# Assignment 1

**Due**. In class Sept. 21 (Friday)

1. Write a MATLAB program for each of the following problems:
   (1) compute the unit of roundoff
   **Solution**

```
u = 1.0;
while 1.0 + u > 1.0
    u = u/2;
end
u,
```

   (2) find the largest (normal) floating-point number on a system. Explain how it works and count the numbers of floating-point additions, multiplications, and divisions.
   **Solution**

```
x = 2.0; xTx = x*x;
while (xTx > x)
   xPre = x;
   x = xTx;
   xTx = x*x;
end                      % at this point, xPre*xPre = infty
x = xPre; xPx = x + x;
while (xPx > x)
   xPre = x;
   x = xPx;
   xPx = x + x;
end                      % at this point, xPre = 2^{e_max}
x = xPre; frac = x/2; xPfrac = x + frac;
while (xPfrac > x)
   xPre = x;
   x = xPfrac;
   frac = frac/2;
   xPfrac = x + frac;
end
xPre,
```

   In double precision, it takes 12 multiplications, 568 additions, and 55 divisions.

2. Consider a three-digit decimal floating-point number system $F$, with numbers

$$x = \pm .d_1 d_2 d_3 \times 10^e,$$

   with $-100 \le e \le 100$ and $0 \le d_i \le 9$. Let $F$ be normalized (i.e., $d_1 \neq 0$ unless $x = 0$). The two zeros ($\pm 0$) have the reprentation $\pm .000 \times 10^{-100}$.

(a) How many different real numbers can be exactly represented in $F$?

(b) Find examples of $x$, $y$, $z$ in $F$ to show that the following statements are *not* generally true, even if the result is within the range of $F$:

- $((x \otimes y) \otimes z) = (x \otimes (y \otimes z))$.
- $((x \oplus y) \oplus z) = (x \oplus (y \oplus z))$.

(c) Find an example where $((x \oplus y) \oplus z)$ has relative error of at least 50%.

**Solution** Assuming the nearest even rounding.

(a) In each interval $[10^e, 10^{e+1})$, there are 900 numbers (from 0.100 to 0.999). There are 201 intervals (from $-100$ to 100). Thus there are $201 \times 900 = 180900$ positive numbers. The total is $2 \times 180900 + 2 = 361802$.

- When $x = 0.550$, $y = 0.101$, $z = 0.350$, $((x \otimes y) \otimes z) = 0.198 \times 10^{-1}$ and $(x \otimes (y \otimes z)) = 0.195 \times 10^{-1}$.
- When $x = 0.999$, $y = 0.499$, $z = 0.456 \times 10^3$, $((x \oplus y) \oplus z) = 0.458 \times 10^3$ and $(x \oplus (y \oplus z)) = 0.457 \times 10^3$.
- When $x = 0.249 \times 10^{-1}$, $y = 0.142 \times 10^1$, $z = -0.145 \times 10^1$, $((x \oplus y) \oplus z) = 0.100 \times 10^{-1}$. The exact result is $0.51 \times 10^{-2}$. The relative error is 96%.

3. What output is produced when the following MATLAB program is run on your computer? Explain (quantitatively) why.

```
x = 0.0;
h = 0.1;
for i = 1:10
    x = x + h;
end;
y = 1.0 - x;
x, y,
```

**Solution.** The decimal number

$$0.1_{10} = 1.100110011001... \times 2^{-4}$$

cannot be exactly represented in binary. When it is stored as a double precision floating-point number $h$, it may be rounded to

$$h = 1.10011001100 \cdots 110011010_2 \times 2^{-4},$$

which is larger than $0.1_{10}$. Due to rounding errors in addition, after the first five iterations $x = 1.0 \times 2^{-1} = 0.5_{10}$. In the subsequent iterations, each time when $h$ is added to $x$, the result is rounded down by about $2^{-51} \times 2^{-4} = 2^{-55}$, because the three least significant bits of $h$ are shifted and rounded off. Accumulated in the last five iterations, the total error is about $5 \times 3 \times 2^{-57} \approx 1.1 \times 10^{-16}$.

4. Find the condition number for adding two numbers.
**Solution**

$$\frac{|a(1+\epsilon_1) + b(1+\epsilon_2) - (a+b)|}{|a+b|} = \frac{|a\epsilon_1 + b\epsilon_2|}{|a+b|} \leq \frac{|a|+|b|}{|a+b|} \max(|\epsilon_1|, |\epsilon_2|)$$

The condition number is $(|a| + |b|)/|a + b|$.

5. Answer the following questions for a small floating-point system where base is 2, precision 3, $e_{\min} = -2$, and $e_{\max} = 3$.

   (a) How many floating-point numbers $x$ satisfying $1 \le x < 2$? How many of these satisfy $1 \le x < 3/2$ and how many satisfy $3/2 \le x < 2$?

   (b) How many floating-point numbers $y$ satisfying $1/2 < y \le 1$? How many of these satisfy $1/2 < y \le 2/3$ and approximately how many satisfy $2/3 < y \le 1$?

   (c) Does it follow that there must exist two different floating-point numbers $x_1$ and $x_2$ for which the computed reciprocals $\mathrm{fl}(1/x_1)$ and $\mathrm{fl}(1/x_2)$ are the same (rounded to the same format)? Are you thinking of $x_1$ and $x_2$ between 1 and 3/2 or between 3/2 and 2? Is this true regardless of the rounding mode?

**Solution** In our small system, $\beta = 2$, $t = 3$, $e_{\min} = -2$, and $e_{\max} = 3$,

- Floating-point numbers in $[1, 2)$:

| interval | small system |
|---|---|
| $[1, 2)$ | four floating-point numbers |
| $[1, 3/2)$ | $1.00 \times 2^0$, $1.01 \times 2^0$ |
| $[3/2, 2)$ | $1.10 \times 2^0$, $1.11 \times 2^0$ |

- Floating-point numbers in $(1/2, 1]$:

| interval | small system |
|---|---|
| $(1/2, 1]$ | four floating-point numbers |
| $(1/2, 2/3]$ | $1.01 \times 2^{-1}$ |
| $(2/3, 1]$ | $1.10 \times 2^{-1}$, $1.11 \times 2^{-1}$, $1.00 \times 2^0$ |

- There are two numbers in $[3/2, 2)$. Their reciprocals are in $(1/2, 2/3]$ (or $[1/2, 2/3]$ or $(1/2, 3/4]$ depending on rounding mode), but there is only one number in $(1/2, 2/3]$. So, there must exist $x_1$ and $x_2$ such that $x_1 \ne x_2$ and $\mathrm{fl}(1/x_1) = \mathrm{fl}(1/x_2)$. (Pigeon hole priciple.) In the small system

  $x_1 = 1.10 \times 2^0, \quad x_2 = 1.11 \times 2^0, \quad \mathrm{fl}(1/x_1) = \mathrm{fl}(1/x_2) = 1.01 \times 2^{-1}$ (nearest rounding).

  All $\mathrm{fl}(1/x)$, $x \in [1, 2)$, are listed in the following table.

| $x$ | | $1(1.00 \times 2^0)$ | $5/4(1.01 \times 2^0)$ | $3/2(1.10 \times 2^0)$ | $7/4(1.11 \times 2^0)$ |
|---|---|---|---|---|---|
| $1/x$ | $-\infty$, 0 | $1(1.00 \times 2^0)$ | $3/4(1.10 \times 2^{-1})$ | $5/8(1.01 \times 2^{-1})$ | $1/2(1.00 \times 2^{-1})$ |
| | nearest | $1(1.00 \times 2^0)$ | $3/4(1.10 \times 2^{-1})$ | $5/8(1.01 \times 2^{-1})$ | $5/8(1.01 \times 2^{-1})$ |
| | $+\infty$ | $1(1.00 \times 2^0)$ | $7/8(1.11 \times 2^{-1})$ | $3/4(1.10 \times 2^{-1})$ | $5/8(1.01 \times 2^{-1})$ |

  Similarly, there exist two numbers in $(2/3, 1]$:

  $y_1 = 1.11 \times 2^{-1}, \quad y_2 = 1.10 \times 2^{-1}, \quad \mathrm{fl}(1/y_1) = \mathrm{fl}(1/y_2) = 1.01 \times 2^0$ (nearest rounding).

  We also list all $\mathrm{fl}(1/y)$, $y \in (1/2, 1]$:.

| $y$ | | $1(1.00 \times 2^0)$ | $7/8(1.11 \times 2^{-1})$ | $3/4(1.10 \times 2^{-1})$ | $5/8(1.01 \times 2^{-1})$ |
|---|---|---|---|---|---|
| $1/y$ | $-\infty$, 0 | $1(1.00 \times 2^0)$ | $1(1.00 \times 2^0)$ | $5/4(1.01 \times 2^0)$ | $3/2(1.10 \times 2^0)$ |
| | nearest | $1(1.00 \times 2^0)$ | $5/4(1.01 \times 2^0)$ | $5/4(1.01 \times 2^0)$ | $3/2(1.10 \times 2^0)$ |
| | $+\infty$ | $1(1.00 \times 2^0)$ | $5/4(1.01 \times 2^0)$ | $3/2(1.10 \times 2^0)$ | $7/4(1.11 \times 2^0)$ |