Homework No.1, 553.385, Due September 17, 2021.

1. (a) Use the definition of limit of a sequence to show that, if $\theta$ is a real number satisfying $0 < \theta < 1$, then $\lim_{n \to \infty} \theta^n = 0$.

(b) Consider the matrix

$$A = \begin{pmatrix} \theta & 1 \\ 0 & 0 \end{pmatrix}$$

for a real number $\theta$. Calculate the matrix powers $A^2$, $A^3$ and then derive the result $A^n$ for general integer $n \geq 1$.

(c) Using your results in parts (a) and (b), show that $\lim_{n \to \infty} A^n$ exists if $0 < \theta < 1$ and determine its matrix value. We shall discuss later in the course what it means precisely for a general sequence of matrices $A_n$ to have a limit as $n \to \infty$.

2. Construct a Taylor series approximation around $x = 0$ for the function $f(x) = \frac{1}{1-x}$ and establish a bound on the error when truncating after $n$ terms.

3. Convert the following numbers to their decimal equivalent:

   (a) $\left(4031.23\right)_5$
   (b) $\left(c73a.4fa\right)_{16}$
   (c) $\left(0.0\overline{1}\right)_3$
   (d) $\left(abc123\right)_{16}$

4. Show that to convert from hexadecimal representation to binary, each hecadecimal digit $b$ is replaced, in order, by the four binary digits $a, a', a'', a'''$ that satisfy

$$b = 2^3a + 2^2a' + 2a'' + a''' .$$

Use this fact to construct a simple algorithm to convert a hexadecimal representation to a binary representation for both integers and fractions. Write a pseudocode to make the conversion $b \rightarrow (a, a', a'', a''')$, using the function $\text{mod}(x, 2)$, which gives the remainder of $x$ when divided by 2.

$Hint$: $a''' = \text{mod}(b, 2)$ and $b' \equiv (b - a''')/2 = 2^2a + 2a' + a''$.

5. In IEEE Standard Floating-Point representation for numbers, a 12-bit binary number is used first to encode the sign $\sigma$ and exponent $E$, followed by a 52-bit binary to encode the fraction $F$. This can also be represented by a 3-digit hexadecimal number, followed by a 13-digit hexadecimal number, as in MATLAB. Consider the following IEEE Standard Floating-Point number in hexadecimal form:

$$x=3ffbb67ae8584caaa$$
(a) Use the first 3 hexadecimal digits to determine the sign $\sigma$ and exponent $E$ of $x$ in decimal format.

(b) Convert the remaining 13 hexadecimal digits to a decimal representation of the fraction $F$ for $x$.

(c) Calculate the decimal representation of the double-precision number $x$.

6. Define $M$ to be the largest positive integer such that it and all smaller positive integers are exactly represented in floating point arithmetic. That is, $fl(m) = m$ if $0 < m \leq M$ but $fl(M + 1) \neq M + 1$. Show that $M = \beta^t$ in base-$\beta$ arithmetic with precision $t$ (using pre-IEEE-standard conventions).

7. For the following numbers $x_A$ and $x_T$, how many significant digits are there in $x_A$ with respect to $x_T$?

   (a) $x_A = -0.42, \ x_T = \cos(2)$
   (b) $x_A = -0.99, \ x_T = \cos(3)$
   (c) $x_A = -0.6594503, \ x_T = \cos(4)$

8. Show that

   $\text{Rel}(\hat{x}/\hat{y}) = \frac{\text{Rel}(\hat{x}) - \text{Rel}(\hat{y})}{1 + \text{Rel}(\hat{y})} \approx \text{Rel}(\hat{x}) - \text{Rel}(\hat{y})$

   for $|\text{Rel}(\hat{x})|, |\text{Rel}(\hat{y})| \ll 1$.

9. (a) For each of the following expressions

   (ii) $x^3 - y^3$,  (b) $\sqrt{x} - \sqrt{y}$,  (iii) $\tan(x) - \tan(y)$

   find a method to avoid loss of precision when $x \approx y$.

   (b) Consider in MATLAB single precision numbers $x=single(2), y=single(1.99999)$. Evaluate the expressions in (a) using both the original definition and also your improved formulation. How many significant figures do you obtain using the original definition and the improved formulation in single precision arithmetic?

10. Consider the linear problem $Ax = b$ for $b = (1, -1)^\top$ and each of the matrices

    (i) $A = \begin{pmatrix} 1 & -1 \\ 2 & -2 \end{pmatrix}$
    (ii) $A = \begin{pmatrix} 1 & -2 \\ -1 & 2 \end{pmatrix}$
    (iii) $A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$

   (a) For which of (i)-(iii) is the problem well-posed? Explain your answers.

   (b) We shall discuss later in the course the matlab function $\text{cond}$ which calculates the condition number for the linear problem above as $\text{cond}(A)$. Calculate the condition numbers for the three problems in (a). Are the results consistent with your answers in (a)? If not, explain any discrepancies.