Computação com Ponto Flutuante IEEE

Graduação em Engenharia Elétrica

Erivelton Geraldo Nepomuceno

Departamento de Engenharia Elétrica
Universidade Federal de São João del-Rei

Agosto de 2014.

Plano de Ensino

Ementa

1. Introdução.
2. Os números reais
3. Representação de Números no Computador.
4. Padrão IEEE de ponto flutuante.
5. Arredondamento
6. Arredondamento correto de operações com ponto flutuante.
7. Exceções
8. Os microprocessadores da Intel.
9. Linguagens de Programação
10. Ponto Flutuante em C
11. Cancelamento.
12. Condicionamento de problemas.
13. Estabilidade de algoritmos.
14. Introdução à Computação por Intervalos.
15. Introdução à Precisão Arbitrária.

Referências

- Overton, M. L. (2001), *Numerical Computing with IEEE floating point arithmetic*, SIAM.
- Institute of Electrical and Electronic Engineering (2008), 754-2008 – *IEEE standard for floating-point arithmetic*.
- Rudin, W. (1976), *Principles of mathematical analysis*, McGraw-Hill New York.
- Moore, R. E. (1979), *Methods and Applications of Interval Analysis*, Philadelphia: SIAM.
- Muller, J.-M.; Brisebarre, N.; De Dinechin, F.; Jeannerod, C.-P.; Lefevre, V.; Melquiond, G.; Revol, N.; Stehlé, D.; Torres, S. & others (2010), *Handbook of floating-point arithmetic*, Springer.
- Goldberg, D. (1991), What Every Computer Scientist Should Know About Floating-point Arithmetic, *Computing Surveys* 23(1), 5–48.

Avaliação

- N1 = 30 pontos: Prova 1: Unidades de 1 a 7.
  - Data: 16/09/2014.
- N2 = 30 pontos: Prova 1: Unidades de 8 a 15.
  - Data: 02/12/2014
- N3 = 40 pontos: Trabalhos computacionais e exercícios em sala de aula.
  - Data: ao longo do curso.
- N4 = 60 pontos. Exame especial: Unidades 1 a 15.
  - Data: 11/12/2014.
- Nota Final: NF=N1+N2+N3.
  - **Se** N1 $\geq$ 18, N2 $\geq$ 18 e NF $\geq$ 60 **então**
    - Aprovado.
  - **senão se** N4 $\geq$ 24 e (N4+N3) $\geq$ 60 **então**
    - Aprovado.
  - **senão**
    - Reprovado.

- Computação numérica é uma parte vital da infraestrutura tecnológica e científica da atualidade.
- Praticamente toda computação numérica utiliza aritimética de ponto flutuante.
- Quase todos os computadores fazem uso da norma IEEE para aritmética de ponto flutuante.
- Entretanto, percebe-se que aspectos importantes da norma IEEE ainda não são compreendidos por vários estudantes e profissionais.
- Computação numérica significa computação com números.
- É uma área tão antiga quanto a própria civilização humana.
- Em torno de 1650, os egipícios já empregava técnicas de computação.
- Contar pedras e gravetos foi utilizado há anos para contar e armazenar números.
- O ábaco foi utilizado na Europa até a introdução da notação posicional decimal.

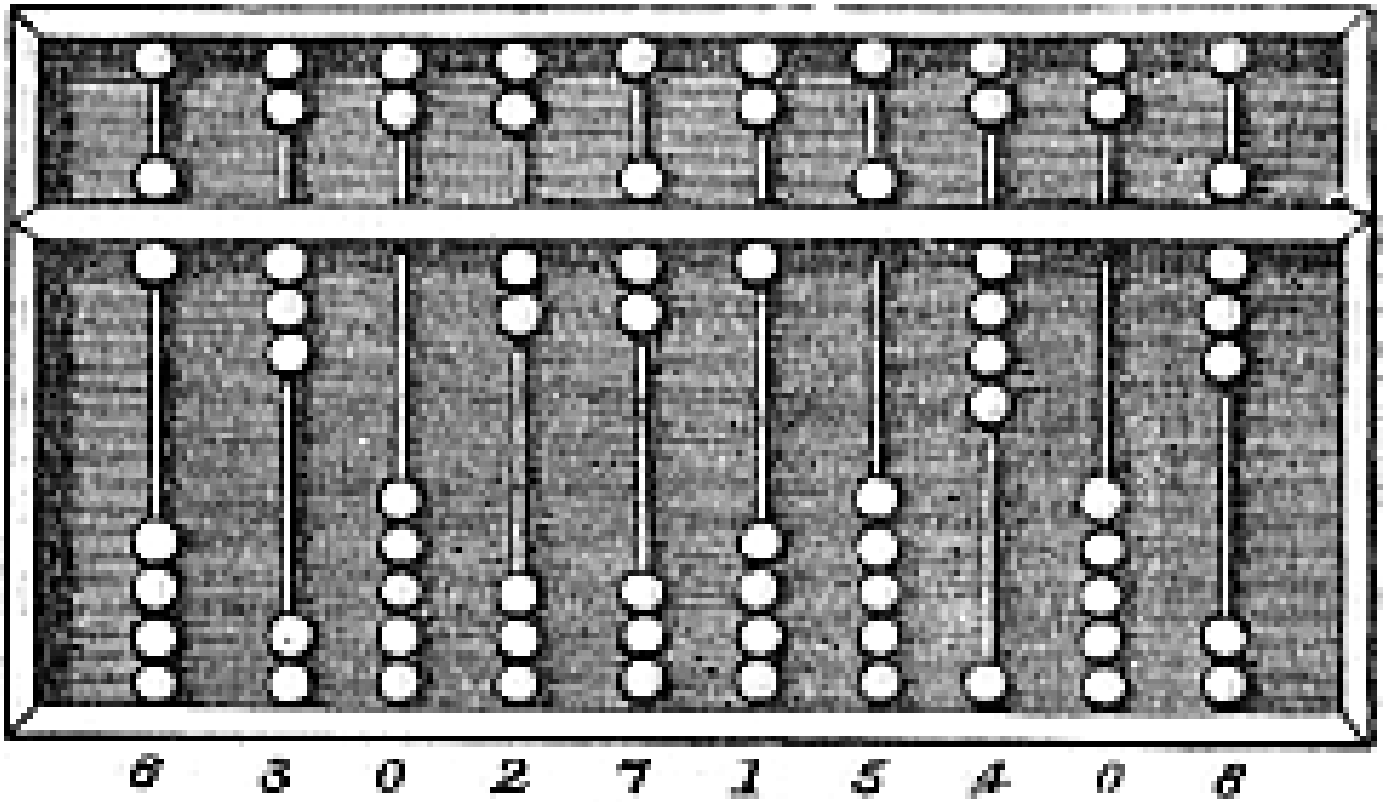

Figure 1: Primeira calculadora utilizada pelo homem: um ábaco representando o número 6302715408. Fonte: Wikipedia.

- A partir do séc. XVI, o sistema decimal se tornou base em toda a Europa.
- O próximo grande avanço foi a tabulação de logaritmos por John Napier no início do séc. XVII.
- Com logaritmos é possível substituir divisões e multiplicações por subtrações e adições.
- Isaac Newton e Leibniz desenvolveram o cálculo no séc. XVII e técnicas numéricas para a solução de vários problemas.
- Outros grandes matemáticos, tais como Euler, Lagrange, Gauss foram responsáveis por grandes desenvolvimentos na computação numérica.
- Um outro dispositivo utilizado foi a régua de deslizamento até a década de 70 do século passado.
- Dispositivos mecânicos para cálculo foram inventados por Schickard, Pascal e Leibniz.
- Charles Babbage iniciou o desenvolvimento de equipamentos sem intervenção humana.

- Durante Segunda Guerra Mundial houve um grande desenvolvimento de dispositivos para cálculos, e pode-se afirmar que mais ou menos nessa época começou-se a era da computação.
- Uma das primeiras máquinas consideradas como computador foi o Z3, construído pelo engenheiro Konrad Zuse na Alemanha entre os anos de 1939 e 1941. O Z3 usava dispositivos eletromecânicos e já empregava números binários de ponto flutuante.
- O governo britânico desenvolveu nessa mesma época um dispositivo eletrônico chamado Colossus usado para decodificar mensagens secretas.

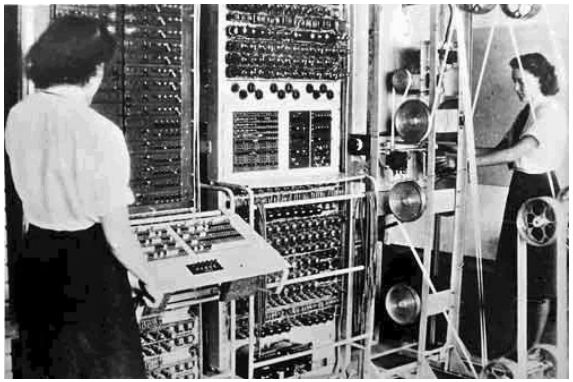

Figure 2: Colossus Mark 2 desenvolvido em 1944. Fonte: Wikipedia.

- Considera-se como primeiro computador eletrônico o ENIAC (*Electronic Numerical Integrator And Computer*). É um dispositivo de cerca de 18000 válvulas e foi construído na Universidade da Pensilvânia entre os anos de 1943 e 1945 por Eckert e Mauchly.

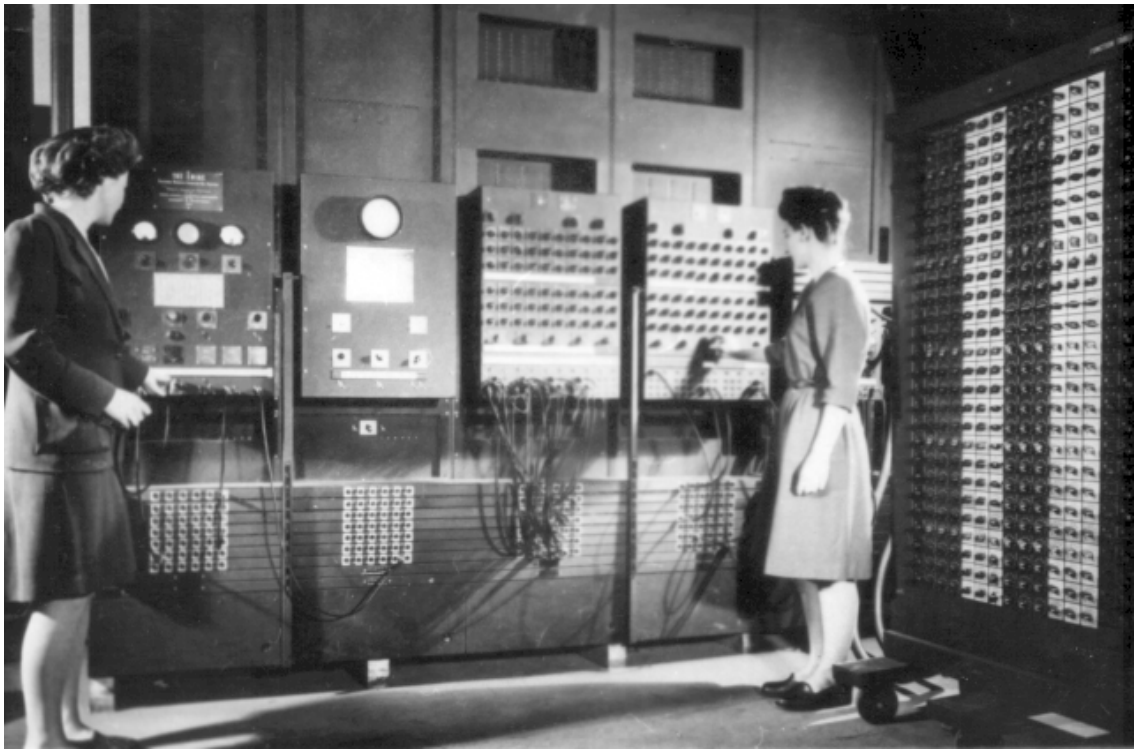

Figure 3: Eniac desenvolvido em 1946. Fonte: Wikipedia.

- Os dois principais cientistas que influenciaram os padrões de desenvolvimento dos dispositivos computacionais foram Alan Turin e John von Neumann.

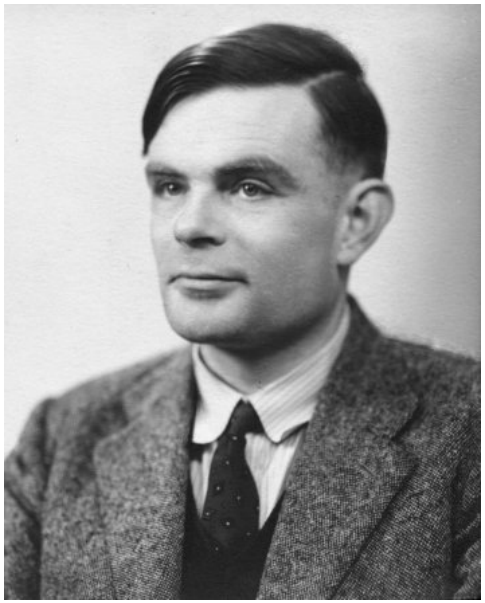

Figure 4: Alan Turin.
Fonte: Wikipedia.

Figure 5: John von Newmann.
Fonte: Wikipedia.

- Durante a década de 1950, o principal uso dos computadores foi para computação numérica.
- A partir de 1960, os computadores passaram a ser usados também em grandes empresas para processar informação, tais como, texto e imagem.
- Usuários frequentemente não estão cientes de que a manipulação de texto, som ou imagem envolve computação numérica.
- Os computadores são usados para resolver equações que modelam os mais diferentes sistemas: da expansão do universo à micro-estrutura do átomom; processamento de imagens e análise estatística de dados médicos; predição de clima; simulação de circuitos para projetos de computadores menores e mais rápidos; modelagem de aeronaves para testes e treinamento de pilotos; confiabilidade de sistemas elétricos.
- Os resultados numéricos são comparados com os resultados experimentais.
- Em síntese: todas áreas da ciência e engenharia utilizam fortemente a computação numérica.

- Os números reais são representados por uma linha.

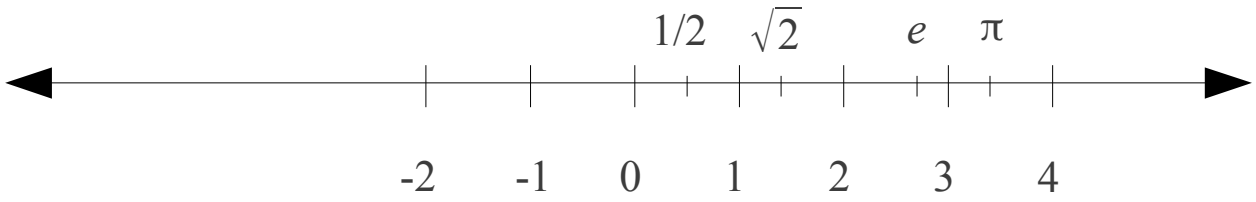

Figure 6: A linha que representa os números reais.

- A linha se extende infinitamente em direção a $-\infty$ e $+\infty$.
- Os símbolos $-\infty$ e $+\infty$ não são considerados como números.

Definition 1

O sistema de números real extendido consite do campo real R e dois símbolos: $+\infty$ e $-\infty$. Preserva-se a ordem original em R, e define-se

$$+\infty < x < -\infty$$

para qualquer $x \in R$. Utiliza-se o símbolo $\bar{R}$.

- Há infinitos mas contáveis números inteiros $0, 1, -1, 2, -2, 3, -3, \ldots$.
- Os números racionais (Q) são aqueles que consistem da razão de dois inteiros, tais como: $1/2, 2/3, 6/3$.
- Os números racionais são infinitos mas contáveis.

Exercício 1

Mostre que os números racionais são contáveis. Dica: utilize uma tabela e faça uso da diagonal.

- Os números irracionais são os números reais que não são racionais. Exemplos: $\sqrt{2}, \pi, e$.

Exemplo 1

O número e é o limite de

$$\left(1 + \frac{1}{n}\right)^n$$

quando $n \to \infty$.

- As investigações para a definição de e começaram no séc. XVII.
- Todo número irracional pode ser definido como o limite de uma sequência de números racionais.
- O conjunto de números irracionais é dito ser incontável.

Exercício 2

Mostre que $\sqrt{2}$ é um número irracional.

- Número romano: MDCCCCLXXXV = 1985.
- O sistema posicional faz uso de um aspecto essencial: o zero é representado por um símbolo.
- Os babilônios em 300 a.C. usavam um símbolo para representar o zero.
- O nosso sistema numérico foi desenvolvido na Índia por volta de 600 d.C.
- Após o ano de 1200 iniciou-se o uso dos números arábicos, notadamente devido a obra "Liber Abaci" (ou Livro do Cálculo) escrito pelo matemático italiano Leonardo Pisano Bigollo, mais conhecido como Fibonacci.

- O sistema decimal requer 10 símbolos (0 a 10). A escolha é pautada em função do número de dedos.
- Os babilônios tinham um outro sistema em base 60.
- O zero é necessário para distinguir 601 de 61.
- Sistema decimal foi utilizado inicialmente apenas para números inteiros.
- Embora o sistema decimal é conveniente para pessoas, o mesmo não acontece para computadores.
- O sistema binário é mais útil, no qual cada número é representado por uma palavra de bits.
- Cada bit corresponde a uma potência de 2.
- Um bit pode ser visto como um entidade física que está ligado ou desligado. Em eletrônica sabemos que o bit é representado por um nível de tensão baixo ou alto.
- Bits são organizados em grupos de 8, chamado de *byte*.
- Cada byte pode representar $2^8 = 256$ diferentes números.

- O número $(71)_{10} = 7 \times 10 + 1$ tem sua representação binária como

  $(1000111)_2 = 1 \times 64 + 0 \times 32 + 0 \times 16 + 0 \times 8 + 1 \times 4 + 1 \times 2 + 1 \times 1.$

- O número fracionário também pode ser representado. Considere o exemplo

$$\frac{11}{2} = (5{,}5)_{10} = 5 \times 1 + 5 \times \frac{1}{10}$$

e

$$\frac{11}{2} = (101.1)_2 = 1 \times 4 + 0 \times 2 + 1 \times 1 + 1 \times \frac{1}{2}.$$

Exercício 3

Faça a transformação de $\frac{1}{10}$ para o sistema binário.

- Números irracionais possuem expansão decimal e binária infinita e sem repetição.

Exercício 4

Faça a expansão decimal e binária dos seguintes números: $\sqrt{2}, \pi, e$.

- Qual é a melhor maneira de representar números em um computador?
- Inteiros são representados por 32 bits. O inteiro 71 seria armazenado como

$$00000000000000000000000001000111.$$

- Faixa: 0 a $2^{32} - 1$ ou 0 a 4294967295.
- O número 2^{32} não é possível de ser representado.

Exercício 5

Qual é o número mínimo de bits necessários para representar o número 50000?

- É necessário representar números positivos e negativos.
- A idéia mais simples é representar o número com duas partes sinal e módulo.
- Neste caso, utiliza-se 1 bit para o sinal e 31 bits para para armazenar o módulo do número.
- Entretanto, o método mais comum é o complemento de 2.
- Seja x tal que $0 \leq x \leq 2^{31} - 1$ é representado em sua forma binária.
- Já o o valor negativo $-y$ tal que $-1 \leq y \leq 2^{31}$ é armazenado na representação binária do inteiro positivo

$$2^{32} - y.$$

Exercício 6

Coloque o número 71 na forma binária de complemento de 2.

- Em um sistema de 32 bits, se dois números positivos forem adicionados e o resultado for maior que $2^{31} - 1$ ocorre o chamado integer overflow.
- Subtração de dois números inteiros representados na representação do complemento de 2 não necessita de hardware adicional.

- Números racionais podem ser representados por pares de inteiros: o numerador e denominador.
- Esta representação é precisa, mas é inconveniente do ponto de vista aritmético.
- Sistemas que representam os números racionais dessa forma tem sido chamados de simbólicos.
- Para a maioria dos casos, os números reais, entretanto, são armazenados usando representação binária.
- Há dois métodos para a representação binária: ponto fixo e ponto flutuante.
- Ponto fixo: 1 bit para o sinal, um grupo de bits para representar o número antes do ponto binário e um grupo de bits para representar o número após o ponto binário.

Exemplo 2

Para uma precisão de 32 bits o número 11/2 pode ser representado como

$$|0|000000000000101|1000000000000000|.$$

Exemplo 3

Represente o número 1/10 na representação binária de ponto fixo com 32 bits.

- Faixa: aproximadamente de 2^{-16} a 2^{15}.
- O ponto fixo é bastante limitado quanto a faixa que se pode armazenar.
- É atualmente pouco usado para computação numérica.
- Entretanto microcontroladores com ponto fixo são mais econômicos, possuem circuitos internos mais simples e necessitam de menos memória [1].

[1] Anoop, C. V. and Betta, C. *Comparative Study of Fixed-Point and Floating Code for a Fixed-Point Micro*. dSPACE User Conference - India, 2012.

- Ponto flutuante é baseado na notação exponencial ou científica.
- Um número x é representado por

$$x = \pm S \times 10^E, \quad \text{em que} \quad 1 \leq S < 10,$$

- em que E é um inteiro. Os números S e E são chamados de significante ou mantissa e expoente, respectivamente.

Exemplo 4

A representação exponencial de 0,00036525 é 3.6525×10^{-4}.

- O ponto (vírgula) decimal flutua para a posição imediatemente posterior ao primeiro dígitio não nulo. Está é a razão para o nome ponto flutuante.
- No computador, utiliza-se a base 2. Assim x é escrito como

$$x = \pm S \times 10^E, \quad \text{em que} \quad 1 \leq S < 2. \tag{1}$$

- A expansão binária do significante é

$$S = (b_0 b_1 b_2 b_3 \ldots) \quad \text{com} \quad b_0 = 1. \tag{2}$$

Exercício 7

O número 11/2 é representado como

$$\frac{11}{2} = (1{,}011)_2 \times 2^2.$$

- Os bits após o ponto binário são chamados de parte fracionária do significando.
- As Eq. (1) e (2) são representações normalizadas de x e o processo de obtenção desta representação chama-se normalização.
- Para representar um número normalizado, a sua representação binária é dividida em três partes: sinal, expoente E e o significante S , nesta ordem.

Exemplo 5

Um sistema de 32 bits pode ser dividido nos seguintes campos: 1 bit para o sinal, 8 para para o expoente e 23 bits para o significante.

- Sinal: 0 para positivo e 1 para negativo.
- E pode ter os valores entre -128 e 127 (usando complemento de 2).
- S utiliza 23 bits para armazenar o valor após o ponto.

- Não é necessário armazenar b_0 (bit escondido) pois é sempre o valor de 1.
- Se $x \in R$ pode ser armazenado exatamente em um computador então x é chamado de número ponto flutuante. Senão, x deve ser arredondado.

Exemplo 6

O número 71 é representado por $(1.000111)_2 \times 2^6$ e é armazenado

$$|0|\ \text{ebits}(6)\ |00011100000000000000000|.$$

- ebits(6) representa a conversão de 6 para o número binário no expoente.
- Se um número x não tem uma expansão binária finita, é necessário terminar esta expansão. Esse processo é chamado de truncamento.

Exemplo 7

Considere o número 1/10, cuja expansão é

$$\frac{1}{10} = (0{,}0001100110011\ldots)_2$$

. Primeiro devemos normalizar e em seguida truncar. Assim a representação de 1/10 é

$$|0|\ \text{ebits}(-4)\ |10011001100110011001100|.$$

- A precisão (p) de um sistema de ponto flutuante é o número de bits do significante (incluindo o bit escondido).
- No sistema de 32 bits, $p = 24$, sendo 23 bits no significante e 1 bit escondido.
- Qualquer ponto flutuante normalizado com precisão p pode ser expresso como

$$x = \pm(1,b_1 b_2 \ldots b_{p-2} b_{p-1})_2 \times 2^E. \tag{3}$$

- O menor ponto flutuante x que é maior que 1 é

$$(1,000\ldots1)_2 = 1 + 2^{-(p-1)}.$$

- Dá-se um especial nome machine epsilon (epsilon da máquina) a distância entre este número e o número 1:

$$\varepsilon = (0,000\ldots01)_2 = 2^{-(p-1)}. \tag{4}$$

- De modo mais geral, para um ponto flutuante x dado por (3) nós definimos

$$\mathrm{ulp}(x) = (0,00\ldots01)_2 \times 2^E = 2^{-(p-1)} \times 2^E = \varepsilon \times 2^E. \tag{5}$$

- Ulp é abreviação de unit in the last place ou unidade da última posição.
- Se $x > 0$ então ulp(x) é a distância entre x e o próximo maior ponto flutuante maior.
- Se $x < 0$ então ulp(x) é a distância entre x e o próximo menor ponto flutuante.

Exercício 8

Seja a precisão $p = 24$ tal que $\varepsilon = 2^{-23}$. Determine ulp(x) para x igual aos seguintes valores: a) 0,25; b) 2; c) 3; d) 4; e) 10; f) 100; g) 1030. Dê sua resposta em potência de 2.

- Até o momento, foi discutido apenas números não nulos.
- O zero não pode ser representado com o uso do bit escondido. 0000 . . . representa 1.
- Até 1975, resolvia esta questão não utilizando o bit escondido.
- A norma IEEE reduz em 1 o expoente e utiliza um caracter especial para identificar o zero.

- Considere o seguinte sistema binário fictício em que todos os números podem ser representados da seguinte forma

$$\pm(b_0 b_1 b_2)_2 \times 2^E). \tag{6}$$

- b_0 é permitido ser 0 se b_1 e b_2 forem também zero. Neste caso, o número decimal representado é o zero.
- O número E pode ser -1, 0 ou 1.
- O conjunto de número representados pode ser visto na Figura

Figure 7: Conjunto de números representados por (6). Fonte: Livro (Overton, 2001, p. 15).

- A precisão de (6) é $p = 3$.
- O maior número é $(1{,}11)_2 \times 2^1 = (3{,}5)_{10}$.
- O menor número positivo é $(1{,}00)_2 \times 2^{-1} = (0{,}5)_{10}$.
- O ponto flutuante seguinte ao número 1 é 1,25, assim $\varepsilon = 0{,}25$.
- A distância entre cada número é dada por

$$\text{ulp}(x) = \varepsilon \times 2^E.$$

- A distância entre 0 e $\pm 0{,}5$ é maior do que a distância entre $\pm 0{,}5$ e ± 1. Essa distância pode ser reduzida com a introdução dos números subnormalizados.