553.385 Numerical Linear Algebra, Fall 2021
Homework 1 Solutions

Lowen Peng
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 \( n \to \infty \).

(a) Let \( \varepsilon > 0 \). Let \( n > \log \varepsilon / \log \theta \). Then since \( \log \theta < 0 \) we have \( n \log \theta < \log \varepsilon \) or \( \theta^n < \varepsilon \). Since \( \varepsilon \) is arbitrary the definition of the limit lets us conclude that \( \theta^n \to 0 \).

(b) Observe that
\[
A^2 = \begin{pmatrix} \theta^2 & \theta \\ 0 & 0 \end{pmatrix} = \theta A, \quad A^3 = \theta A^2 = \theta^2 A.
\]
Generally we have \( A^n = \theta^{n-1} A \).

(c) By our result above we have \( A^n = \theta^{n-1} A \to 0 \).
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.

Observe that

$$f(x) = \frac{1}{1-x}, \quad f^{(1)}(x) = \frac{1}{(1-x)^2}, \quad f^{(2)}(x) = \frac{2}{(1-x)^3}, \quad f^{(3)}(x) = \frac{6}{(1-x)^4}, \quad \cdots$$

In general

$$f^{(n)}(x) = \frac{n!}{(1-x)^{n+1}},$$

and $f^{(n)}(0) = n!$. By Taylor’s theorem we have the $n$th-order approximation

$$f(x) = \frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots + x^{n-1} + \frac{x^n}{(1-\xi)^{n+1}},$$

for $|\xi| \in [0, |x|]$. 


3. Convert the following numbers to their decimal equivalent:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>(4031.23)_5</td>
<td>(b) (c73a.4fa)_16</td>
</tr>
<tr>
<td>(c)</td>
<td>(0.01)_3</td>
<td>(d) (abc123)_16</td>
</tr>
</tbody>
</table>

(a) Write \((4031.23)_5 = 4 \times 5^3 + 0 \times 5^2 + 3 \times 5^1 + 1 \times 5^0 + 2 \times 5^{-1} + 3 \times 5^{-2} = 516.32\).

(b) We have \((c73a.4fa)_16 = 51002.31103515625\).

(c) Observe that
\[
(0.01)_3 = \sum_{k=1}^{\infty} 0 \times 3^{-2k+1} + \sum_{k=1}^{\infty} 1 \times 3^{-2k} = \sum_{k=1}^{\infty} \frac{1}{9^k} = \frac{1/9}{1 - 1/9} = 0.125.
\]

(d) We have \((abc123)_16 = 11256099\).
4. Show that to convert from hexadecimal representation to binary, each hexadecimal digit \( b \) is replaced, in order, by the four binary digits \( a, a', a'', a''' \) that satisfy
\[
b = 2^3 a + 2^2 a' + 2 a'' + a'''.
\]
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 \( b' \equiv (b - a''')/2 = 2^2a + 2a' + a'' \).

The representation is justified as any hexadecimal number admits a binary representation:

\[
\sum_{k \in \mathbb{Z}} b_k 16^k = \sum_{k \in \mathbb{Z}} [2^3 a_k + 2^2 a'_k + 2 a''_k + a'''_k] 16^k
= \sum_{k \in \mathbb{Z}} [2^3 a_k + 2^2 a'_k + 2 a''_k + a'''_k] 2^{4k}
= \sum_{k \in \mathbb{Z}} [2^{4k+3} a_k + 2^{4k+2} a'_k + 2^{4k+1} a''_k + 2^{4k} a'''_k].
\]

Following the hint, note that \( a''' = (b - a''')/2 \mod 2 \), thus we set \( b'' = (b' - a''')/2 = 2a + a' \), \( a' = (b' - a'')/2 \mod 2 \), and \( b''' = (b'' - a')/2 = a \). This computes the binary digits \( a'''', a''', a', a \) for a hexadecimal digit \( b \). In a hexadecimal number we iterate across all digits and replace with the 4 associated binary digits in this way.
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:

$$x = \text{3ffbb67ae8584caa}$$

(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 $x$.

(a) We find $(3f)_{16} = (001111111111)_2$. We thus have $\sigma = 0$. For the rest $E = (011111111111)_2 = 1023$.

(b) Write $F = (0.\text{bb67ae8584caa})_{16} = 0.732050807568877$.

(c) Write $x = (-1)^\sigma 2^{E-1023}(1 + F) = 1.732050807568877$.

Note that this is a double-precision approximation to $\sqrt{3}$. 

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).

Let $m$ have base-$\beta$ representation

$$m = \sum_{i=0}^{s-1} m'_i \beta^i = \beta^s \sum_{i=1}^{s} m_i \beta^{-i} = (0.m_1 \cdots m_s)_{\beta} \beta^s, \quad m'_i = m_{s-i}.$$ 

First suppose $m < \beta^t$. Then $s = t$ and $fl(m) = (0.m_1 \cdots m_t)_{\beta} \beta^t = m$. Next let $m = \beta^t$. Then $s = t + 1$, $m_1 = 1$, and $m_2, \ldots, m_{t+1} = 0$. We have again $fl(m) = (0.m_1 \cdots m_t)_{\beta} \beta^{t+1} = m$.

Lastly let $m = \beta^t + 1 = (\beta^{-1} + \beta^{-t-1}) \beta^{t+1} = (0.m_1 \cdots m_{t+1})_{\beta} \beta^{t+1}$, with $m_1 = 1$, $m_{t+1} = 1$ and all other $m_i = 0$. Then $fl(m) = (0.m_1 \cdots m_t)_{\beta} \beta^{t+1}$ with $m_1 = 1$ and all other $m_i = 0$, so that $fl(m) = \beta^t \neq m$. 
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) \( x_A = -0.6594503, \ x_T = \cos(4) \)

(a) The relative error is \(|x_A - x_T|/x_T = |1 - x_A/x_T| = 0.009259143923400\ldots\). The largest integer \( n \) such that \(|1 - x_A/x_T| < 5 \times 10^{-n}\) is \( n = 2 \). Thus there are 2 significant digits.

(b) Write \(|1 - x_A/x_T| = 7.57924891381379\ldots \times 10^{-6}\). There are 5 significant digits.

(c) Write \(|1 - x_A/x_T| = 0.00888355512246\ldots\). There are 2 significant digits.
8. Show that

\[ \operatorname{Rel}(\hat{x}/\hat{y}) = \frac{\operatorname{Rel}(\hat{x}) - \operatorname{Rel}(\hat{y})}{1 + \operatorname{Rel}(\hat{y})} = \operatorname{Rel}(\hat{x}) - \operatorname{Rel}(\hat{y}) \]

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

Write

\[ \operatorname{Rel}(\hat{x}/\hat{y}) = \frac{\hat{x}/\hat{y} - x/y}{x/y} = \frac{\hat{x}y - xy}{x\hat{y}} - 1 = \frac{\operatorname{Rel}(\hat{x}) + 1}{\operatorname{Rel}(\hat{y}) + 1} - 1 = \frac{\operatorname{Rel}(\hat{x}) - \operatorname{Rel}(\hat{y})}{\operatorname{Rel}(\hat{y}) + 1}. \]

Note \(\operatorname{Rel}(\hat{y}) + 1 \approx 1\), which concludes our argument.
9. (a) For each of the following expressions

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

find a method to avoid loss of precision when \(x = y\).

(b) Consider in MATLAB single precision numbers \(x = \text{single}(2), y = \text{single}(1.99999)\). 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?

(a) The objective is to find decompositions to improve precision. Write: (i) \(x^3 - y^3 = (x^2 + xy + y^2)(x - y)\), (ii) \(\sqrt{x} - \sqrt{y} = (x - y)/(\sqrt{x} + \sqrt{y})\), and (iii) using the angle-sum identities \(\tan(x) - \tan(y) = \tan(x - y)[1 + \tan(x)\tan(y)]\).

(b) The following MATLAB program computes the original and improved expressions (as above), and finds relative errors:

```matlab
1  x=2; y=1.99999; d=x-y;
2  x4=single(x); y4=single(y); d4=single(d);
3  bad1=x4^3-y4^3
4  good1=(x4^2+x4*y4+y4^2)*d4
5  true1=(x^2+x*y+y^2)*d
6  relb1=double(bad1)/true1-1
7  reg1=double(good1)/true1-1
8
9  bad2=sqrt(x4)-sqrt(y4)
10  good2=d4/((sqrt(x4)+sqrt(y4))
11  true2=d/((sqrt(x)+sqrt(y))
12  relb2=double(bad2)/true2-1
13  reg2=double(good2)/true2-1
14
15  bad3=tan(x4)-tan(y4)
16  good3=tan(d4)*(1+tan(x4)*tan(y4))
17  true3=tan(d)*(1+tan(x)*tan(y))
18  relb3=double(bad3)/true3-1
19  reg3=double(good3)/true3-1
```

By inspection of the relative errors in the output, we find the respective number of significant figures to be: (i) 4 with the original, 7 with the improved; (ii) 3 with the original, 10 with the improved; and (iii) 4 with the original, 8 with the improved.
10. Consider the linear problem \( \mathbf{Ax} = \mathbf{b} \) for \( \mathbf{b} = (1, -1)^T \) and each of the matrices

\[
(i) \quad \mathbf{A} = \begin{pmatrix} 1 & -1 \\ 2 & -2 \end{pmatrix} \quad (ii) \quad \mathbf{A} = \begin{pmatrix} 1 & -2 \\ -1 & 2 \end{pmatrix} \quad (iii) \quad \mathbf{A} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}
\]

(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 \texttt{cond} which calculates the condition number for the linear problem above as \texttt{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.

(a) The \( \mathbf{A} \) in (i) is singular and no \( \mathbf{x} \) satisfies the linear problem. The \( \mathbf{A} \) in (ii) is also singular but multiple \( \mathbf{x} \) satisfy the linear problem. The \( \mathbf{A} \) in (iii) is invertible and a unique \( \mathbf{x} \) satisfies the linear problem.

(b) The following program computes the condition numbers of the respective matrices:

```matlab
A=[1 -1; 2 -2];
condA=cond(A)

B=[1 -2; -1 2];
condB=cond(B)

C=[2 -1; -1 2];
condC=cond(C)
```

The condition numbers of \( \mathbf{A} \) in (i) and (ii) are on the order of \( 10^{16} \), indicating as we expect a (nearly or exactly) singular matrix. The condition number of \( \mathbf{A} \) in (iii) is on the order of 1, which makes sense as the problem is well-posed.