

This assignment is due at the **start** of your lecture/tutorial on Friday, 7 October 2011.

1. [6 marks]

What are the absolute and relative errors in approximating π by the following values?

- (a) 3.1
- (b) 3.1415
- (c) 3.1415926535

For the purposes of this question, you can assume that the “true” value of π is 3.141592653589793.

2. [10 marks]

In a floating-point number system with parameters $\beta = 10$, $p = 3$, $L = -20$ and $U = +20$ that uses the *round-to-nearest* rounding rule and allows gradual underflow with subnormal (i.e., denormalized) numbers, what is the result of each of the following floating-point arithmetic operations?

- (a) $1.00 \cdot 10^1 + 3.00 \cdot 10^{-1}$
- (b) $1.00 \cdot 10^3 - 3.00 \cdot 10^1$
- (c) $1.00 \cdot 10^2 + 3.00 \cdot 10^{-3}$
- (d) $2.00 \cdot 10^4 - 6.00 \cdot 10^6$
- (e) $3.00 \cdot 10^{10} - 6.00 \cdot 10^3$
- (f) $1.56 \cdot 10^2 + 4.64 \cdot 10^3$
- (g) $2.45 \cdot 10^{10} \times 6.53 \cdot 10^{-15}$
- (h) $-3.21 \cdot 10^{10} / (1.75 \cdot 10^{-15})$
- (i) $3.24 \cdot 10^{-10} \times 1.56 \cdot 10^{-12}$
- (j) $2.54 \cdot 10^{-10} \times 6.73 \cdot 10^{-15}$

Write each answer as a normalized 3-digit decimal floating-point number if possible. If that is not possible, write your answer as a subnormal (i.e., denormalized) 3-digit decimal floating-point number if that is the most accurate representation. If that is not possible either, then write your answer as +Inf, -Inf or NAN, whichever best represents your answer.

[Note: your textbook uses L to denote the smallest possible exponent in the floating-point number system and U to denote the largest possible exponent in the floating-point number system. In class, I used n_{\min} for L and n_{\max} for U .

3. [5 marks]

Consider the function $f(x) = x^{1/5}$ for real positive x (i.e., $x > 0$).

Is this function well-conditioned or ill-conditioned in a relative sense with respect to small relative changes in the value of the input argument x ?

Justify your answer.

4. [10 marks: 5 marks for each part]

Assume throughout this question that there are no overflows or underflows.

Which of the two mathematically equivalent expressions

$$\frac{1}{1-x} - \frac{1}{1+x} \quad \text{or} \quad \frac{2x}{(1-x)(1+x)}$$

can be evaluated more accurately in floating-point arithmetic?

- (a) Show that one of the expressions always gives a very accurate answer in a relative error sense.
- (b) Give an example which illustrates that the other expression can give an answer that is much less accurate in a relative error sense.

5. [20 marks: 5 marks for each part]

In your calculus course, you may have learnt that

$$e = \lim_{n \rightarrow \infty} (1 + 1/n)^n \tag{1}$$

where $e \approx 2.718281828459$ is the base of the natural logarithms.

- (a) Write a MatLab program to approximate e by computing $(1 + 1/n)^n$ for $n = 10^k$ and $k = 1, 2, 3, \dots, 20$.

You can compute a very accurate approximation to e in MatLab by evaluating $\exp(1)$ and use this value to approximate the absolute error in $(1 + 1/n)^n$. For each $k = 1, 2, 3, \dots, 20$ your program should print both your approximation to e computed from $(1 + 1/n)^n$ and an approximation to the associated absolute error using $(1 + 1/n)^n - \exp(1)$.

Hand in your MatLab program and its output.

In exact arithmetic, the magnitude of the absolute error in your approximation should decrease as k increases. However, you should see that the magnitude of the absolute error in your computed approximation decreases as k increases for $k = 1, 2, 3, \dots, 8$, but, for $k > 8$, the magnitude of the absolute error does not decrease. When I performed this computation, I found that the magnitude of the absolute error was roughly constant for $k = 9, 10, 11$; the magnitude of the absolute error increased for $k = 11, 12, \dots, 16$; and the magnitude of the absolute error remained constant at about -1.7183 for $k = 16, 17, \dots, 20$.

- (b) Why does the magnitude of the absolute error remain roughly constant at about -1.7183 for $k = 16, 17, \dots, 20$?
- (c) Why does the magnitude of the absolute error reach a minimum at about $k = 8$?
- (d) Explain why the absolute error is approximately proportional to $1/n$ for $n = 10^k$ and $k = 1, 2, 3, \dots, 8$.