Assignment #1
Numerical Computing (COMP 350)

Guillaume Labranche (260585371)

due on 23 September 2015

1. **No**. Let x be a real number with finite binary representation and x_i be the i$^{\text{th}}$ bit after the point. The integer part can be represented in finite decimal form in all cases. For all x_i in the fractional part, if the digit is 1, its value within x is $1/(2^i)$ and that in decimal form is finite, so adding all x_i's leads to a finite decimal representation.

2. $\mathbf{-23}$

3. (a) $(2^4)/2 - 1 = 16/2 - 1 = 8 - 1 = \mathbf{7}$

   (b) Normalized nonnegative range:

	Sign	Exponent	Fraction	Decimal Value
N_{min}	1	$(0001)_2 = 1$	$(00000)_2$	$(1.00000)_2 \cdot 2^{-6} = \mathbf{(1/64)_{10}}$
N_{max}	1	$(1110)_2 = 14$	$(11111)_2$	$(1.11111)_2 \cdot 2^{7} = \mathbf{(252)_{10}}$

   (c) Subnormal nonnegative range:

	Sign	Exponent	Fraction	Decimal Value
Smallest	1	$(0000)_2 = 1$	$(00001)_2$	$(0.00001)_2 \cdot 2^{-6} = \mathbf{2^{-11}}$
Largest	1	$(0000)_2 = 15$	$(11111)_2$	$(0.11111)_2 \cdot 2^{-6} = \mathbf{31/2^{-11}}$

   (d) $\epsilon = \mathbf{2^{-5}}$

   (e) **10.75** and **11.15**

Sign	Exponent	Fraction	Decimal Value
1	$(1010)_2 = 10$	$(01011)_2$	$(1.01011)_2 \cdot 2^3 = 10.75$
1	$(1010)_2 = 10$	$(01100)_2$	$(1.01100)_2 \cdot 2^3 = 11.0$
1	$(1010)_2 = 10$	$(01101)_2$	$(1.01101)_2 \cdot 2^3 = 11.25$

   (f)

$$
\begin{aligned}
x &= -(1.0110101)_2 \cdot 2^0 \\
x_+ &= -(1.01101)_2 \cdot 2^0 \\
x_- &= -(1.01110)_2 \cdot 2^0
\end{aligned}
$$

   - Round down: $x_- = -(1.01110)_2 \cdot 2^0$

- Round up: $x_+ = -(1.01101)_2 \cdot 2^0$
- Round towards zero: $x_+ = -(1.01101)_2 \cdot 2^0$
- Round to nearest: $x_+ = -(1.01101)_2 \cdot 2^0$

4. (a) **True**. When adding a number x to itself, we are doubling its value ($2x$). In binary representation, this comes down to shifting the decimal point 1 position to the right, so the significant stays the same but E is increased by 1 when x is in the form $(b_0.b_1b_2 \ldots b_{23})_2 \cdot 2^E$.

$$
\begin{aligned}
x \oplus x &= 2x \\
\text{round}(x + x) &= 2x \\
\text{round}(2x) &= 2x \\
\text{round}(2 \cdot (b_0.b_1b_2 \ldots b_{23})_2 \cdot 2^E) &= 2 \cdot (b_0.b_1b_2 \ldots b_{23})_2 \cdot 2^E \\
\text{round}((b_0.b_1b_2 \ldots b_{23})_2 \cdot 2^{E+1}) &= (b_0.b_1b_2 \ldots b_{23})_2 \cdot 2^{E+1}
\end{aligned}
$$

(b) **False**. Counter-example (rounding mode is **round down**):

$$
\begin{array}{rrlr}
x = & (& 1.00000000000000000000000| &)_2 \cdot 2^0 \\
y = & -(& 0.11111111111111111111111|1 &)_2 \cdot 2^0 \\
x - y = & (& 0.00000000000000000000000|1 &)_2 \cdot 2^0 \\
\boldsymbol{x \ominus y} = & (& 0.00000000000000000000000| &)_2 \cdot 2^0 \\
y - x = & -(& 0.00000000000000000000000|1 &)_2 \cdot 2^0 \\
y \ominus x = & -(& 0.00000000000000000000001| &)_2 \cdot 2^0 \\
\boldsymbol{-(y \ominus x)} = & (& 0.00000000000000000000001| &)_2 \cdot 2^0
\end{array}
$$

5.
- $\infty/0 = \boldsymbol{\infty}$ (because $a/0 = \infty$)
- $\infty/(-\infty) =$ **NaN** (any operation involving NaN results in NaN)
- $1^{\text{NaN}} =$ **NaN** (same reason)
- $-0/\text{NaN} =$ **NaN** (same reason)