Question:	1	2	3	4	5	6	7	Total
Points:	15	10	20	10	20	15	10	100

Computer representation of integer numbers

1. Convert the following numbers to decimal representation. Show your work.

   (a) (5 points) 10001100_2

   76543210

   $10001100_2 = 2^7 + 2^3 + 2^2 = 128_{10} + 8_{10} + 4_{10} = 140_{10}$

   (b) (5 points) $B01_{16}$

   A-10, B-11, C-12, D-13, E-14, F-15

   210

   $B01_{16} = 11 \cdot 16^2 + 16^0 = 11 \cdot 256 + 1 = 2817_{10}$

   (c) (5 points) 301_8

   210

   $301_8 = 3 \cdot 8^2 + 8^0 = 3 \cdot 64 + 1 = 193_{10}$

2. You are designing a specialized computer memory that is supposed to store **only non-negative** integers. You are planing to store your integers in eight bits.

   (a) (5 points) What is the largest number you can store?

   $\underset{76543210}{11111111} = \underbrace{2^7 + 2^6 + \ldots + 2^2 + 2^1 + 2^0}_{\text{geometric progression}} = 2^8 - 1 = 255$

   (b) (5 points) How many different integer numbers are in your system?

   Every bit combination is a valid integer.
   Eight bits → 2^8 combinations → 2^8 integers

Computer representation of floating point numbers

3. You are developing a specialized microprocessor to store the results of your measurement. The specification requires that the processor is using chopping when operating with floating point numbers. Furthermore, it is required that the floating point numbers are stored in a manner similar to IEEE754 standard: one bit for the sign, several bits for the exponent, and another group of bits for the fractional part of the number. You do **not** need to reserve special bit combinations, e.g. for zero, infinity, and NaN.

   The key requirements to your processor are as following: (a) the largest (in absolute value) number that you expected to process is $\sim 2^9$; (b) the relative error that is produced when a measurement is stored in the system is $\sim 2^{-8}$

   (a) (5 points) What is machine ϵ in your system? Explain.

   Relative error $\simeq$ Machine epsilon

   $\varepsilon = 2^{-8}$

   (b) (5 points) How many bits you reserve for the fractional part of a floating point number? Explain.

   $\varepsilon = \frac{0}{2} + \frac{0}{2^2} + \frac{0}{2^3} + \dots + \frac{0}{2^7} + \frac{1}{2^8} = 0.\underbrace{00000001}_{\text{eight bits in the fractional part}}$

   (c) (5 points) How many bits you reserve for the exponent? Explain.

   will not be on the test

   $r = (-1)^s \cdot (1+f) \cdot 2^e$; $r_{max} = (1+f_{max}) \cdot 2^{e_{max}}$;

   $f_{max} \approx 1$; $r_{max} = 2^{e_{max}+1} = 2^9$; $e_{max} = 8 \Rightarrow e_{min} = -7$ $\Rightarrow$ four bits

   (d) (5 points) What is the smallest positive floating point number in your system? Explain.

   $r_{min} = (1+f_{min}) \cdot 2^{e_{min}}$; $f_{min} = 0$

   $r_{min} = 2^{e_{min}} = 2^{-7}$

When your answer is a floating point number, provide it as powers of 2.

4. (10 points) As you know the use of the expression

$$\pi - \sqrt{\pi^2 - x},$$

if used with finite-precision floating point arithmetic and for small values of x, $|x| \ll 1$, leads to loss of significance (known as *catastrophic cancellation*).

Rewrite the expression above to fix the catastrophic cancellation problem:

$$\pi - \sqrt{\pi^2 - x} =$$

$$= \frac{(\pi - \sqrt{\pi^2 - x})(\pi + \sqrt{\pi^2 - x})}{\pi + \sqrt{\pi^2 - x}} = \frac{\pi^2 - (\pi^2 - x)}{\pi + \sqrt{\pi^2 - x}} = \frac{x}{\pi + \sqrt{\pi^2 - x}}$$

Matlab

The chemical equation

$$x_1\,[Ca(OH)_2] + x_2\,[HNO_3] \rightarrow x_3\,[Ca(NO_3)_2] + x_4\,[H_2O]$$

indicates that x_1 molecules of calcium hydroxide $Ca(OH)_2$ combine with x_2 molecules of nitric acid HNO_3 to yield x_3 molecules of calcium nitrate $Ca(NO_3)_2$ and x_4 molecules of water H_2O.

Since atoms are not destroyed or created in chemical reactions, the balance of calcium atoms requires that

$$x_1 = x_3.$$

The balance of oxygen atoms requires that

$$2x_1 + 3x_2 = 6x_3 + x_4.$$

The balance of hydrogen atoms requires that

$$2x_1 + x_2 = 2x_4.$$

The balance for nitrogen atoms requires that

$$x_2 = 2x_3$$

5. (a) (5 points) Rewrite the balance equations above in matrix form $Ax = b$:

$$A = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 2 & 3 & -6 & -1 \\ 2 & 1 & 0 & -2 \\ 0 & 1 & -2 & 0 \end{pmatrix}; \quad b = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}; \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

Downloads

(b) (5 points) Write matlab function (call it **chemreaction()**) that accepts no parameters and returns the 4×4 matrix A and 4×1 column vector b your found in Step (a). When called, your function **must** print absolutely nothing.

(c) (5 points) Write a matlab script (call it `chem.m`) that calls your function to initialize A and b and tries to solve the linear equation using Matlab's backslash operator. (Describe what happened in your GitLab's README file.)

(d) (5 points) Add some code to your script that verifies that the vector [1;2;1;2] is the solution of your system of equations.

Git and Gitlab

6. (15 points) Upload the code you wrote for this exam:

   1. Use gitlab web interface to create a new project called **midterm1-sample** (the name is case sensitive, must be exactly as shown)
   2. Use gitlab web interface to add *README* file and edit it to add some meaningful content
   3. Use gitlab web interface to upload your matlab code to your project
   4. Use gitlab web interface to grant the access to your project (with the permission of the *reporter*) to the user `michael.rozman`

Systems of linear equations

7. (10 points) You wrote your own function to solve a system of linear equations. It takes about 10 seconds (on a slow computer) to solve the system of 10 equations with 10 unknowns. **Estimate** how long it would take to solve a system of 11 linear equations with 11 unknowns if

   1. your code implements Cramer's algorithm

   $T_c(n) \sim n!$; $\frac{T_c(11)}{T_c(10)} = \frac{11!}{10!} = 11$; $T_c(11) = 10s \cdot 11 = 110\,s$

   2. your code implements gaussian elimination method

   $T_G(n) \sim n^3$; $\frac{T_G(11)}{T_G(10)} = \left(\frac{11}{10}\right)^3 = (1+.1)^3 \approx 1 + 3 \cdot 0.1 = 1.3$; $T_c(11) = 13\,s$