Exercise Sheet 2

1. Consider the normalized floating point number system $\mathbb{F} = \mathbb{F}(2, 3, -1, 3)$.
   (a) Compute all the positive elements of $\mathbb{F}$.
   (b) Find the unit roundoff and the spacing between adjacent floating point numbers, using Theorem 1.1.
   (c) Compute explicitly the subnormal numbers of $\mathbb{F}$.
2. Find examples to show that the floating point multiplication, in general, is neither associative nor distributive (over floating point addition). You may use the calculator of Example 1.1.
3. Compute $\kappa_2(A)$, i.e. the condition number with respect to the spectral norm, of the matrix
$$A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}.$$
4. Show that the floating point multiplication and division are backward stable.
5. Compute the first order Taylor approximations of the following expressions
$$\frac{1}{1+\epsilon}, \quad \sqrt{1+\epsilon}, \quad \log(1+\epsilon), \quad \sin\epsilon.$$
6. Consider the problem $f : x \mapsto \cos x$ for $x \approx 0$. Is it well- or ill-conditioned? Suppose you have a perfect implementation of the cosine that gives you $\tilde{f}(x) = \mathrm{fl}(\cos x)$. Is the evaluation of $\tilde{f}$ (backward) stable?
7. Consider the problem $f : x \mapsto \sqrt{x+1} - 1$ for $x \approx 0$. Is it well- or ill-conditioned? Rewrite f in a way that avoids cancellation.
8. Consider the linear system $Ax = b$ with
$$A = \begin{bmatrix} 1 & 4 \\ 1 & 4.001 \end{bmatrix}, \quad b = \begin{bmatrix} 3 \\ 4 \end{bmatrix}.$$
   Check if the system is well- or ill-conditioned. For a perturbation of the right-hand side $\Delta b = [0.01,\, 0]^\top$, compute the relative error of the solution.
9. The linear system
$$\begin{bmatrix} a_{11} & a_{12} \\ 0 & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

can be easily solved by the backward substitution. On a machine, this method results to a solution of the form

$$\tilde{x}_2 = b_2 \oslash a_{22},$$
$$\tilde{x}_1 = (b_1 \ominus \tilde{x}_2 \otimes a_{12}) \oslash a_{11}.$$

Show that the computed solution (given Assumption 1.1) satisfies a linear system of the form

$$\begin{bmatrix} \tilde{a}_{11} & \tilde{a}_{12} \\ 0 & \tilde{a}_{22} \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

where in addition holds

$$\frac{|\tilde{a}_{ij} - a_{ij}|}{|a_{ij}|} \leq c\, u,$$

for u being the unit roundoff and c a small constant.