Aritmética con redondeo

Requisitos. Notación científica, formatos float32 y float64.

Redondeo. El estándar IEEE-754 exige que el resultado de las operaciones sea el mismo que se obtendría si se realizasen con precisión absoluta y después se redondease. Hay 4 modos de redondeo:

- redondeo a cero (*truncamiento*);
- redondeo al más cercano (al par en caso de empate), llamado brevemente *redondeo*;
- redondeo a más infinito (por exceso);
- redondeo a menos infinito (por defecto).

1. Redondear los siguientes números binarios a 3 dígitos después del punto, usando varios modos de redondeo:

$$1.010001 \qquad 1.01111 \qquad 1.0001 \qquad 1.0011$$

Suma de números de punto flotante

2. Sumar los siguientes números binarios y redondear el resultado a 3 dígitos después del punto, usando el redondeo al más cercano:

$$1.001 \cdot 2^5 + 1.111 \cdot 2^4; \qquad 1.010 \cdot 2^{-3} + 1.010 \cdot 2^{-5};$$
$$1.001 \cdot 2^1 + 1.101 \cdot 2^{-1}; \qquad 1.000 \cdot 2^7 + 1.011 \cdot 2^5.$$

3. Sumar los siguientes números binarios y redondear el resultado a 3 dígitos después del punto, usando el redondeo al más cercano:

$$1.000 + 1.000 \cdot 2^{-3}, \qquad 1.000 + 1.000 \cdot 2^{-4}.$$

Errores de redondeo (ejemplos)

4. Para cada una de las siguientes expresiones, calcular el valores exacto, el valor en aritmética con tres dígitos decimales (dos dígitos después del punto flotante), el error absoluto y el error relativo:

- $3.50 \cdot 3.50$
- $9.90 \cdot 10^3 + 1.50 \cdot 10^2$
- $(1.00 + 4.70 \cdot 10^{-2}) - (1.00 + 3.20 \cdot 10^{-2})$
- $\left(\frac{2}{9}\right) \cdot \left(\frac{9}{7}\right)$. $0.222 \cdot 1.28 \approx 0.284$; 0.2857; $1.7 \cdot 10^{-3}$, $6.0 \cdot 10^{-3}$

5. Tarea. Lo mismo para las siguientes expresiones:

- $\dfrac{\pi - \frac{22}{3}}{\frac{1}{17}}$
- $(121 - 0.327) - 119$
- $(121 - 119) - 0.327$

Épsilon de la máquina

6. Observación. El número de dígitos binarios en la mantisa es acotado. Por eso las sumas de la forma $1 + x$, donde x es muy pequeño, se redondean al número 1.

7. Definición. ε de la máquina es el número x positivo más pequeño tal que $1 + x$ se puede representar de manera precisa en la máquina.

Proposición. En el sistema binario el épsilon de la máquina es igual a 2^{-n}, donde n es la longitud de la mantisa sin tomar en cuenta el bit implicito antes del punto flotante).

8. Programa en C que calcula el épsilon de la máquina.

```
#include <stdio.h>

int main() {
    double x = 1.0;
    int n = 0;
    while (1.0 + (x * 0.5) > 1.0) {
        ++n; x *= 0.5;
    }
    printf("Epsilon de la maquina en forma binaria = 2^(-%d)\n", n);
    printf("Epsilon de la maquina en forma decimal = %G\n", x);
    return 0;
}
```

Violación de las leyes de la aritmetica común en la aritmética con redondeo

9. Violación de la ley asociativa en la aritmética con cuatro cifras. Usando la aritmética con cuatro cifras (tres cifras después del punto), calcular las expresiones

$$(a + b) + c \qquad \text{y} \qquad a + (b + c)$$

para $a = 1.000$, $b = c = 1.000 \cdot 2^{-4}$.

10. Violación de la ley asociativa en Mathematica. Ejecutar el siguiente programa en Mathematica y explicar los resultados:

```
x = 1.0
eps = 0.5 ^ 53
y = (x + eps) + eps
z = x + (eps + eps)
y - 1
z - 1
```

11. Tarea creativa: violación de la ley distributiva. Inventar un ejemplo cuando en aritmética de punto flotante con redondeo a tres dígitos no se cumple la ley distributiva, i.e. hallar a, b, c tales que

$$\text{fl}(\text{fl}(a + b) * c) \neq \text{fl}(\text{fl}(a * c) + \text{fl}(b * c)).$$