Stat-6207 **Homework Set 2** **Fall 2018 (Modarres)**

1) Consider a floating point representation that uses a 32-bit word, with one bit for the sign, 7 bits for the exponent (e) in excess-64 notation, and 24 bits for the mantissa, $x = \pm(\sum_{i=1}^{6} \frac{d_i}{16^i})16^{e-64}$ where $0 \leq d_i \leq 15$. Carry out the following floating-point operations by hand. Show the intermediate and final values of all variables in hexadecimal.

$A = -462.3$

$B = 2.0/16.0$

$C = 5.0/16.0^7 + 1.0/16.0^8$

$D = B - C$

$E = 2.0/16.0 + 11.0/16.0^4$

$F = 7.0/16.0^2 + 12.0/16.0^5$

$G = E * F$

$R = 5.0/16.0 + 3.0/16.0^5$

$P = 5.0/16.0$

$Q = R/P$

2) Perform the following calculations: a) $\frac{4}{5} + \frac{1}{3}$, b) $\frac{4}{5} \times \frac{1}{3}$ c) $(\frac{1}{3} - \frac{3}{11}) + \frac{3}{20}$ d) $(\frac{1}{3} + \frac{3}{11}) - \frac{3}{20}$

- Exactly,
- Using 3-digit arithmetic with chopping of the exact final value,
- Using 3-digit arithmetic with rounding of the exact final value,
- Compute the relative errors.

3) Let $f(x) = x\exp(x^2)$.

- a) Find the fourth Taylor polynomial $P_4(x)$ for $f(x)$ near $x_0 = 0$.
- b) Find an upper bound for $|f(x) - P_4(x)|$ for $x \in [0, 0.4]$.
- c) Approximate $\int_0^{0.4} f(x)dx$ using $\int_0^{0.4} P_4(x)dx$.
- d) Find an upper bound for in the computation of part a) by using your answer to part b).
- e) Approximate $f'(0.2)$ by computing $P_4'(0.2)$. Use the correct answer to $f'(0.2)$ to 5 decimal places to compute the relative error in your computation.

4) Let $X = (X_1, \ldots, X_d)$ be a random vector with i.i.d. components and let $\|X\|_p = (\sum_{i=1}^{d} |X_i|^p)^{1/p}$ be its p-norm, for $p > 0$. The behavior of p-norm as d tends to infinity is of interest in many high dimensional applications, including analysis of nearest neighbor search algorithms, data mining, pattern recognition, and fraud detection. Let $p = 2$ and $\|X\|_2 = \sqrt{\sum_{i=1}^{d} {X_i}^2}$ be its Euclidean norm. Using the following sampling experiment show that the coefficient of variation $CV(\|X\|_2) = \sqrt{\mathrm{Var}(\|X\|_2)/E(\|X\|_2)}$ tends to zero as d tends to infinity. Generate $n = 10$ vectors from a d-variate (i.i.d. components) normal distribution and report estimates of $E(\|X\|_2)$, $\mathrm{Var}(\|X\|_2))$ and $CV(\|X\|_2)$. Obtain 10 replications of the experiment for each $d \in \{50, 500, 5000, 50000, 500000\}$.

5) Let n and m be positive integers and consider the binary floating point system F of the form $\pm 1.a_1 a_2 \ldots a_n 2^e$ where $e \in \{0, -1, -2, \ldots, -m\}$ and $a_k \in \{0.1\}$ for all k.

- a) In terms of n and m, how many numbers are in this system?
- b) Let $m = 2$ and $n = 5$. What is the machine epsilon if we approximate all numbers by rounding to the closest number in F?
- c) Let $m = 2$. Find the smallest value of n that will ensure that some number in F is greater that 1.8.

6) Consider the finite Taylor series for computing $e^x = \sum_{i=0}^{\infty} x^i/i!$. Write a function to compute e^x and evaluate it at $x = \pm 1, \pm 10, \pm 20$, and ± 30. Compare your computing values to the built-in function exp(x). If the values do not agree, explain the cause and suggest a remedy.