Homework 1

Theory:
1. Consider the following while loop:

```plaintext
y = 1;
end
```

Using what you know about floating point arithmetic, figure out how many times the loop will iterate and what the final value of \( x \) will be. Explain your results. What changes if we initialize \( x \) and \( y \) to something else?

Solution:

Recall that in floating point arithmetic, numbers are represented in the form \( 2^p z \), with \( z \in [1, 2) \). The real numbers for which the stored exponent is \( p \) are exactly those in the interval \( [2^p, 2^{p+1}) \). If \( x = 2^{px} z_x \) and \( y = 2^{py} z_y \), then the sum \( x + y \) is stored on the machine as

\[
x + y = 2^p z, \quad \text{with} \quad p = \max\{p_x, p_y\}.
\]

Note that if \(|p_x - p_y| > m\), where \( m + 1 \) is the number of bits allocated to the mantissa, then \( x + y = x \) (if \( x > y \)) or \( x + y = y \) (if \( y > x \)).

With this in hand: on a 64-bit machine, 53 bits are allocated for the mantissa, one of them for the sign. The number \( y = 1 \) corresponds to exponent \( p_y = 0 \). If we initialize \( x = y = 1 \) and divide \( x \) by two on the \( n \)th iteration, then on this iteration

\[
x = 2^{-n}, \quad \text{(value of} \ x \ \text{on} \ n \text{th iteration}),
\]

i.e. it corresponds to exponent \( p_x = -n \). On the 53rd iteration, \(|p_y - p_x| > 52\), so \( x + y = y \) and the loop will terminate.

An identical argument can be brought to bear on any initial values for \( x \) and \( y \). In general, if \( y \) is a number with machine exponent \( p_y \) and \( x \) is a number with machine exponent \( p_x \), then we consider several cases:

- Case 1: \(|p_y - p_x| > 52\). Then the loop terminates immediately.
- Case 2: \(|p_y - p_x| \leq 52\). In this case, the loop terminates after 53 - \(|p_y - p_x| \) iterations.
2. Show that for matrices $A$ and $B$, $\|AB\|_p \leq \|A\|_p \|B\|_p$ where $1 \leq p \leq \infty$.

Solution:

This result follows from the following chain of inequalities:

\[
\|AB\|_p = \sup_{\|x\|_p = 1} \|(AB)x\|_p \\
= \sup_{\|x\|_p = 1} \|A(Bx)\|_p \\
\leq \sup_{\|y\|_p = \|B\|} \|Ay\|_p \\
\leq \|A\|_p \|B\|_p.
\]

Note that these calculations use the fact that for any vector $x$ and any matrix $A$, $\|Ax\|_p \leq \|x\|_p \|A\|_p$.

3. Bonus: Let $A \in \mathbb{R}^{n \times m}$, i.e. of the form

\[
A = \begin{pmatrix}
a_{11} & \cdots & a_{1m} \\
\vdots & \ddots & \vdots \\
a_{n1} & \cdots & a_{nm}
\end{pmatrix},
\]

be a matrix with real valued entries. Prove that

\[
\|A\|_1 = \max_{1 \leq j \leq m} \sum_{i=1}^n |a_{ij}|.
\]

Solution:

The proof will make use of the following lemma, whose truth is transparent:

Lemma 1. Suppose $w_1, \cdots, w_m$ are non-negative that sum to 1. Then for any numbers $a_1, \cdots, a_m$,

\[
\sum_{i=1}^m a_i w_i \leq \max_i |a_i|.
\]

Proof of problem:

Let $x$ satisfy $|x_1| + \cdots + |x_n| = 1$, i.e. $x$ is an $n$–vector of 1-norm equal to 1. Then

\[
\|Ax\|_1 = \max_{1 \leq j \leq m} \left| \sum_{i=1}^n a_{ij} x_j \right| \\
\leq \sum_{i=1}^m \sum_{j=1}^n |a_{ij} x_j| \\
= \sum_{j=1}^n |x_j| \left( \sum_{i=1}^m |a_{ij}| \right).
\]
It follows from the Lemma that
\[ \|Ax\|_1 \leq \max_j \sum_{i=1}^m |a_{ij}|. \]

On the other hand, if \( j_0 \) is the index for which this maximum is attained, then by setting \( x_{j_0} = 1 \) and all other \( x_j = 0 \), we have
\[ \|Ax\|_1 = \|A(:, j_0)\| = \max_j \sum_{i=1}^m |a_{ij}|, \]

i.e. we can actually attain equality. Together, these results prove the problem. \( \square \)
Computation:

1. Recall that you can approximate the function \( \sin x \) via its Taylor series:

\[
\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots
\]

The following Matlab function uses the series to compute \( \sin x \):

```matlab
function sum = taylorsin(x)
sum = 0;
term = x;
n = 1;
while sum + term ~= sum
    sum = sum + term;
    term = -x.^2.*term ./((n+1)*(n+2));
    n = n+2;
end
```

Answer the following questions for \( x = \pi/2, 11\pi/2, 21\pi/2, \) and \( 31\pi/2 \):

(a) How accurate is the computed result?
(b) How many terms are required?
(c) What is the largest term in the series?

Can you draw any conclusions about the use of floating-point arithmetic and power series to evaluate functions?

**Solution:**

I modified the function `taylorsin.m` to provide the information asked for by the problem. My modified function looked like:

```matlab
function [sum, n, maxterm, n_maxterm] = taylorsin(x)
% function [sum, n, maxterm, nmaxterm] = taylorsin(x)
% Description: Calculates the sin of x via a Taylor series
% Inputs:
%  x   a scalar
% Outputs:
%  sum  partial sum
%  n   number of terms that went into that sum
%  maxterm  the largest term
%  n_maxterm  the index of the largest term

sum = 0;
term = x;
n = 1;
maxterm = abs(x);
n_maxterm = n;
while sum + term ~= sum
    sum = sum + term;
end
```

4
term = -x.^2.*term ./((n+1)*(n+2));
n = n+2;
if abs(term) > maxterm
    maxterm = abs(term);
    n_maxterm = n
end
end
end

I then ran this function with a script, called hw1_comp1.m. My calling script looked like this:

```matlab
% script: hw1_comp1.m
% Description: a small script for solving HW1, computation problem 1

x = [pi/2, 11*pi/2, 21*pi/2, 31*pi/2]; % defining a vector containing inputs
data = zeros(length(x),4); % define data storage
for i = 1:length(x)
    [sum n maxterm n_maxterm]=taylorsin(x(i))
    data(i,1) = sum; % value of sum
    data(i,2) = n; % number of terms
    data(i,3) = maxterm; % the largest term
    data(i,4) = n_maxterm; % the index of the largest term
end %for
accuracies = abs(data(:,1) - 1);
```

The information can be displayed in a table:

<table>
<thead>
<tr>
<th>value of $x$</th>
<th>sum</th>
<th>number of terms</th>
<th>maximum term</th>
<th>index of maximum term</th>
<th>error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi/2$</td>
<td>1.0000</td>
<td>23</td>
<td>1.57</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>$11\pi/2$</td>
<td>-1.0000</td>
<td>75</td>
<td>3.0665e + 06</td>
<td>17</td>
<td>2.0000</td>
</tr>
<tr>
<td>$21\pi/2$</td>
<td>0.9996</td>
<td>121</td>
<td>1.4673e + 13</td>
<td>33</td>
<td>0.0004</td>
</tr>
<tr>
<td>$31\pi/2$</td>
<td>833.9055</td>
<td>159</td>
<td>7.9890e + 19</td>
<td>49</td>
<td>832.9055</td>
</tr>
</tbody>
</table>

The conclusions I draw from this experiment are that it is a bad idea to use Fourier series to calculate the sin function.