# 1 Erros

## 1.1 Introdução

De uma maneira geral, a resolução de um problema em qualquer área do conhecimento científico passa inicialmente por uma fase de observação e entendimento do fenômeno físico envolvido. Com conhecimentos já estabelecidos, buscamos, usando simplificações, quando necessárias, a construção de um modelo matemático que represente o problema que desejamos tratar com a maior fidelidade possível. Esta etapa é caracterizada como **fase de modelagem** do problema real.

Com o problema representado por um modelo matemático, buscamos, para a sua resolução, um método exato quando possível, ou, quando não, um método numérico aproximado.

Mesmo quando utilizamos na resolução do modelo matemático um método exato, isto é, um método que apresenta a solução exata, pelo fato de este envolver um número muito grande de operações elementares como adição, multiplicação, subtração e divisão e sendo estas processadas em equipamento com capacidade limitada para armazenar dados, podemos cometer erros.

Por outro lado, quando optamos, em razão da complexidade do modelo matemático, pela resolução usando um método numérico, além dos erros no processamento anteriormente mencionados, podemos também cometer erros provenientes do fato de utilizarmos um algoritmo aproximado. Esta etapa é caracterizada como **fase de resolução** do modelo matemático.

Podemos entender as duas fases descritas anteriormente utilizando o esquema representado na Figura 1.1.

Neste capítulo apresentamos os principais erros que podem ocorrer na fase da resolução de um problema. Os erros cometidos devido à mudança da base de representação de números devido ao sistema utilizado pelos computadores para armazenar dados numéricos; ainda os erros de arredondamento e truncamento, erros absolutos e relativos e, finalmente, a propagação dos erros.

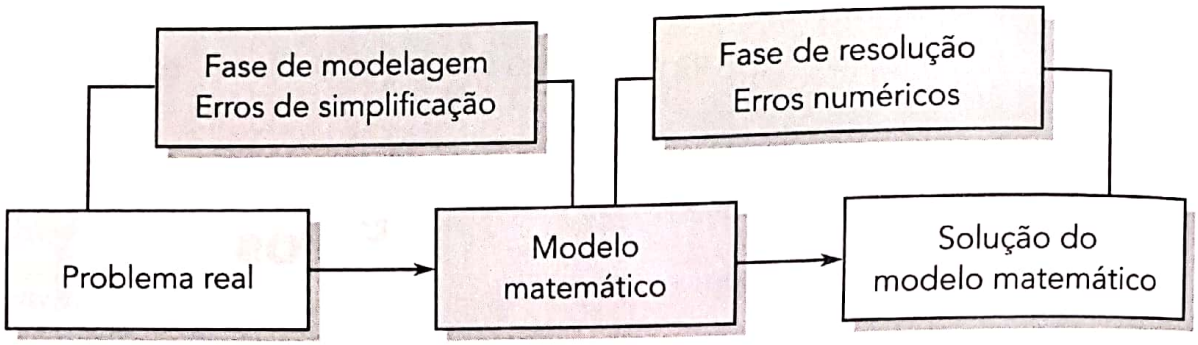



**Figura 1.1**

## 1.2 Erros na fase da modelagem

São os erros decorrentes de simplificações, muitas vezes necessárias, para que o fenômeno da natureza que estivermos observando possa ser representado por um modelo matemático e que tenha condições de ser tratado com as ferramentas matemáticas disponíveis.

## 1.3 Erros na fase de resolução

São erros provenientes da utilização de um computador, para processarmos os cálculos necessários na obtenção de uma solução para o modelo matemático. Tais erros ocorrem devido ao fato de os equipamentos terem capacidade limitada para armazenar dados numéricos utilizados nas operações elementares de adição, multiplicação, subtração e divisão.

Os erros nesta fase de resolução podem ser classificados em erros na mudança de base e erros de representação de números, apresentados a seguir:

### 1.3.1 Erros de mudança da base

A maioria dos equipamentos computacionais representa os valores numéricos no sistema binário. Assim, quando os dados numéricos presentes nos modelos matemáticos são lidos, estes são transformados em uma outra base de representação.

Acontece, muitas vezes, que esta transformação pode ser acometida de erros, em razão da limitação da representação do equipamento computacional que estamos utilizando para o processamento dos dados numéricos.

Dado um número real, $N$, é sempre possível representá-lo em qualquer base $b$, da seguinte forma:

$$N_b = \sum_{i=n}^{m} a_i \times b^i$$

em que $a_i \in \{0, 1, 2, 3, \ldots, (b-1)\}$, com $n$ e $m$ inteiros.

### Base binária

$$N_2 = \sum_{i=n}^{m} a_i \times 2^i, \quad a_i \in \{0,1\}$$

**Exemplo 1.1**

a) $(1011)_2 = 1 \times 2^0 + 1 \times 2^1 + 0 \times 2^2 + 1 \times 2^3$

Neste caso, o binário só tem a parte inteira, isto é, $i = 0,1,2,3$ e temos:

$$a_0 = 1, \quad a_1 = 1, \quad a_2 = 0, \quad a_3 = 1$$

b) $(111.01)_2 = 1 \times 2^{-2} + 0 \times 2^{-1} + 1 \times 2^0 + 1 \times 2^1 + 1 \times 2^2$

Neste caso, o binário tem parte inteira e parte fracionária, isto é, $n = -2$ e $m = 2$ e, portanto:

$$a_{-2} = 1, \quad a_{-1} = 0, \quad a_0 = 1, \quad a_1 = 1, \quad a_2 = 1$$

### Base decimal

$$N_{10} = \sum_{i=n}^{m} a_i \times 10^i, \quad a_i \in \{0,1,\ldots,9\}, \text{ com } n \text{ e } m \text{ inteiros.}$$

**Exemplo 1.2**

a) $(231)_{10} = 1 \times 10^0 + 3 \times 10^1 + 2 \times 10^2$

Neste caso, o número na base decimal é inteiro, $i = 0, 1, 2$ e temos:

$$a_0 = 1, \quad a_1 = 3, \quad a_2 = 2$$

b) $(231.35)_{10} = 5 \times 10^{-2} + 3 \times 10^{-1} + 1 \times 10^0 + 3 \times 10^1 + 2 \times 10^2$

Neste caso, o número na base decimal tem parte inteira e parte fracionária, $n = -2$ e $m = 2$ e temos:

$$a_{-2} = 5, \quad a_{-1} = 3, \quad a_0 = 1, \quad a_1 = 3, \quad a_2 = 2$$

Assim, dado um número real qualquer numa base $b$, podemos escrevê-lo em uma outra base $b'$, a partir de adequação conveniente de seus coeficientes,

$$a_i = 0,1,2,3,\ldots,(b-1)$$

e de uma potência adequada na nova base $b'$.

### *Mudança da base binária para a base decimal*

***Procedimento:*** multiplicar o dígito binário por uma potência adequada de 2.

**Exemplo 1.3**

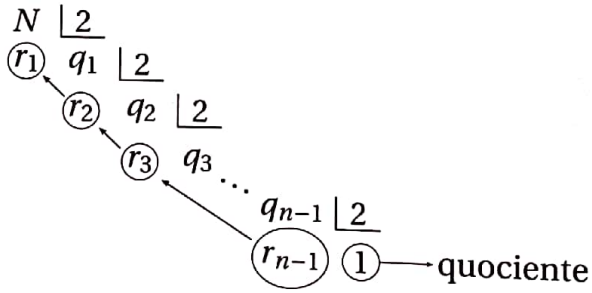
a) $(1101)_2 = 1\times 2^0 + 0\times 2^1 + 1\times 2^2 + 1\times 2^3 = (13)_{10}$

b) $(111.011)_2 = 1\times 2^{-3} + 1\times 2^{-2} + 0\times 2^{-1} + 1\times 2^0 + 1\times 2^{-1} + 1\times 2^2 = (7.375)_{10}$

### *Mudança da base decimal para a base binária (número na base decimal tem somente a parte inteira)*

***Procedimento:*** divisões sucessivas.

O procedimento consiste na divisão do número $N$ na base decimal sucessivamente por 2, armazenando, a cada passo, os restos, $r_i$ $i = n-1, n-2, \ldots, 1$ até que o quociente da divisão seja igual a 1. O número binário é constituído do quociente 1 e pelos restos das divisões $r_{n-1}, r_{n-2}, \ldots, r_1$ nesta ordem como segue:

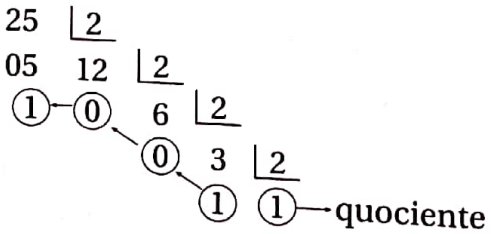



Assim, temos:

$$N_{10} = (1r_{n-1}r_{n-2}\ldots r_2 r_1)_2$$

**Exemplo 1.4**

a) O número 25 na base 10 para a base 2,



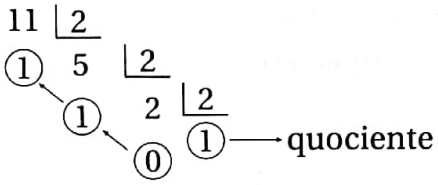

Assim,

$$(25)_{10} = (11001)_2 = 1\times 2^0 + 0\times 2^1 + 0\times 2^2 + 1\times 2^3 + 1\times 2^4$$

b) O número 11 na base 10 para a base 2,

$$(11)_{10} = (1011)_2 = 1 \times 2^0 + 1 \times 2^1 + 0 \times 2^2 + 1 \times 2^3$$

pois




***Mudança da base decimal para a base binária***
***(número na base decimal tem somente a parte fracionária)***

*Procedimento:* multiplicações sucessivas.

O procedimento é constituído dos seguintes passos:

a) Multiplicamos o número fracionário por 2.

b) Do resultado do passo a), a parte inteira é o primeiro dígito binário.

c) Do resultado do passo b), a parte fracionária é novamente multiplicada por 2.

d) O processo continua até que a parte fracionária seja nula.

*Exemplo 1.5*

a) $(0.1875)_{10} = (0.0011)_2 = 0 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3} + 1 \times 2^{-4} = (3/16)_{10}$, isto é:

$(0.1875) \times (2) = 0.375$ → parte inteira = 0 e parte fracionária = 0.375

$(0.375) \times (2) = 0.75$ → parte inteira = 0 e parte fracionária = 0.75

$(0.75) \times (2) = 1.5$ → parte inteira = 1 e parte fracionária = 0.5

$(0.5) \times (2) = 1.0$ → parte inteira = 1 e parte fracionária = 0

b) $(13.25)_{10} = (13)_{10} + (0.25)_{10} = (1101)_2 + (0.01)_2 = (1101.01)_2$

c) $(0.2)_{10} = (0.001100110011...)_2$

Observe que $(0.2)_{10}$ é uma dízima periódica de período (0.0011). Assim, o decimal $(0.2)_{10}$ não tem uma representação binária exata, isto é, a representação é aproximada e, portanto, apresenta erro.

### 1.3.4 Erro absoluto

Definimos erro absoluto por

$$E_{abs} = |a_{ex} - a_{aprox}|$$

em que $a_{ex}$ é o valor exato da grandeza considerada e $a_{aprox}$ é o valor aproximado da mesma grandeza numérica.

### 1.3.5 Erro relativo

Definimos erro relativo por

$$E_{rel} = \frac{|a_{ex} - a_{aprox}|}{|a_{ex}|}$$

em que $a_{ex}$ é o valor exato da grandeza considerada e $a_{aprox}$ é o valor aproximado da mesma grandeza.

Podemos observar que o erro relativo nos fornece mais informações sobre a qualidade do erro que estamos cometendo num determinado cálculo, uma vez que no erro absoluto não é levada em consideração a ordem de grandeza do valor calculado, enquanto no erro relativo esta ordem é contemplada.

*Exemplo 1.10*

a) Consideremos o valor exato

$$a_{ex} = 2345.713$$

e o valor aproximado

$$a_{aprox} = 2345.000.$$

Então,

$$E_{abs} = |a_{ex} - a_{aprox}| = |2345.713 - 2345.000| = 0.713$$

$$E_{rel} = \frac{|a_{ex} - a_{aprox}|}{|a_{ex}|} = \frac{0.713}{2345.713} = 0.00030396$$

b) Consideremos o valor exato $a_{ex} = 1.713$ e o valor aproximado $a_{aprox} = 1.000$.
Então,

$$E_{abs} = 0.713$$

$$E_{rel} = 0.416229$$

Observe que nos exemplos **a)** e **b)** o erro absoluto é o mesmo, embora o erro cometido pela aproximação seja muito mais significativo no exemplo **b)**. No exemplo **a)**, o erro relativo é da ordem de 0.03%, e no exemplo **b)**, é da ordem de 41.6%.

Como usar as definições de $E_{abs}$ e $E_{rel}$ na resolução de problemas se quando aplicamos um procedimentos numérico obtemos uma solução aproximada $a_{aprox}$, e não temos o valor exato $a_{ex}$?

Em geral, nos procedimentos numéricos geramos uma sequência de soluções aproximadas $x_0, x_1, \ldots, x_k, x_{k+1} \ldots$ que converge ou não para a solução desejada $\bar{x}$ do problema.

Quando esta sequência for convergente para a solução $\bar{x}$, na medida em que $k$ cresce $(k \to \infty)$ os valores de $x_k$ e $x_{k+1}$ tendem a ficar próximos, isto é,

$$|x_{k+1} - x_k| \simeq 0 \quad \text{ou} \quad \frac{|x_{k+1} - x_k|}{|x_{k+1}|} \simeq 0.$$

As definições de erro absoluto e erro relativo são usadas como critério de parada nessas sequências de aproximações, com uma tolerância $\varepsilon > 0$ pré-fixada.

Em geral, o erro relativo é preferível, devido às observações nos exemplos anteriores.

### *Exemplo 1.11*

Para resolver a equação $f(x) = x^2 - a = 0$, com $a > 0$, podemos utilizar o seguinte processo iterativo:

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right), \quad n = 0, 1, 2, \ldots$$

Assim, dado o valor $x_0$, usamos a expressão anterior para gerar a sequência de soluções aproximadas $x_1, x_2, \ldots$

Dado que a propriedade de convergência da sequência de aproximações esteja estabelecida e uma tolerância pré-fixada $\varepsilon > 0$ foi definida para o cálculo de uma raiz da equação $\boldsymbol{f(x) = 0}$, podemos verificar se a sequência de aproximações atingiu a precisão anterior $\varepsilon$, realizando o seguinte teste:

Se $|x_{n+1} - x_n| < \varepsilon$ for verdadeiro, dizemos que $x_{n+1}$ é a raiz da equação $f(x) = 0$ com tolerância $\varepsilon$; caso contrário, devemos calcular outro elemento da sequência.

Podemos ainda de forma alternativa realizar o seguinte teste: Se

$$\frac{|x_{n+1} - x_n|}{|x_{n+1}|} < \varepsilon$$

for verdadeiro, concluímos que $x_{n+1}$ é a raiz da equação com a tolerância $\varepsilon > 0$ e, em caso contrário, devemos proceder ao cálculo de outro termo da sequência. No primeiro teste, usamos $E_{abs}$ e, no segundo, $E_{rel}$.

### *Exemplo 1.12*

No Exemplo 1.11, considere $a = 2$ e temos neste caso a equação $x^2 - 2 = 0$ para ser resolvida. Utilizando o processo iterativo:

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right), \quad n = 0, 1, 2\ldots$$

e uma tolerância $\varepsilon = 0.0001$, temos, a partir de uma solução $x_0 = 1$, a seguinte sequência de soluções aproximadas convergindo para a solução

$$\bar{x} = \sqrt{2} \simeq 1.414213562:$$

$x_0 = 1 \rightarrow$ solução inicial

$$x_1 = \frac{1}{2}\left(x_0 + \frac{2}{x_0}\right) = 1.50000, \quad \frac{|x_1 - x_0|}{|x_1|} = 0.33333 > \varepsilon$$

$$x_2 = \frac{1}{2}\left(x_1 + \frac{2}{x_1}\right) = 1.41667, \quad \frac{|x_2 - x_1|}{|x_2|} = 0.05882 > \varepsilon$$

$$x_3 = \frac{1}{2}\left(x_2 + \frac{2}{x_2}\right) = 1.41422, \quad \frac{|x_3 - x_2|}{|x_3|} = 0.00173 > \varepsilon$$

$$x_4 = \frac{1}{2}\left(x_3 + \frac{2}{x_3}\right) = 1.41421, \quad \frac{|x_4 - x_3|}{|x_4|} = 0.00001 < \varepsilon$$

Assim, como o critério de parada está satisfeito, temos que $x_4 = 1.41421 \simeq \sqrt{2}$.

## 1.3.6 Erro de truncamento

Quando representamos uma função utilizando uma série infinita e, por limitações do sistema de armazenamento de dados do equipamento, considerarmos apenas um número finito de termos na série, dizemos que estamos cometendo um erro de truncamento.

### *Exemplo 1.13*

a) Consideramos a representação de uma função $f(x)$ utilizando a Série de Taylor, nas vizinhanças do ponto $\bar{x}$:

$$f(x) = f(\bar{x}) + f^{(1)}(\bar{x})\frac{(x-\bar{x})}{1!} + f^{(2)}(\bar{x})\frac{(x-\bar{x})^2}{2!} + \cdots + f^{(n)}(\bar{x})\frac{(x-\bar{x})^n}{n!} + \cdots$$

em que $f^{(n)}(\bar{x})$ é o valor da $n$-ésima derivada da função $f(x)$ no ponto $\bar{x}$. Quando truncamos a série no 3º termo, isto é, considerando apenas os termos até a derivada de ordem 2, na expressão anterior, temos um erro cometido nesta aproximação, como segue:

$$f(x) \simeq f(\bar{x}) + f^{(1)}(\bar{x})\frac{(x-\bar{x})}{1!} + f^{(2)}(\bar{x})\frac{(x-\bar{x})^2}{2!}$$

b) Consideremos o desenvolvimento da função $f(x) = e^x$ em Série de Taylor, em torno do ponto $x = 0$, isto é:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots$$

ou, de forma compacta:

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

Suponha que o equipamento utilizado para trabalhar numericamente com a série seja capaz de armazenar somente dados referentes aos 4 primeiros termos da série, isto é:

$$e^x \simeq 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!}$$

Neste caso, desprezamos todos os termos de potências maiores que 4, isto é, truncamos a série no termo de potência de ordem 3.

Destacando os quatro primeiros termos da série, podemos escrevê-la da seguinte maneira:

$$e^x = \frac{1}{6}(x^3 + 3x^2 + 6^x + 6) + \sum_{n=4}^{\infty} \frac{x^n}{n!}$$

Vamos supor que desejamos calcular o valor de $e^x$ para $x = 2$ usando apenas os quatro primeiros termos da série, isto é, a série truncada.

Neste caso, temos $e^2 = 6.33333$, que é um valor com erro absoluto bem significativo quando comparado com o valor $e^2 = 7.38906$ obtido numa calculadora científica que armazena uma quantidade maior de termos da série.

## 1.3.7 Propagação dos erros

Quando desenvolvemos ou utilizamos um procedimento numérico para buscar a solução de um determinado problema, normalmente o processamento envolve um número muito grande de operações elementares. Assim, na maioria das vezes, o erro cometido em uma operação isolada pode não ser muito significativo para a solução do

problema que estamos tratando, mas, sim, é necessário estar ciente como os erros se propagam quando tratamos com muitas operações no processamento.

O crescimento subsequente de um erro gerado é chamado de erro propagado.

Neste caso, é fundamental termos o conhecimento de como estes erros estão se propagando, isto é, caso estejam se acumulando a uma taxa crescente, dizemos que o erro é ilimitado, e a sequência de operações é considerada instável.

Se, por outro lado, os erros estão se acumulando a uma taxa decrescente, dizemos que o erro é limitado e, portanto, a sequência de operações é considerada estável.

Ilustramos a influência do erro de arredondamento na sequência de operações, com os seguintes exemplos:

*Exemplo 1.14*

Considere a soma

$$\sum_{k=1}^{100} 0.1 + 10,$$

base decimal, 2 dígitos e arredondamento.

a) $S = \underbrace{(0.1 + 0.1 + \ldots + 0.1)}_{100 \text{ termos}} + 10 = 10 + 10 = 20$

Neste caso, somamos inicialmente os números pequenos (100 termos), que, somados com o número 10, não são desprezíveis, produzindo o resultado igual ao número 20.

b) Se fizermos

$$S = 10 + \sum_{k=1}^{100} 0.1$$

e usando arredondamento com 2 dígitos em todas as operações, temos:

$$S = 10 + \underbrace{0.1 + 0.1 + \ldots + 0.1}_{100 \text{ termos}}$$

$$= \underbrace{(10 + 0.1)}_{10} + \underbrace{0.1 + 0.1 + \ldots + 0.1}_{99 \text{ termos}}$$

$$= \underbrace{(10 + 0.1)}_{10} + \underbrace{0.1 + 0.1 + \ldots + 0.1}_{98 \text{ termos}}$$

$$\vdots$$

$$= 10$$

Neste caso, somamos em cada parcela o número 10 com o número 0.1, que no resultado da adição com arredondamento tornou-se desprezível.

Assim, obtemos resultados muito diferentes para a soma S. Observe que, a sequência de como executamos as operações com 2 dígitos, os erros de arredondamento são introduzidos de formas diferentes, influenciando o resultado final.

A propagação (acúmulo) dos erros, neste caso, decorre da adição de um número grande com números pequenos.

Concluímos que, na adição envolvendo muitas parcelas, devemos somar primeiro os números de grandeza numérica pequena.

### *Exemplo 1.15*

Usando aritmética num sistema de ponto flutuante de 4 dígitos, base decimal e arredondamento, calcule o valor da seguinte soma:

$$S = \sum_{i=1}^{4} (x_i + y_i), \text{ sendo } x_i = 0.46709 \text{ e } y_i = 3.5678$$

Para $i = 1$, na aritmética definida, realizamos inicialmente a operação que resulta no seguinte valor aproximado:

$$S_1 = (x_1 + y_1) = 0.4035 \times 10^1$$

Considerando o valor exato com todos dígitos $a_{ex} = 4.03489$ e $a_{aprox} = 4.035$, temos o erro absoluto:

$$E_{abs1} = |4.03489 - 4.035| = 0.00011 = 0.11 \times 10^{-3}$$

Para $i = 2$, realizamos a operação que resulta no seguinte valor aproximado:

$$S_2 = (x_1 + y_1) + (x_2 + y_2) = 0.8070 \times 10^1$$

cujo erro absoluto é dado por:

$$E_{abs2} = |8.06978 - 8.070| = 0.00022 = 0.22 \times 10^{-3}$$

Observe que, ao realizarmos a mesma operação de adição por duas vezes, cometemos um erro absoluto maior.

Para $i = 3$, realizamos a operação que resulta no seguinte:

$$S_3 = (x_1 + y_1) + (x_2 + y_2) + (x_3 + y_3) = 0.1211 \times 10^2$$

cujo erro absoluto é dado por:

$$E_{abs3} = |12.10467 - 12.11| = 0.00533 = 0.533 \times 10^{-2}$$

Para $i = 4$, repetindo o mesmo procedimento, obtemos o seguinte valor para a soma:

$$S_4 = 0.1614 \times 10^2$$

que apresenta o seguinte erro absoluto:

$$E_{abs4} = |16.13956 - 16.14| = 0.00044 = 0.44 \times 10^{-3}$$

Para $i = 5$, temos o seguinte valor para a soma:

$$S_5 = 0.2018 \times 10^2$$

e o erro absoluto:

$$E_{abs5} = |20.17445 - 20.18| = 0.005550 = 0.5550 \times 10^{-2}$$

Para $i = 6$, temos o seguinte valor para a soma:

$$S_6 = 0.2421 \times 10^2$$

e o erro absoluto:

$$E_{abs6} = |24.20934 - 24.21| = 0.000660 = 0.6600 \times 10^{-3}$$

Para $i = 7$, temos o seguinte valor para a soma:

$$S_7 = 0.2825 \times 10^2$$

e o erro absoluto:

$$E_{abs7} = |28.24423 - 28.25| = 0.005770 = 0.5770 \times 10^{-2}$$

Observe que, neste caso, os erros aumentaram nas primeiras somas das parcelas, mas depois oscilaram, diminuindo e aumentando. De uma maneira geral, os erros podem se acumular bem como se cancelar durante o procedimento numérico.

### *Exemplo 1.16*

Para resolver a equação $f(x) = x^2 - 2 = 0$, do Exemplo 1.12, utilizamos o processo iterativo:

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right), \quad n = 0, 1, 2, \ldots$$

Neste procedimento, em cada iteração estão envolvidas as operações de adição, multiplicação e divisão, que são repetidas até que se calcule o valor aproximado $x_n$ para solução da equação com uma precisão $\varepsilon$ desejada.

Desta forma, observamos que o valor final $x_n$ está sujeito a um determinado tipo de erro, e a cada iteração realizada este erro pode se propagar ao longo do processo. Como o procedimento é convergente para a solução $\bar{x} = \sqrt{2}$ da equação, apesar dos erros cometidos, temos que a sequência de operações tornou-se estável, pois este processo de refinamento, apesar de muitas operações, tende a corrigir os erros.

## Exercícios

**1.** Representar na base binária os seguintes números decimais:

a) 13

b) 29.75

c) 17.6

d) 0.46875

**2.** Considere os números $x = 12.4$, $y = 4.18$ e $z = 6.05$, base decimal, 3 dígitos e arredondamento em todas as operações. Calcule:

a) $(x + y) + z$; $x + (y + z)$

b) $(x - y) - z$; $x - (y + z)$

c) $\left[\frac{x}{y}\right] \times z$; $\left[\frac{x \times z}{y}\right]$

d) $x \times (y + z)$; $(x \times y) + (x \times z)$

**3.** Represente o número decimal (0.2) na base binária com 4, 8, 12 e 16 dígitos.

**4.** Considerando que a base 16 é representada pelos dígitos 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F, represente:

a) $(27D)_{16}$ na base decimal

b) $(27D.9)_{16}$ na base decimal

c) $(32E.32)_{16}$ na base decimal

**5.** Representar os seguintes números na forma normalizada:

a) $(100)_{10}$