

In-class Exercises and Homework, Sept. 18

Due Tuesday, Sept. 25 on plain old paper.

The *unit in the last place* (ulp) is a convenient way to quantify the absolute error in terms of the available floating-point precision. Consider 6-digit decimal arithmetic: If $\hat{x} = 1.23458$ is an approximation to the “true” value $x = 1.23456$, then the error in \hat{x} is 2 ulp. Similarly, if $x = 123.456$ and $\hat{x} = 123.458$, then the error is still 2 ulp, because $x = 1.23456 \times 10^2$ and $\hat{x} = 1.23458 \times 10^2$; the error is two units in the last place in the significand of the normalized 6-digit base-10 floating-point representation of x .

1. Euler’s constant is $\gamma = 0.5772156649\dots$. What is the floating-point representation of γ as a normalized base-10 floating-point number with 6 digits of precision if the rounding mode is
 - (a) round to nearest?
 - (b) round to zero?
 - (c) round to ∞ ?
 - (d) round to $-\infty$?
2. Compute the approximate absolute and relative error in each of the floating-point approximations in Exercise 1.
3. Express the approximate absolute error in each of the floating-point approximations in Exercise 1 in terms of units in the last place (ulp).
4. The IEEE 754 floating-point standard, used in nearly all general-purpose computers nowadays, provides floating-point numbers in base-2 format. IEEE single-precision numbers have a 24-bit significand and an exponent between -126 and $+127$. That is, they have the form

$$a_0.a_1a_2\cdots a_{23} \times 2^e$$

where $-126 \leq e \leq 127$. A nonzero floating-point number is *normalized* if $a_0 = 1$ and e is in the allowed range.

- (a) What is the normalized single-precision representation of the number 1?
- (b) If we change the significand by 1 ulp, by how much does the value of the floating-point representation of 1 change?
- (c) What is the single-precision floating-point representation of 2.0? 0.5? 1.125?
- (d) By how much does the value of each floating-point representation change in (c) if it is altered by 1 ulp?

5. What is the absolute error in the IEEE single-precision representation of $\frac{1}{10}$ when rounding to nearest?
6. Suppose that your compiler's documentation states that the computed value of $\sin x$ is accurate to 1 ulp if $|x| \leq \pi/4$. What is the maximum absolute error in the computed IEEE single-precision value of the sine function if $|x| \leq \pi/4$? (Hint: $|\sin x| \leq 1/\sqrt{2}$ when $|x| \leq \pi/4$, and 1 ulp is largest when the corresponding result is largest in absolute value.)
7. Consider the IEEE 754 single-precision format.
 - (a) To what value does the smallest positive normalized floating-point number correspond? Give an "exact" answer in terms of powers of 2 as well as a decimal approximation.
 - (b) To what value does the largest positive normalized floating-point number correspond? Give an "exact" expression and a decimal approximation.
8. Answer the same questions as Exercise 7 for the IEEE 754 double-precision format, which is $a_0.a_1a_2\cdots a_{52} \times 2^e$ where $-1022 \leq e \leq 1023$.
9. (a) Prove Theorem 2 in the notes. (b) Re-state Theorem 2 in terms of ulp.
10. Find the binary representation of $1/3$. Then determine the representation of $1/3$ in IEEE single precision and the corresponding absolute and relative errors in the result.