

This assignment is due at the **start** of your tutorial on 2 October 2009.

1. [6 marks]

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

- (a) 3.0
- (b) 3.141
- (c) 3.14159266745

For the purposes of this question, you can assume that the “true” value of π is 3.14159265358979.

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 numbers, what is the result of each of the following floating-point arithmetic operations?

- (a) $1 + 10^{-2}$
- (b) $1 - 10^{-2}$
- (c) $1 + 10^{-5}$
- (d) $10^4 - 10^6$
- (e) $10^{12} - 10^5$
- (f) $1.26 \cdot 10^1 + 2.56 \cdot 10^2$
- (g) $1.25 \cdot 10^{10} \times 4.23 \cdot 10^{-15}$
- (h) $10^{10}/10^{-15}$
- (i) $1.25 \cdot 10^{-10} \times 4.23 \cdot 10^{-12}$
- (j) $1.25 \cdot 10^{-10} \times 4.23 \cdot 10^{-15}$

Write each answer as a normalized 3-digit floating-point number if possible. If that is not possible, write your answer as a subnormal 3-digit 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.

3. [10 marks: 5 marks for each part]

Consider the function $f(x) = \log_e(x)$ for real positive x .

- (a) 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 for x close to 1?
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 for x close to 10?
Justify your answer for each case.

(b) Write a little MatLab program to verify your predictions from part (a).

Hand in your MatLab program and its output.

Also include a brief explanation of why you believe your computational results from part

(b) support your theoretical predictions from part (a).

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

$$(x - y)(x + y) \quad \text{or} \quad x^2 - y^2$$

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

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

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 error in $(1 + 1/n)^n$. For each k , your program should print both your approximation to e computed from $(1 + 1/n)^n$ and an approximation to the associated error using $\exp(1) - (1 + 1/n)^n$.

Hand in your MatLab program and its output.

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

(b) Why does the error remain roughly constant at about 1.7183 for $k = 16, 17, \dots, 20$?

(c) Why does the magnitude of the error reach a minimum at about $k = 8$?

(d) Explain why the error is approximately proportional to $1/n$ for $n = 10^k$ and $k = 1, 2, 3, \dots, 8$.