Homework 1
Math 146 Spring 2023
Due Monday, February 6, 11:59 PM

1. (10 points) For $\mathbf{x}$, an arbitrary $n \times 1$ column vector, and an $n \times n$ matrix $A = [A_{ij}]$, $\mathbf{x}^T A\mathbf{x}$ is a scalar, which is called a *quadratic form*. Express $\mathbf{x}^T A\mathbf{x}$ in terms of the components of A and $\mathbf{x}$. (Hint: $\mathbf{x}^T A\mathbf{x} = \mathbf{x} \cdot (A\mathbf{x})$)

2. (16 points) If A is a symmetric $n \times n$ matrix and B is $n \times m$, prove that $B^T AB$ is a symmetric matrix.

3. (16 points) Prove that $(AB)^T = B^T A^T$ for $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$.

4. (18 points) Assume we have a floating point number system, represented in binary, which allocates 1 bit to sign, 4 bits to the mantissa (q) and 4 bits to the characteristic(c), and our numbers are of the form $\pm(1+q)(2^{c-7})$
   (a) What is machine epsilon?
   (b) Find the floating point representation of $\frac{1}{7}$. Also include the binary representation for q and c.
   (c) Find the floating point representation of π. Also include the binary representation for q and c.
   (d) Comment on the relative errors of your representations from (b) and (c)

5. (18 points) Near certain values of x, each of the following functions cannot be accurately computed using the formula as given due to cancellation error. Identify the values of x and propose a reformulation to remedy the problem.
   (a) $f(x) = 1 + \cos x$
   (b) $f(x) = 1 - 2\sin^2 x$
   (c) $f(x) = \ln(x) - 1$

6. (18 points)
   (a) For which positive integer(s) α can the number $5 + 2^{-\alpha}$ be represented *exactly* in double precision floating point arithmetic?
   (b) Find the largest integer α for which $fl(19 + 2^{-\alpha}) > fl(19)$ in double precision floating point arithmetic.

7. (24 points)

(a) What is the relative condition number of evaluation of the function $f(x) = e^{\cos x}$ at the point x? About how many digits of accuracy would you expect to lose when performing this operation at $x = 1000$?

(b) Suppose that f and g are continuously differentiable functions. Let $\kappa_f(x)$ denote the relative condition number of evaulation of the function f at x and similarly for $\kappa_g(x)$ Find a relationship between the condition number of evaluation of $h(x) = f(x)g(x)$ and κ_f and κ_g. Comment on why this might be useful.

(c) Let $\kappa_f(x)$ denote the relative conditon number of evaluation of the function f at the point x. Find a function f which is infinitely differentiable on the interval (0,1) (but which may have singularities at $x = 0$) such that

$$\lim_{x \to 0^+} \kappa_f(x) = \infty.$$

8. (18 points) You are running a simulation that updates time every 0.1 seconds. The time in the simulation is kept by incrementing a time variable. See the below code.

```
dt = 0.1;          % time step
t  = 0;            % initialize time
Nsteps = 864000;   % number of steps to take

% loop in time
%
for j=1:Nsteps

   %
   % SOME SIMULATION
   %

   % update time
   %
   t = t + dt;
end
```

(a) Suppose you are simulating one day (86,400 s). Implement the above code, and compute the absolute and relative errors in the time at the end of the simulation.

(b) Change the time increment to 0.125 and again run the simulation to 1 day. What are the relative and absolute error in the time?

(c) Explain the difference in the results from parts (a) and (b).

9. (20 points) Write a Matlab function to perform matrix multiplication of two matrices. Your code should take matrices of any size and return an error if they cannot be multiplied. Use `tic` and `toc` to time your function and the built-in Matlab function to multiply A, a 1000×40

matrix with B, a 40×4000 matrix of randomly chosen numbers from 0 to 30. Report these times, as well as the flop count of your function for A and B, both $n \times n$ matrices.

10. (22 points) Suppose that you can only use addition, subtraction, multiplication, division, rounding and integer powers of numbers. You decide to use a Taylor Series to evaluate $y = e^x$ with only these operations, since you learned in calculus that it converges for all x. Below gives an example of such code.

    (a) Explain what the while loop is doing in this code.

    (b) Assess the accuracy of the algorithm below by using it to approximate $y = e^x$ on the interval $x \in [-20, 20]$ by comparing with the built-in library function for the exponential. Compute the absolute and relative errors as a function of x and plot the results (use log scale for the error; i.e. in MATLAB use command semilogy for plotting).

    (c) For what values of x do you see poor performance from the algorithm? Explain the reason for the poor performance.

    (d) Based on your answer from the previous part, modify the algorithm to eliminate the poor performance. Discuss the changes and demonstrate the performance of the modified code by plotting the errors as a function of x.

```
% myexp.m -- function for computing y=exp(x) using a Taylor series
%
function [y,Nterms]=myexp(x);
  oldsum = 0;
  newsum = 1;
  term   = 1;

  n      = 0;
  while newsum~=oldsum
    n = n+1;
    term = term*x/n;
    oldsum = newsum;
    newsum = newsum + term;
  end

  Nterms = n + 1;
  y = newsum;
```

11. Suppose that you can only use addition, subtraction, multiplication, division, rounding and integer powers of numbers. You decide to use a *truncated* Taylor Series to evaluate $y = e^x$ with only these operations.

    (a) (20 points) Create a function that approximates e^x by truncating the series to n terms. Use your function for $n = 12$ to approximate e^x for some values of x between -100 and 100. Repeat for $n = 50$. Comment on your results.

(b) (30 BONUS points) By exploiting properties of the exponential, design an algorithm for accurately computing the value of e^x using your truncated series for x values between -100 and 100. Explain your algorithm and implement it. Report your relative error for e^x for $x = \pm 0.5$ and $x = \pm 100$ for $N = 50$, and compare to the previous part. You should be able to acheive relative errors below 10^{-12}. Hint: Based on the Taylor Series remainder, for what values of x do you expect the series to be the most accurate? Exploit properties of the exponential to make an algorithm that only uses the series on this range of x values.