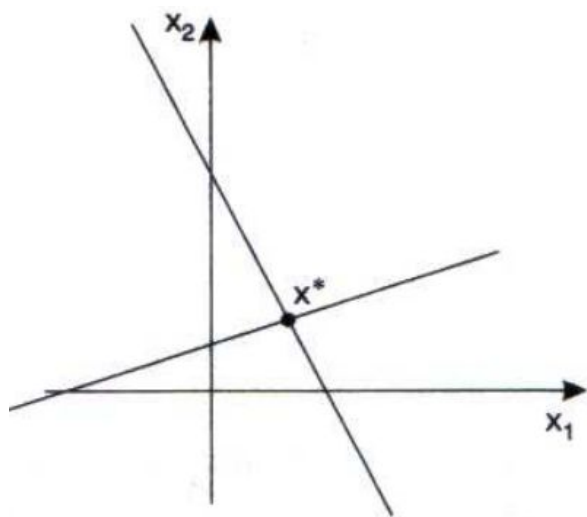
Modelar o problema a seguir utilizando sistema linear:



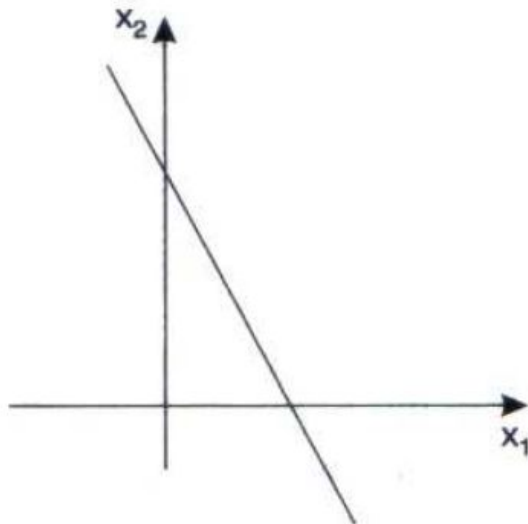
$$\begin{cases} x_4 - x_5 = 35 \\ x_1 + x_2 = 85 \\ x_1 + x_3 = 75 \\ x_2 - x_3 + x_4 = 70 \\ x_5 = 25 \end{cases}$$

Resolução de Sistemas Lineares

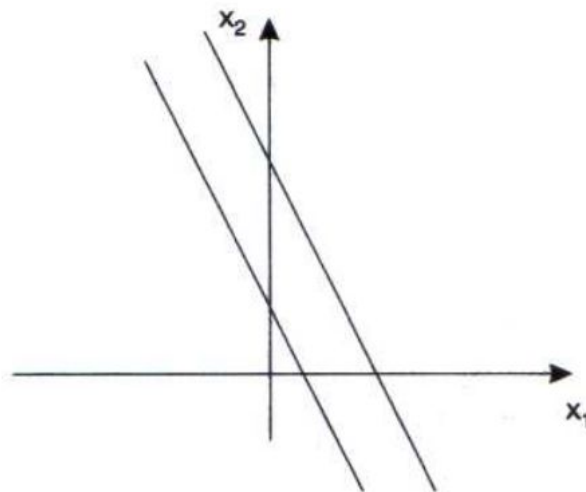
Geometricamente (e intuitivamente), as diversas equações de um sistema linear são representadas por retas em um espaço.



(1) retas concorrentes



(2) retas coincidentes



(3) retas paralelas

A solução(ões) ocorre(m) na(s) interseção(ões) entre as retas.

Resolução de Sistemas Lineares

Também podemos representar matricialmente os sistemas lineares.

$$\mathbf{Ax} = \mathbf{b}$$

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

matriz de coeficientes

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}$$

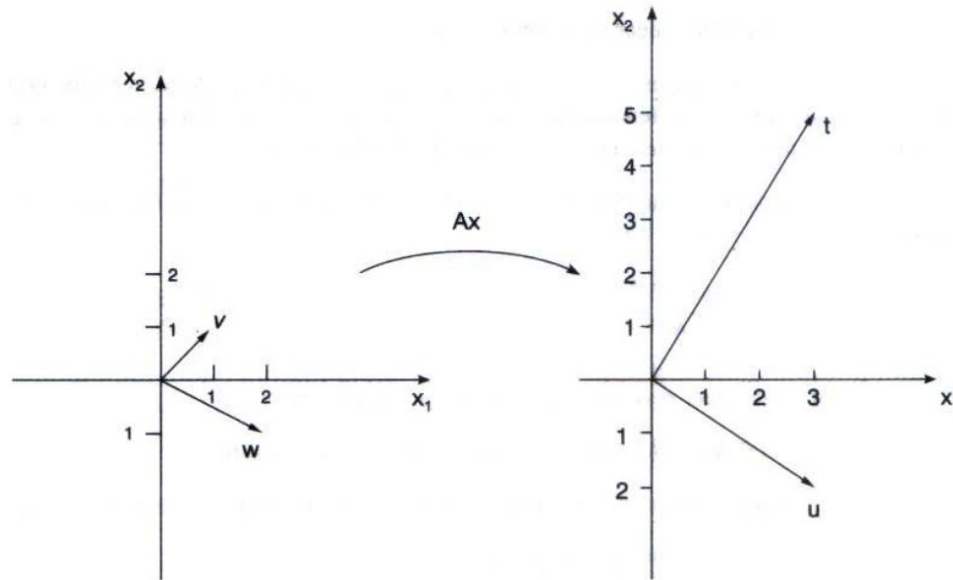
vetor solução

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_m \end{pmatrix}$$

vetor de constantes
do problema

Resolução de Sistemas Lineares

Podemos também visualizar que a matriz de coeficientes **A** mapeia o vetor de soluções **x*** para o vetor de constantes **b**, comportando-se como uma função;



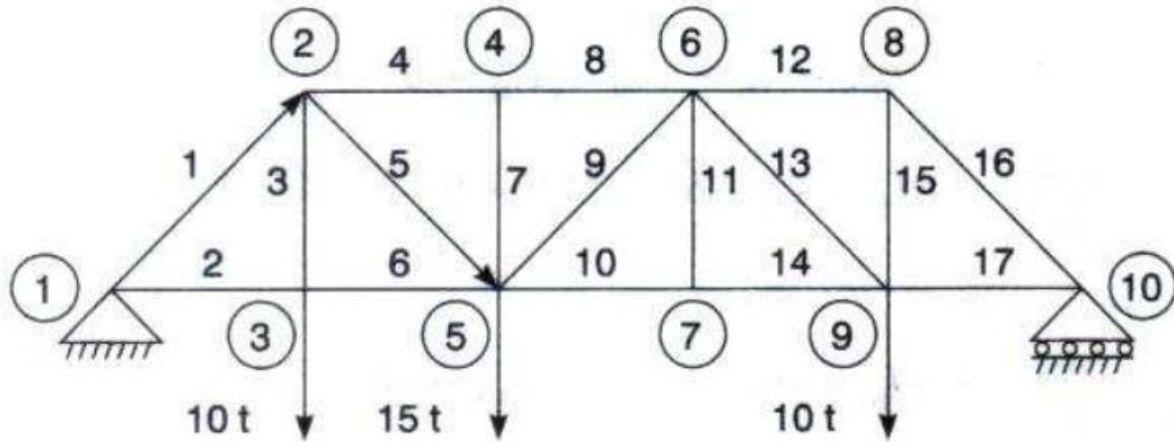
$$A = \begin{pmatrix} 2 & 1 \\ 1 & -3 \end{pmatrix}$$

$$\text{se } v = (1 \ 1)^T \quad \text{então: } u = Av = (3 \ -2)^T;$$

$$\text{se } w = (2 \ -1)^T \quad \text{então: } t = Aw = (3 \ 5)^T;$$

Resolução de Sistemas Lineares

Modelar outro problema mais complexo a seguir utilizando sistema linear:



Dica:

$$\text{Junção 2} \begin{cases} \Sigma F_x = -\alpha f_1 + f_4 + \alpha f_5 = 0 \\ \Sigma F_y = -\alpha f_1 - f_3 - \alpha f_5 = 0 \end{cases}$$

Imagem da matriz de coeficientes (A)

Dado uma matriz A: $m \times n$, a Imagem de A é dada por:

$$\text{Im}(A) = \{ \mathbf{b} \in \mathbb{R}^m \mid \exists \mathbf{x} \in \mathbb{R}^n \mid \mathbf{b} = \mathbf{A}\mathbf{x} \}$$

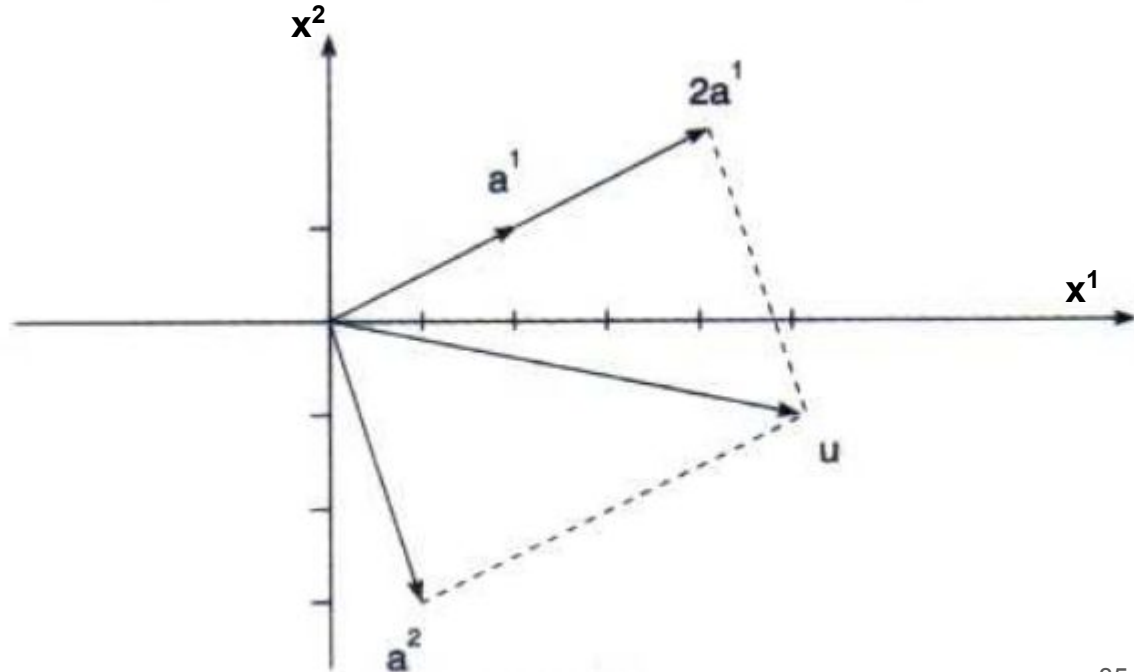
A imagem de A é formada por elementos **b** de um subconjunto \mathbb{R}^m , onde **existe pelo menos um** elemento **x** de outro subconjunto \mathbb{R}^n o qual a relação **$\mathbf{b} = \mathbf{A}\mathbf{x}$** é satisfeita.

Nota: Tudo se resume em representar o vetor **b** como uma combinação linear **$\mathbf{A}\mathbf{x}$** .

Sistema Linear compatível determinado

Observem o sistema linear a seguir:

Descrevam algebricamente este sistema em função de u , $A=\{a^1, a^2\}$, x^1 e x^2 .



Sistema Linear compatível determinado

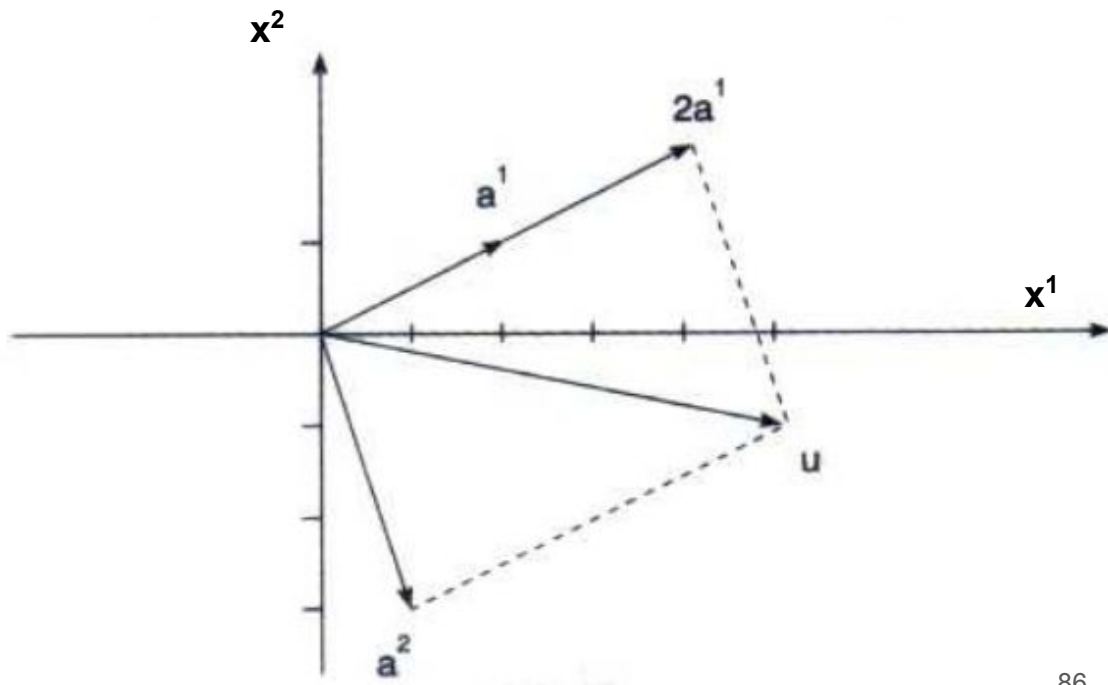
Observem o sistema linear a seguir:

Descrevam algebricamente este sistema em função de u , $A = \{a^1, a^2\}$, x^1 e x^2 .

a^1, a^2 são linearmente independentes.

Dado qualquer valor do vetor $u \in \text{Im}(A)$, x^1 e x^2 serão únicos (solução única).

Solução única.



Sistema Linear compatível determinado

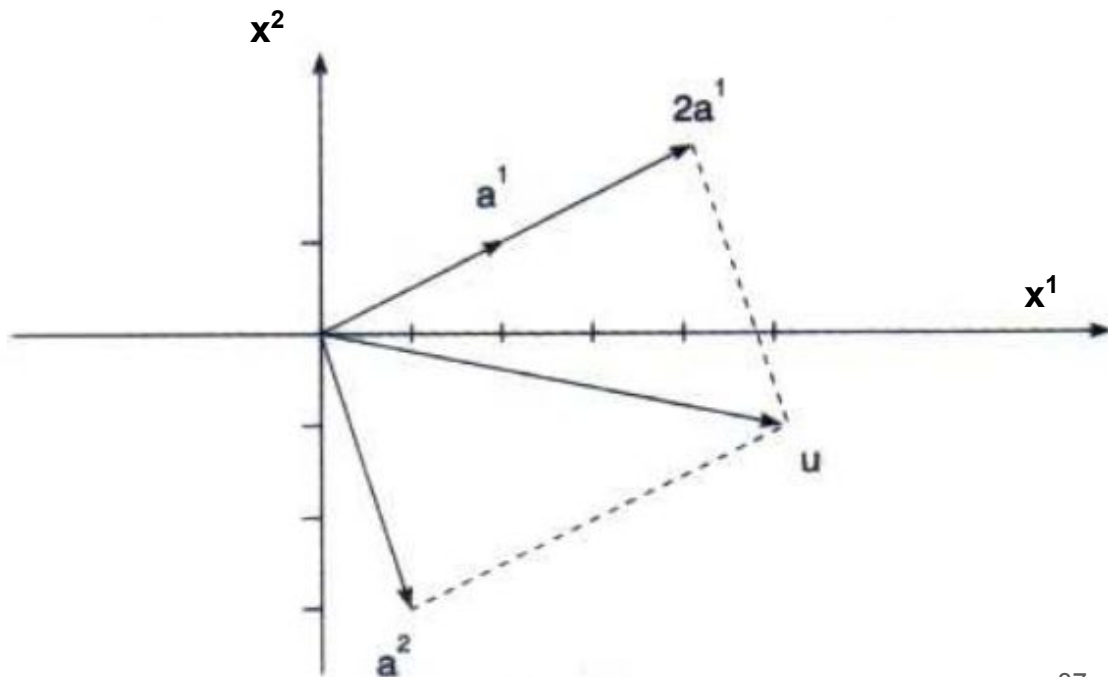
Observem o sistema linear a seguir:

Descrevam algebricamente este sistema em função de u , $A = \{a^1, a^2\}$, x^1 e x^2 .

a^1, a^2 são linearmente independentes.

Dizemos que A forma uma base em \mathbb{R}^2 .

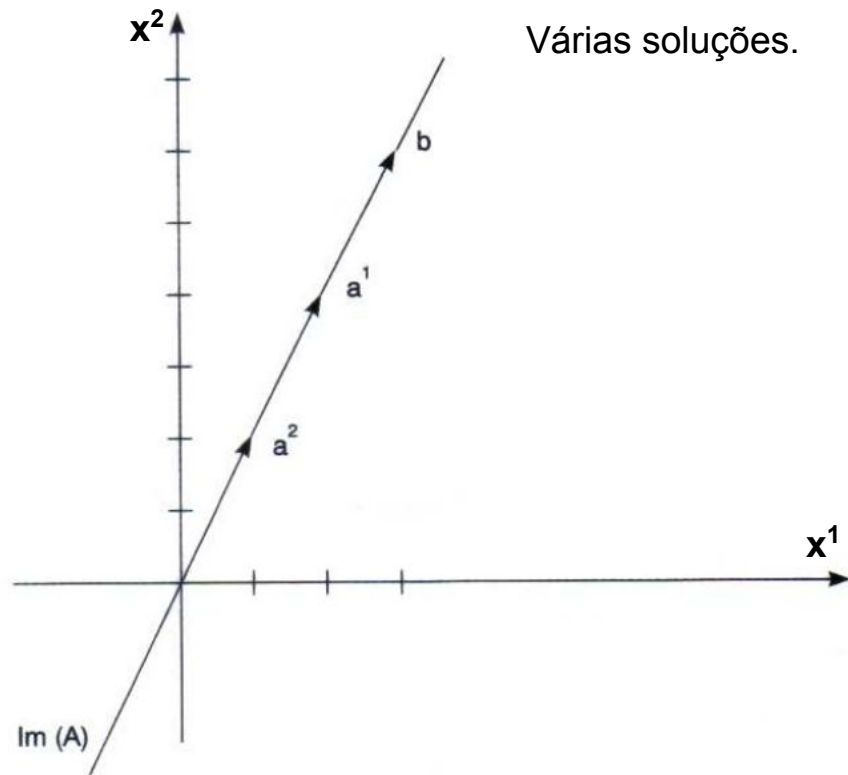
Solução única.



Sistema Linear compatível indeterminado

Observem o sistema linear a seguir:

Descrevam algebricamente este sistema em função de u , $A=\{a^1, a^2\}$, x^1 e x^2 .



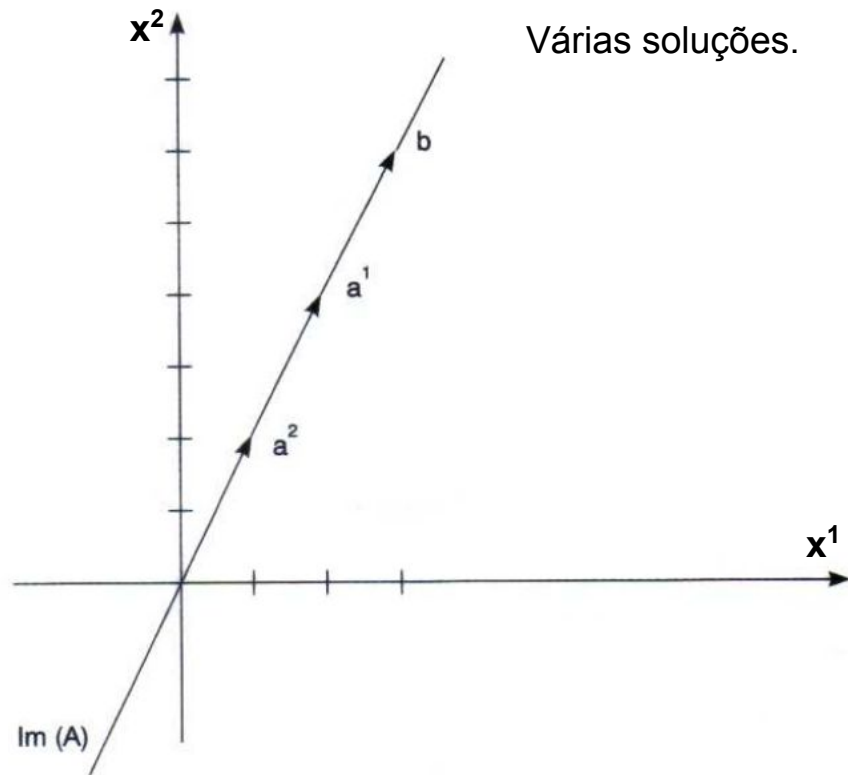
Sistema Linear compatível indeterminado

Observem o sistema linear a seguir:

Descrevam algebricamente este sistema em função de u , $A=\{a^1, a^2\}$, x^1 e x^2 .

a^1, a^2 são linearmente dependentes.

Dado qualquer valor do vetor $u \in \text{Im}(A)$, x^1 e x^2 admitem infinitas soluções.



Sistema Linear compatível indeterminado

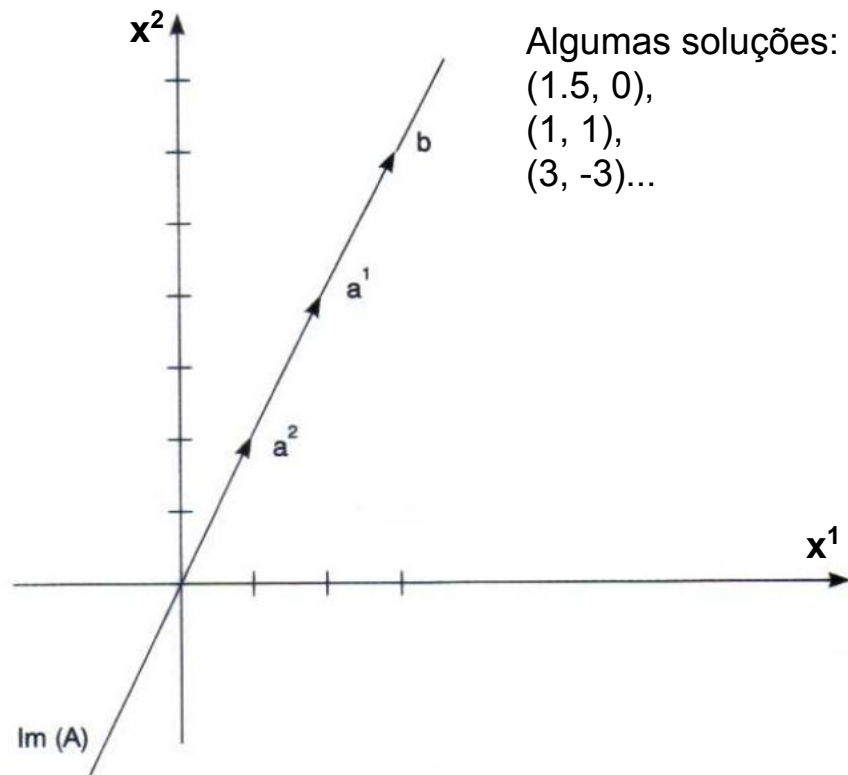
Observem o sistema linear a seguir:

Descrevam algebricamente este sistema em função de u , $A = \{a^1, a^2\}$, x^1 e x^2 .

a^1, a^2 são linearmente dependentes.

Dizemos que **A NÃO** forma uma base em \mathbb{R}^2 .

Observem que $ca^2 = a^1 \mid c \in \mathbb{Z}$.



Sistema Linear incompatível

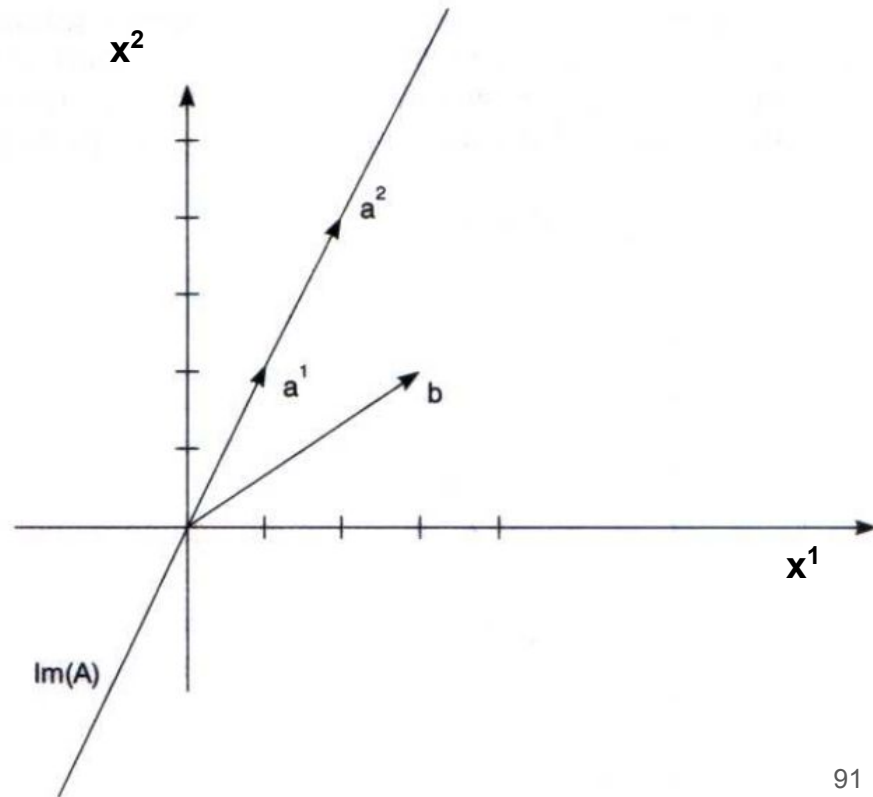
Observem o sistema linear a seguir:

Descrevam algebricamente este sistema em função de u , $A = \{a^1, a^2\}$, x^1 e x^2 .

a^1, a^2 são linearmente dependentes.

Dizemos que A **NÃO** forma uma base em \mathbb{R}^2 .

Observem que $ca^2 = a^1 \mid c \in \mathbb{Z}$
e $u \notin \text{Im}(A)$.



Posto da matriz de coeficientes (A)

O **Posto** de A equivale à dimensão de $\mathbf{b} = \text{Im}(A)$, ou $\text{Dim}(\text{Im}(A))$.

Vimos casos em que $\text{Dim}(\text{Im}(A)) = \text{Dim}(x^*)$, ou a dimensão do vetor solução x^* é igual à dimensão do vetor de constantes b .

A seguir, vamos investigar quando:

- $\text{Dim}(\text{Im}(A)) < \text{Dim}(x^*)$
- $\text{Dim}(\text{Im}(A)) > \text{Dim}(x^*)$

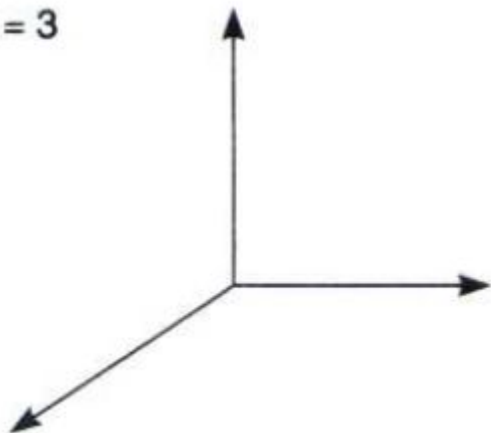
Para esses dois casos, consideramos o Posto de A como:
 $\text{posto}(A) \leq \min\{\text{Dim}(\text{Im}(A)) , \text{Dim}(x^*)\}$

Posto da matriz de coeficientes (A)

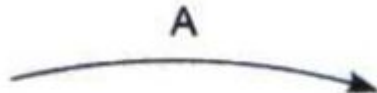
Fazendo $\text{Dim}(\text{Im}(A)) = m$ e $\text{Dim}(x^*) = n$:

Se $m < n$, o sistema linear nunca poderá ter solução única, pois $\text{posto}(A) < n$ sempre. Em outras palavras, n contém muitos m 's;

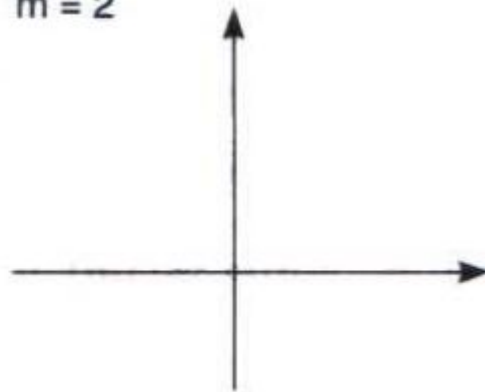
$n = 3$



A



$m = 2$



Posto da matriz de coeficientes (A)

Exemplo para $m < n$:

$$\begin{cases} -x_1 + 2x_2 + 3x_3 = 6 \\ x_2 + x_3 = 9 \end{cases}$$

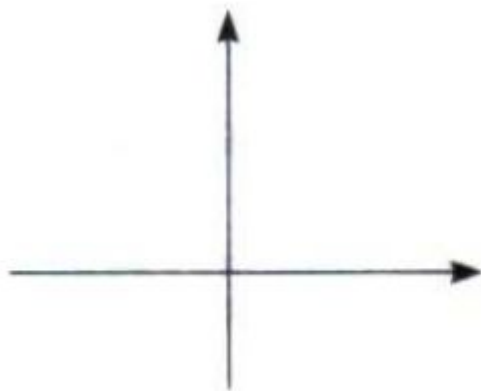
Observem que o sistema é compatível indeterminado.

Posto da matriz de coeficientes (A)

Fazendo $\text{Dim}(\text{Im}(A)) = m$ e $\text{Dim}(x^*) = n$:

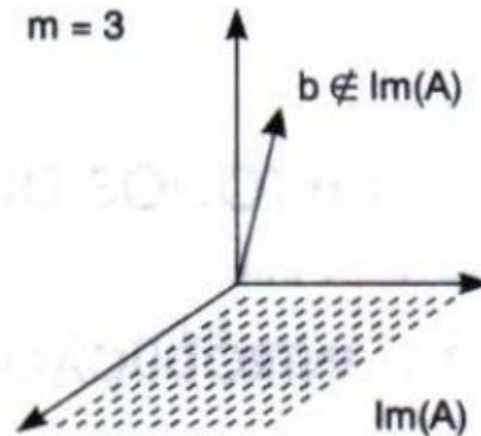
Se $m > n$, mesmo que $n = \text{posto}(A)$, o sistema linear poderá não ter solução, pois frequentemente ocorrerá a situação $b \notin \text{Im}(A)$;

$n = 2$



A

$m = 3$



Resumo de sistemas lineares

Matriz A		$m = n$	$m < n$	$m > n$
Posto Completo		$(\text{posto}(A) = n)$ Compatível determinado	$(\text{posto}(A) = m)$ Infinitas soluções	$(\text{posto}(A) = n)$ $b \in \text{Im}(A)$, solução única $b \notin \text{Im}(A)$, incompatível
Posto Deficiente	$b \in \text{Im}(A)$	Infinitas soluções	Infinitas soluções	Infinitas soluções
	$b \notin \text{Im}(A)$	Incompatível	Incompatível	Incompatível

Resolução de Sistemas Lineares

Métodos Diretos

Métodos Diretos

Imagine um sistema linear $n \times n$, que envolve o cálculo de $(n + 1)$ determinantes de ordem n (Regra de Cramer-Sarrus).

Se $n=20$, o número total de multiplicações seria $21 \times 20! \times 19$, mais um número aproximado de adições .

Um computador que efetue 100 milhões de multiplicações por segundo levaria 3×10^5 anos para realizar tais operações.

Além disso, para resolver um sistema linear com vetor de solução x^* única, faz-se $x^*=A^{-1} \times b$, onde A^{-1} é a matriz inversa da matriz A . O número de operações para calcular a inversa de A e em seguida o produto entre A^{-1} e b é desaconselhável em comparação com os métodos diretos vistos a seguir.

Matriz inversa é única para cada matriz que possui matriz inversa (solução única);

Métodos Diretos

Os métodos diretos calculam a solução de um problema num número finito de passos.

A resposta para estes métodos seriam precisas se considerada a precisão aritmética infinita.

Exemplos incluem a eliminação de Gauss, o método de fatoração QR para sistemas de equações lineares e o método simplex de programação linear.

Método da Eliminação de Gauss

Dado um sistema linear qualquer da forma:

$$a_{11}.x_1 + a_{12}.x_2 + \dots + a_{1n}.x_n = b_1$$

$$a_{21}.x_1 + a_{22}.x_2 + \dots + a_{2n}.x_n = b_2$$

...

...

$$a_{n1}.x_1 + a_{n2}.x_2 + \dots + a_{nn}.x_n = b_n$$

o método de Gauss consiste em fazer operações entre linhas deste sistema até chegarmos a um novo sistema (que terá a mesma solução que o inicial) com a forma triangular:

$$a'_{11}.x_1 + a'_{12}.x_2 + \dots + a'_{1n}.x_n = b'_1$$

$$a'_{22}.x_2 + \dots + a'_{2n}.x_n = b'_2$$

...

...

$$a'_{nn}.x_n = b'_n$$

Método da Eliminação de Gauss

Então, desde que se consiga levar o sistema na forma triangular sem alterar suas soluções, o nosso problema está resolvido.

Regra de Cramer X Eliminação de Gauss

Tanto o método da eliminação de Gauss quanto o de Cramer são métodos diretos que, ao contrário dos métodos iterativos, fornecem um resultado exato, afora grandes erros de arredondamentos computacionais para sistemas muito grandes, onde não são recomendados esses métodos, e sim os iterativos, que apesar de não fornecer um resultado exato, fornecem uma aproximação muito boa.

O método da eliminação de Gauss, considerando n o número de variáveis, tem um custo de ordem cúbica (n^3), este valor é bastante reduzido se comparado com o número de operações que seria necessário efetuar caso resolvêssemos o sistema pela regra de Cramer, que tem um custo assintótico de $\Theta(n!)$. Para $n = 10$, por exemplo, teríamos que efetuar aproximadamente 40 milhões de operações, pela regra de Cramer, ao invés de 430 pelo Método de Gauss.

Método da Eliminação de Gauss

Consideremos outra vez o sistema. Os passos a executar são:

1. considerar a 1a linha como base para a eliminação;
2. zerar todos os coeficientes da 1a coluna abaixo da a_{11} ;
3. zerar todos os coeficientes abaixo da diagonal principal; para isso vamos fazer:
 - 3.1. calcular o elemento $m_{i1} = - (a_{i1} / a_{11})$, $1 \leq i \leq n$.
 - 3.2. vamos somar à 2a equação a 1a , multiplicada pelo coeficiente m_{21} , e colocar o resultado na 2a linha. Isto também não altera a solução do sistema; repetir para as equações abaixo, usando m_{i1} na i -ésima equação;
 - 3.3. calcular o elemento $m_{i2} = - (a'_{i2} / a'_{22})$;
 - 3.4. repetir o passo 3.2 com as demais equações.

Método da Eliminação de Gauss

Para representar todas as mudanças vamos formar uma matriz com duas partes: a 1a será a matriz dos coeficientes e a 2a será o vetor dos termos independentes:

$$A = \left(\begin{array}{cccc|c} 3 & 2 & 1 & \vdots & 6 \\ 1 & 3 & 1 & \vdots & 5 \\ 2 & 2 & 3 & \vdots & 7 \end{array} \right)$$

(esta matriz se chama matriz aumentada).

Método da Eliminação de Gauss-Jordan

Este método é uma complementação ao método de Gauss. Ele transforma o sistema dado em um outro diagonal, isto é, onde todos os elementos fora da diagonal são nulos.

Veremos com o exemplo anterior como funciona o método de Gauss-Jordan.

Como vimos, o sistema tem a matriz aumentada:

$$A = \begin{pmatrix} 3 & 2 & 1 & : & 6 \\ 1 & 3 & 1 & : & 5 \\ 2 & 2 & 3 & : & 7 \end{pmatrix}$$

Eliminação de Gauss-Jordan X método de Gauss

A vantagem deste processo é que um sistema cuja matriz aumentada é uma matriz na forma escalonada reduzida tem solução imediata, enquanto que para resolver um sistema que está apenas na forma escalonada (triangular) ainda é necessário fazer uma série de substituições para obter a solução final.

Método da Eliminação de Gauss-Jordan

Resolver:

$$2x - y - z = 2$$

$$3x + y - 2z = 9$$

$$-x + 2y + 5z = -5$$

Solução :

$$x = 2$$

$$y = 1$$

$$z = -1$$

Método da Eliminação de Gauss-Jordan: inversão de matrizes

Uma análise rápida do método de eliminação de Gauss nos mostra que, ao invés de resolvermos um sistema completo, resolvemos, na realidade, um sistema triangular.

Isso quer dizer que, a partir do sistema dado, geramos outro, que conserva suas soluções originais.

De um modo mais formal, podemos dizer que saímos de um sistema geral

$$AX = B$$

e chegamos a outro sistema

$$A'X = B'$$

cujas matrizes A' e B' são triangulares, mas as soluções são as mesmas. Naturalmente a passagem de uma forma à outra não pode ser arbitrária. Para termos a garantia de que as raízes são preservadas só permitimos operações elementares sobre as linhas do sistema. Cada uma dessas operações, como veremos abaixo, equivale à multiplicação da matriz do sistema por uma matriz elementar. Assim, a cada transformação fica definida uma certa matriz G_i .

Método da Eliminação de Gauss-Jordan: inversão de matrizes

$$\begin{array}{ccc|ccc} 1 & -2 & 3 & 1 & 0 & 0 \\ 3 & -8 & 11 & 0 & 1 & 0 \\ -4 & 6 & -7 & 0 & 0 & 1 \end{array}$$

Pivoteamento Parcial de Gauss

É comum ocorrer que um dos pivôs ou é zero ou um valor muito próximo de zero;

Isso faz com que o multiplicador assuma um valor muito grande e quando este for multiplicado pela variável independente (vetor b) os erros da casa decimal dos valores de b vão aumentar para os dígitos mais significativos.

Por exemplo, seja um valor de $b = 5,26547$, considerando o erro como 0.00047 , se multiplicarmos, por exemplo, $5,26547$ por um valor multiplicador muito grande, como 1000 , o erro passará para os dígitos mais significativos (o erro aumenta):

$5,26547 \times 1000 = 5265,47$, o erro passa a ser igual a $0,74$!

A IDEIA É QUE O MULTIPLICADOR ASSUMA VALORES MENOR DO QUE 1, FAZENDO COM QUE OS ERROS EM b SEJAM MANTIDOS PEQUENOS!

Pivoteamento Parcial de Gauss

Considere novamente o sistema:

$$\begin{cases} E_1 : & 0.003x_1 & + & 59.14x_2 & = & 59.17, \\ E_2 : & 5.291x_1 & - & 6.13x_2 & = & 46.78 \end{cases}$$

A estratégia de pivotamento parcial define primeiro

$$\max\{|a_{11}^{(1)}|, |a_{21}^{(1)}|\} = \max\{|0.003|, |5.291|\} = |5.291| = |a_{21}^{(1)}|.$$

Pivoteamento Parcial de Gauss

A operação $(E_2 - m_{21}E_1) \rightarrow (E_2)$ reduz o sistema para

$$\begin{cases} E_1 : & 5.291x_1 & - & 6.13x_2 & = & 46.78, \\ E_2 : & & & 59.14x_2 & \approx & 59.14. \end{cases}$$

Usando quatro algarismos com arredondamento, os valores resultantes da aplicação da substituição regressiva neste sistema são os valores corretos $x_1 = 10$ e $x_2 = 1$.

Pivoteamento Parcial de Gauss

Na prática, para permutar as linhas da matriz, fazemos a seguinte multiplicação:

$$\begin{bmatrix} 0,00 & 1,00 \\ 1,00 & 0,00 \end{bmatrix}_{(2 \times 2)} \times \begin{bmatrix} 0,00 & 3,00 \\ 2,00 & 5,00 \end{bmatrix}_{(2 \times 2)} = \begin{bmatrix} 2,00 & 5,00 \\ 0,00 & 3,00 \end{bmatrix}_{(2 \times 2)}$$

Pivoteamento Parcial de Gauss: Algoritmo

Eliminação de Gauss com pivoteamento parcial: dados o número n de equações e variáveis, uma matriz aumentada $[A, b]$, com n linhas e $n + 1$ colunas, devolve um sistema linear triangular inferior equivalente ao sistema inicial ou emite uma mensagem de erro.

Passo 1: Para $i = 1, \dots, n - 1$, execute os passos 2 a 4:

Passo 2: Faça p ser o menor inteiro tal que

$|a_{pi}^{(i)}| = \max_{i \leq j \leq n} |a_{ji}^{(i)}|$, $i \leq p \leq n$. Se $a_{pi}^{(i)} = 0$, então escreva “não existe uma solução única” e pare.

Passo 3: Se $p \neq i$ então faça $(E_p) \leftrightarrow (E_i)$.

Passo 4: Para $j = i + 1, \dots, n$, execute os passos 5 e 6:

Passo 5: Faça $m_{ji} \leftarrow \frac{a_{ji}}{a_{ii}}$.

Passo 6: Faça $(E_j - m_{ji}E_i) \rightarrow (E_j)$.

Passo 7: Devolva $[A, b]$ como solução e pare.

Pivoteamento Parcial de Gauss: Algoritmo

Método de substituição regressiva: dados o número n de equações e variáveis, uma matriz aumentada $[A, b]$, com n linhas, $n + 1$ colunas e A triangular inferior, resolve o sistema linear ou emite uma mensagem dizendo que a solução do sistema linear não é única.

Passo 1: Se $a_{nn} = 0$, então escreva “não existe uma solução única” e pare.

Passo 2: Faça $x_n \leftarrow \frac{a_{n(n+1)}}{a_{nn}}$.

Passo 3: Para $i = n - 1, \dots, 1$, execute os passos 4 e 5:

Passo 4: Se $a_{ii} = 0$, então
escreva “não existe uma solução única” e pare.

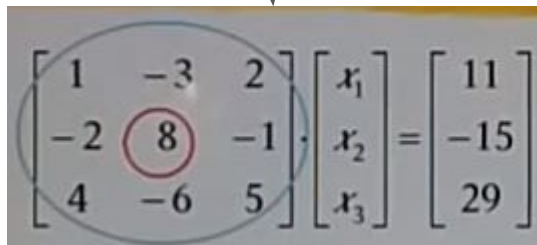
Passo 5: Faça $x_i \leftarrow \frac{a_{i(n+1)} - \sum_{j=i+1}^n a_{ij}x_j}{a_{ii}}$.

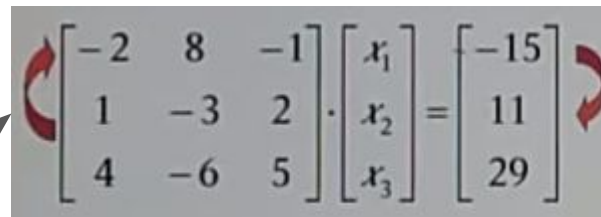
Passo 6: Devolva (x_1, x_2, \dots, x_n) como solução e pare.

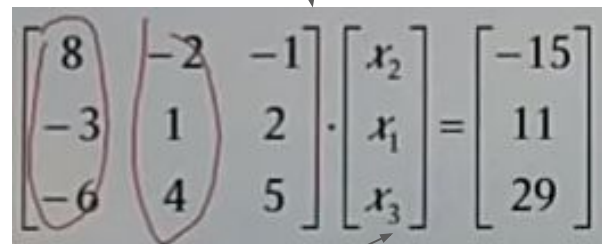
Pivoteamento Total de Gauss

Ao invés de procurar o menor elemento na coluna do pivô, procuramos o menor elemento em toda a matriz A.

$$\begin{bmatrix} 1 & -3 & 2 \\ -2 & 8 & -1 \\ 4 & -6 & 5 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 11 \\ -15 \\ 29 \end{bmatrix}$$


$$\begin{bmatrix} 1 & -3 & 2 \\ -2 & 8 & -1 \\ 4 & -6 & 5 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 11 \\ -15 \\ 29 \end{bmatrix}$$


$$\begin{bmatrix} -2 & 8 & -1 \\ 1 & -3 & 2 \\ 4 & -6 & 5 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -15 \\ 11 \\ 29 \end{bmatrix}$$


$$\begin{bmatrix} 8 & -2 & -1 \\ -3 & 1 & 2 \\ -6 & 4 & 5 \end{bmatrix} \cdot \begin{bmatrix} x_2 \\ x_1 \\ x_3 \end{bmatrix} = \begin{bmatrix} -15 \\ 11 \\ 29 \end{bmatrix}$$

Observe a troca equivalente das posições de x_n !

Pivoteamento Total de Gauss

Na prática, para chegarmos a nova matriz A' , trocamos as colunas multiplicando a matriz A por:

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$PA = A'$$

onde a posição da unidade (na matriz identidade) relacionada a cada coluna passa a assumir a posição na nova coluna.

Assim conseguimos diminuir mais ainda o multiplicador caso o maior elemento da matriz A não esteja na primeira coluna;

Fatoração LU

- É muito comum ter-se que se resolver **não** um sistema $\mathbf{A} \cdot \mathbf{x} = \mathbf{B}$, e sim muitos sistemas onde **só varia o lado direito \mathbf{B} , mantida a matriz \mathbf{A} .**
- O método apresentado, Eliminação de Gauss, obrigaria a resolver tudo desde o início, para cada novo sistema.
- Um aperfeiçoamento desse método aproveita quase tudo o que já foi feito, permitindo que a solução de cada novo sistema, onde só variou o lado direito \mathbf{B} , se dê rapidamente: é o método **LU**.
- Na verdade o vetor \mathbf{B} , em geral, representa a condição de carga da estrutura, ou do circuito elétrico, ou das condições de contorno do problema a ser resolvido.

Fatoração LU

Encontrar a matriz U através de Gauss. A matriz L é formada pelos multiplicadores que geraram U.

$Ax=b$ torna-se $LUx=b$, onde $A=LU$

Passo 1: $Ly = b$, onde determinamos os valores para y.

Passo 2: $Ux = y$, onde determinamos os valores para x.

$$\begin{matrix} & A \\ \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} & = & \begin{matrix} L:(\text{Multiplicadores}) \\ \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} \end{matrix} & \begin{matrix} U:(\text{Gauss}) \\ \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix} \end{matrix} \end{matrix}$$

Fatoração LU

Ou seja, alterando-se \mathbf{b} em $\mathbf{L}\mathbf{y} = \mathbf{b}$, encontramos facilmente \mathbf{y} .

Em seguida, calculamos facilmente \mathbf{x} em $\mathbf{U}\mathbf{x} = \mathbf{y}$.

$$\begin{array}{c} \text{A} \end{array} \qquad \begin{array}{c} \text{L: (Multiplicadores)} \end{array} \qquad \begin{array}{c} \text{U: (Gauss)} \end{array}$$
$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix}$$

Fatoração LU

Exemplo de fatoração LU:

$$5,0 \text{ } \mathbf{x}_1 + 1,0 \text{ } \mathbf{x}_2 - 2,0 \text{ } \mathbf{x}_3 = 10$$

$$3,0 \text{ } \mathbf{x}_1 - 9,4 \text{ } \mathbf{x}_2 + 1,8 \text{ } \mathbf{x}_3 = 22$$

$$1,0 \text{ } \mathbf{x}_1 + 2,2 \text{ } \mathbf{x}_2 + 4,6 \text{ } \mathbf{x}_3 = 10$$

Exercícios solicitados até o momento

Questão 1. Resolver e comparar a solução analítica e a solução numérica considerando os casos apresentados no problema do slide 26.

Questão 2. Escrever pelo menos 3 casos de cada conversão entre binário para decimal, decimal para binário, binário para decimal fracionário e decimal fracionário para binário.

Questão 3. modelar algebricamente o sistema linear compatível indeterminado do slide 88.

Questão 4. Entregar os sistemas de equações apresentados utilizando o método de Gauss, Gauss-Jordan e pivotamento parcial de Gauss (ambos com arredondamento de operações para 3 dígitos inteiros).

$$\text{a) } \begin{cases} 10^{-4}x_1 + x_2 = 1 \\ x_1 + x_2 = 2 \end{cases}$$

$$\text{b) } \begin{cases} 0,0002x_1 + 2x_2 = 5 \\ 2x_1 + 2x_2 = 6 \end{cases}$$

Resolução de Sistemas Lineares

Métodos Iterativos

Métodos Iterativos

- Para sistemas grandes, com grande porcentagem de entradas de zero (sistemas esparsos), essas técnicas aparecem como alternativas mais eficientes.
- Sistemas esparsos de grande porte frequentemente surgem na análise de circuitos, na solução numérica de problemas de valor de limite e equações diferenciais parciais.
- Em alguns casos podem ser aplicados para resolver conjuntos de equações não lineares.

Métodos Iterativos

- Um método é iterativo quando fornece uma sequência de aproximações da solução.
- Cada uma das aproximações é obtida das anteriores pela repetição do mesmo processo.
- É preciso observar se a sequência obtida está convergindo ou não para a solução desejada.

Métodos Iterativos: convergência

- Dado uma sequência de vetores $\{x^{(k)}\}$ candidatos à solução x^* do sistema linear, onde tais candidatos à soluções pertencem à uma norma E sobre um espaço vetorial, dizemos que:
 - a sequência $\{x^{(k)}\}$ **converge** para $x^* \in E$ se $\|x^{(k)} - x^*\| \rightarrow 0$, quando $k \rightarrow \infty$.

obs.: Norma: os vetores estão normalizados.

Métodos Iterativos

- Para determinar a solução de um sistema linear por métodos iterativos, precisamos transformar o sistema dado em um outro sistema onde possa ser definido um processo iterativo.
- A solução obtida para o sistema transformado deve ser também solução do sistema original (sistemas lineares devem ser equivalentes).
- Assim um sistema do tipo $Ax=b$ é transformado em $x^k = Fx^{(k-1)} + d$
Escolhemos uma aproximação inicial x^0 :
Assim, $x^1 = Fx^0 + d$
 $x^2 = Fx^1 + d$

E assim sucessivamente.

Métodos Iterativos: quando parar?

Se a sequência x^k estiver suficientemente próximo de $x^{(k-1)}$ paramos o processo.

- Dada uma precisão ε , quando $\|x(k) - x\| < \varepsilon$ então x^k é a solução do sistema linear.
- Computacionalmente, um número máximo de iterações também é critério de parada.

Gauss-Jacob

X	Y	Z	
15	2	-1	-200
2	12	1	-250
1	2	8	30

$$\begin{cases} 15x + 2y - z = 200 \\ 2x + 12y + z = -250 \\ x + 2y + 8z = 30 \end{cases} \quad \begin{pmatrix} 15 & 2 & -1 \\ 2 & 12 & 1 \\ 1 & 2 & 8 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -200 \\ -250 \\ 30 \end{pmatrix}$$

Valor inicial para as soluções.

Critério de parada:
 $\Delta x, \Delta y, \Delta z < 0.010$

onde $\Delta x, \Delta y, \Delta z$ são as diferenças entre os valores da iteração k menos a iteração k-1.

				PRECISAO		
N	X	Y	Z	ΔX	ΔY	ΔZ
0	0.000	0.000	0.000			
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						

Gauss-Jacob

X	Y	Z	
15	2	-1	-200
2	12	1	-250
1	2	8	30

Iteração 1:

$$x^1 = (b^1 - y^{L1} * Y^0 - z^{L1} * Z^0) / x^{L1}$$

$$y^1 = (b^2 - x^{L2} * X^0 - z^{L2} * Z^0) / y^{L2}$$

$$z^1 = (b^3 - x^{L3} * X^0 - y^{L3} * Y^0) / z^{L3}$$

...

$$\begin{cases} 15x + 2y - z = 200 \\ 2x + 12y + z = -250 \\ x + 2y + 8z = 30 \end{cases} \quad \begin{pmatrix} 15 & 2 & -1 \\ 2 & 12 & 1 \\ 1 & 2 & 8 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -200 \\ -250 \\ 30 \end{pmatrix}$$

				PRECISAO		
N	X	Y	Z	ΔX	ΔY	ΔZ
0	0.000	0.000	0.000			
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						

Gauss-Jacob

X	Y	Z	
15	2	-1	-200
2	12	1	-250
1	2	8	30

$$\begin{cases} 15x + 2y - z = 200 \\ 2x + 12y + z = -250 \\ x + 2y + 8z = 30 \end{cases} \quad \begin{pmatrix} 15 & 2 & -1 \\ 2 & 12 & 1 \\ 1 & 2 & 8 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -200 \\ -250 \\ 30 \end{pmatrix}$$

Iteração 1:

$$x^1 = (b^1 - y^{L1} * Y^0 - z^{L1} * Z^0) / x^{L1}$$

$$y^1 = (b^2 - x^{L2} * X^0 - z^{L2} * Z^0) / y^{L2}$$

$$z^1 = (b^3 - x^{L3} * X^0 - y^{L3} * Y^0) / z^{L3}$$

...

$$\Delta x = x^1 - x^0$$

$$\Delta y = y^1 - y^0$$

$$\Delta z = z^1 - z^0$$

				PRECISAO		
N	X	Y	Z	ΔX	ΔY	ΔZ
0	0.000	0.000	0.000			
1	-13.333	-20.833	3.750	-13.333	-20.833	3.750
2						
3						
4						
5						
6						
7						
8						
9						
10						

Gauss-Jacob

X	Y	Z	
15	2	-1	-200
2	12	1	-250
1	2	8	30

O algoritmo para após seis iterações.

Caso a condição de parada fosse que todas as variáveis Δx , Δy e Δz se igualassem à 0, o algoritmo pararia na oitava iteração.

$$\begin{cases} 15x + 2y - z = 200 \\ 2x + 12y + z = -250 \\ x + 2y + 8z = 30 \end{cases} \quad \begin{pmatrix} 15 & 2 & -1 \\ 2 & 12 & 1 \\ 1 & 2 & 8 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -200 \\ -250 \\ 30 \end{pmatrix}$$

				PRECISAO		
N	X	Y	Z	ΔX	ΔY	ΔZ
0	0.000	0.000	0.000			
1	-13.333	-20.833	3.750	-13.333	-20.833	3.750
2	-10.306	-18.924	10.625	3.028	1.910	6.875
3	-10.102	-20.001	9.769	0.204	-1.078	-0.856
4	-10.015	-19.964	10.013	0.087	0.037	0.244
5	-10.004	-19.999	9.993	0.011	-0.035	-0.020
6	-10.001	-19.999	10.000	0.003	0.000	0.007
7	-10.000	-20.000	10.000	0.001	-0.001	0.000
8	-10.000	-20.000	10.000	0.000	0.000	0.000
9	-10.000	-20.000	10.000	0.000	0.000	0.000
10	-10.000	-20.000	10.000	0.000	0.000	0.000

Gauss-Seidel

Ao invés de usar os valores iniciais de $X=0$ para calcular y^1 , o método utiliza o valor de x^1 para realizar o cálculo de y^1 .

O mesmo para calcular z^1 : utilizamos os valores de x^1 e de y^1 ao invés de usarmos os valores iniciais $X=Y=0$.

Gauss-Seidel

Se aplicarmos Gauss-Seidel no exemplo anterior (Gauss-Jacob) teremos uma economia de 2 iterações. Em alguns casos, onde Gauss-Jacob demandaria muitas iterações ,poderia economizar uma quantidade muito maior de iterações .

				PRECISAO		
N	X	Y	Z	ΔX	ΔY	ΔZ
0	0.000	0.000	0.000			
1	-13.333	-18.611	10.069	-13.333	-18.611	10.069
2	-10.181	-19.976	10.016	3.153	-1.365	-0.053
3	-10.002	-20.001	10.001	0.178	-0.025	-0.016
4	-10.000	-20.000	10.000	0.002	0.001	-0.001
5	-10.000	-20.000	10.000	0.000	0.000	0.000
6	-10.000	-20.000	10.000	0.000	0.000	0.000
7	-10.000	-20.000	10.000	0.000	0.000	0.000

Exercícios solicitados até o momento

Questão 5. Implementar em uma linguagem de programação à sua escolha os métodos de Gauss, Gauss-Jordan, Gauss-Jacobi e Gauss-Seidel. Comparar a quantidade de iteração entre esses algoritmos para o caso a seguir, considerando apenas Gauss-Jacobi e Gauss-Seidel:

$$\begin{bmatrix} 1 & 2 & -1 & 0 \\ 0 & 1 & 3 & -1 \\ -1 & 0 & 1 & 4 \\ 5 & -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ w \end{bmatrix} = \begin{bmatrix} 8 \\ 3 \\ -20 \\ 9 \end{bmatrix}$$

Obs.: Aplicar o método de pivoteamento total caso seja necessário.