BLG 202E HOMEWORK #1

1) In general floating point system which is consist of β, t, L, U

   $Rounding\ unit = \frac{1}{2}\beta^{1-t}$

   For $\beta = 2\ and\ t = 52$:

   $Rounding\ unit = \frac{1}{2}\beta^{1-t} \rightarrow \frac{1}{2}2^{-51} \rightarrow \mathbf{2^{-52}}$

2) $Real\ answer\ of\ x - y = 0{,}00016$
   a. With rounding:

      $x = 0{,}8532\ \ y = 0{,}8530 \rightarrow x - y = 0{,}0002$

      $$Absolute\ Error = |real - found| = |0{,}00016 - 0{,}0002| = \mathbf{0{,}00004}$$

      $$Relative\ Error = \frac{|real - found|}{real} = \frac{|0{,}00016 - 0{,}0002|}{0{,}00016} = \mathbf{0.25}$$

   b. With chopping:

      $x = 0{,}8531\ \ y = 0{,}8530 \rightarrow x - y = 0{,}0001$

      $$Absolute\ Error = |real - found| = |0{,}00016 - 0{,}0001| = \mathbf{0{,}00006}$$

      $$Relative\ Error = \frac{|real - found|}{real} = \frac{|0{,}00016 - 0{,}0001|}{0{,}00016} = \mathbf{0.375}$$

3)
   a. When we do matrix multiplication on the left side, we get:

      $$ax + by = 1\ \ bx + ay = 0$$

      If we add second equation to the first one and do some proper calculations:

      $$ax + ay + bx + by = 1$$

      $$a(x + y) + b(x + y) = 1$$

      $$(x + y)(a + b) = 1$$

      $$(x + y) = \frac{\mathbf{1}}{\boldsymbol{a + b}}$$

   b.
      i. While determining a condition of linear system solving, we should obtain expressions for the variables which are given,

      $$\begin{bmatrix} a & b \\ b & a \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \rightarrow ax + by = 1, bx + ay = 0$$

If we solve the equation with two unknowns, we find:

$$x = \frac{a}{a^2 - b^2} \; and \; y = \frac{b}{b^2 - a^2}$$

We should determine effects of small changes in a and b on x and y. To illustrate,

If $a = 1.00001$ and $b = 1 \rightarrow x \approx 50000, y \approx -49999$

If $a = 1.00002$ and $b = 1 \rightarrow x = 0.00004, y \approx -0.00004$

As we saw, small changes in values causes too big differences. Because of that solving linear system is ill-conditioned.

ii. As we found in part a:

$$(x + y) = \frac{1}{a + b}$$

To reach a result in condition checking, we should change values very little.

If $a = 1.00001$ and $b = 1 \rightarrow x + y \approx 0.49999 \approx \frac{1}{2}$

If $a = 1.00002$ and $b = 1 \rightarrow x + y \approx 0.49999 \approx \frac{1}{2}$

Proximity of values does not affect our results significantly. Because of that solving linear system is well-conditioned.

c. In general floating systems, β, t, L, U determine limits of the systems. With the help of these values, we are able to calculate truth accuracy of our results, biggest and smallest numbers that we can represent and our calculation sensitivity.

d. Standard formula for roots of the quadratic equation is: $\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

i. $x_1 = \frac{10^5 + \sqrt{(-10^5)^2 - 4*1*1}}{2*1}$ $\quad$ $x_2 = \frac{10^5 - \sqrt{(10^5)^2 - 4*1*1}}{2*1}$

While calculating x_2, value of the square root approximately equal to 10^5. This means share of the fraction value is calculated as 0 and cancellation error occurs.

Solution: To avoid cancellation error, we can use $\boldsymbol{x_1 * x_2 = c/a}$ formula. Firstly, we should calculate x_1 and using formula we can obtain x_2.

$$x_1 = \frac{10^5 + \sqrt{(-10^5)^2 - 4*1*1}}{2*1}, \; x_2 = \frac{c}{a*x_1} = \frac{1}{x_1}$$

ii.

$$x_1 = \frac{-(5 * 10^{30}) + \sqrt{(\mathbf{5 * 10^{30}})^2 - 4 * (\mathbf{6 * 10^{30}}) * (\mathbf{-4 * 10^{30}})}}{2 * (6 * 10^{30})}$$

$$x_2 = \frac{-(5 * 10^{30}) - \sqrt{(\mathbf{5 * 10^{30}})^2 - 4 * (\mathbf{6 * 10^{30}}) * (\mathbf{-4 * 10^{30}})}}{2 * (6 * 10^{30})}$$

While calculating inside of the square root, numbers exceed $\mathbf{10^{60}}$. But our upper bound is $\mathbf{10^{50}}$. Therefore, we cannot calculate and represent results of the multiplications.

Solution: To avoid overflow, we should **rescale our equation with any s** $\neq 0$,

$$c = \sqrt{b^2 - 4ac} \rightarrow c = s\sqrt{\left(\frac{b}{s}\right)^2 - 4\frac{a}{s}\frac{c}{s}}$$

If we apply this method with s $= \mathbf{10^{30}}$,

$$x_1 = \frac{-(5 * 10^{30}) + \mathbf{10^{30}} * \sqrt{(\mathbf{5})^2 - 4 * (\mathbf{6}) * (\mathbf{-4})}}{2 * (6 * 10^{30})}$$

$$x_2 = \frac{-(5 * 10^{30}) - \mathbf{10^{30}}\sqrt{(\mathbf{5})^2 - 4 * (\mathbf{6}) * (\mathbf{-4})}}{2 * (6 * 10^{30})}$$