Task 2. Determination pf macine precision, number of bits in mantissa and the range of floating point numbers

e-mail: andrzej.kedziorski@fizyka.umk.pl
pokój: 485B
http://www.fizyka.umk.pl/˜tecumseh/EDU/MNII/

Task 2

Write a program that determines the machine precision (unit roundoff) as well as the number of mantissa bits for a floating point number in single and double precision in the IEEE 754 standard. In addition, the program is to calculate the possible range of exponents for single and double precision numbers. What is the effect of applying a gradual underflow on the results?

Floating-point number $x = \pm m\beta^e$

- β base (radix), $\beta = 2$
- m - mantissa of length t (number of bits in mantissa)

  $$m = d_0 + \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \ldots + \frac{d_{t-1}}{\beta^{t-1}},$$

  where $0 \leqslant d_i \leqslant \beta - 1$, $i = 0, \ldots, t - 1$
- e exponent, $L \leqslant e \leqslant U$
- Normalization: $1 \leqslant m \leqslant \beta$

Definitions of machine precision (unit roundoff) ϵ_{mach}

- Rounding toward 0 (truncation) $\epsilon_{\mathrm{mach}} = \beta^{1-t}$
- Rounding to nearest $\epsilon_{\mathrm{mach}} = \frac{1}{2}\beta^{1-t}$
- The smallest real number ϵ that satisfies *fl*$(1+\epsilon) > 1$, where *fl*$(\ldots)$ means calculation within floating-point arithmetic with finite precision
- Evaluate ϵ_{mach} using the last "definition"; as a byproduct, we can find the number of bits in the mantissa t
- With the aid of t, we can compare the evaluated ϵ_{mach} with the above definitions (which definitions are compatible with each other?)
- Perform the above calculations in single and double precision (make sure that single-precision calculations were made)
- For example, you may check the value of the expression $|3(4/3-1)-1|$ that also gives the value of ϵ_{mach}

Unit roundoff ϵ_{mach} - output

- On the output, the program prints the iteration number in columns, ϵ and $(1 + \epsilon)$, where ϵ represents successive approximations to ϵ_{mach}
- At the end the program prints the results: t and the values of ϵ_{mach} calculated in different ways
- Compare the results with the IEEE 754 standard

A range of floating point numbers

1. Find the smallest positive floating point number
2. Find the smallest positive floating point number with a normalized mantissa
3. For the above task, the unit roundoff ϵ can be used, defined as the smallest number such that $fl(1+\epsilon) > 1$
4. Find the largest possible floating point number (less than ∞)
5. Perform the above calculations in single and double precision (make sure that single-precision calculations were made)
6. On the output, the program prints (three) extreme values determined for single and double precision, respectively.
7. Compare the results to the extreme values of exponent e $(L \leqslant e \leqslant U)$ and to the underflow and overflow levels defined in the IEEE 754 standard