Exercise 3.1. (Baby floating-point calculations.)

(a) How would you express the fraction (in decimal) 1/8 in the form \( b_1.b_2b_3 \cdots \times 2^p \) in base 2, where each \( b_i \) is a bit and \( b_1 \neq 0 \).

(b) The decimal value 0.1 (one-tenth) can be written as the infinite sum

\[
0.1 = \frac{1}{10} = \frac{1}{16} + \frac{9}{(16)^2} + \frac{9}{(16)^3} + \cdots = \frac{1}{16} + \frac{9}{(16)^2} \sum_{k=2}^{\infty} \frac{1}{(16)^k}.
\]

It follows that there is a repeating normalized infinite representation of 0.1 in base 16 with the form

\[
0.1 = (1.\{h_1h_2h_3\})_{16} \times 16^p,
\]

where each \( h_i \) is a hexadecimal digit. Give the value of \( p \) (in decimal) and the values (in hexadecimal) of \( \{h_i\} \).

(c) Suppose that we wish to use the expression (3.1) of 0.1 from part (b) to produce \( t \), a correctly rounded version of 0.1 that is representable in IEEE double precision. This corresponds to defining \( t \) in the form

\[
t = (1.h_1h_2\ldots h_{13})_{16} \times 16^p.
\]

(Note that the “1” appearing before the hexadecimal point would be the hidden bit.) Please write \( t \) as in (3.2), giving the exponent \( p \) and the value of each \( h_i \) (in hexadecimal), and comment on how you determined the value of \( h_{13} \).

Exercise 3.2. Given a real nonzero number \( x \), by convention the value \( y = x^0 \) (“\( x \) raised to the zero power”) is equal to one. Using IEEE double-precision arithmetic,

(a) Compute \( y = x^0 \) when (a) \( x = 0 \), (b) \( x = \text{Inf} \), and (c) \( x = \text{NaN} \), printing \( x \) and \( y \) in two formats in each case: (i) \texttt{long e} and (ii) hexadecimal. Explain whether the results seem sensible (or not) to you.

(b) Do the same as in (a) for (i) \( 1.0 \)\( (\text{Inf}) \), (ii) \( -1.0 \)\( (\text{Inf}) \), (iii) \( \log(0.0) \), (iv) \( \log(-\text{Inf}) \), and (v) \( \exp(-\text{Inf}) \). Explain whether the results seem sensible (or not) to you.

(c) Devise your own example of a “non-standard” calculation. Determine what happens when it is performed in IEEE double-precision arithmetic, and comment on the results.

Exercise 3.3. Here are the hexadecimal representations of the correctly rounded representable versions, in IEEE double-precision format, of three familiar real numbers. What are those three numbers (in decimal)? Explain how you derived each answer, showing your work, and check your answers by setting \texttt{format hex} and printing the numbers.

c059000000000000
3f847ae147ae147b
400921fb544442d18
Exercise 3.4. The Taylor series expansion of a smooth function \( f \) around the point \( \bar{x} \) is
\[
f(\bar{x} + h) = f(\bar{x}) + hf'(\bar{x}) + \frac{1}{2}h^2 f''(\xi),
\]
where \( \xi \) lies between \( \bar{x} \) and \( \bar{x} + h \). We define the forward-difference approximation \( \phi(\bar{x}, h) \) of the derivative \( f' \) evaluated at \( \bar{x} \) as
\[
\phi(\bar{x}, h) = \frac{f(\bar{x} + h) - f(\bar{x})}{h}.
\]
It follows from the Taylor series that the error in \( \phi(\bar{x}, h) \) satisfies the exact relation
\[
\phi(\bar{x}, h) - f'(\bar{x}) = \frac{1}{2}hf''(\xi) = e_r,
\]
where \( e_r \) is called the truncation error because \( \phi(\bar{x}, h) \) arises from truncating the Taylor series, and \( \xi \) is unknown.

Consider the function \( f(x) = \sin(x) \), where \( x \) is in radians.

(a) Give an upper bound on the truncation error \( |e_r| \) of (3.4) that is valid for all \( \bar{x} \), expressed in terms of the finite-difference interval \( h \). What does this bound suggest about how to choose \( h \) so that the error in a forward-difference approximation to \( f'(\bar{x}) \) is as small as possible?

(b) Give the mathematical form of \( f'(x) \). At \( \bar{x} = 2.25 \), print the numerical values of \( f(\bar{x}) \) and \( f'(\bar{x}) \), computed with IEEE double-precision arithmetic.

(c) For \( k = 4, \ldots, 15 \), define \( h_k = 10^{-k} \). Let \( \tilde{\phi}_k \) denote the computed version of \( \phi(\bar{x}, h_k) \), obtained from the formula (3.3) with \( \bar{x} = 2.25 \) and \( h = h_k \).

(i) For each \( k \), print, using a standard scientific decimal format, \( k, h_k, \tilde{\phi}_k, \) and the difference \( \tilde{\phi}_k - f'(\bar{x}) \).

(ii) Then print in hexadecimal (1) the exact value of \( f(\bar{x}) \), and (also in hexadecimal) for \( k = 4, \ldots, 15 \), the quantities (2) \( h_k \), (3) the computed difference \( f(\bar{x} + h_k) - f(\bar{x}) \), and (4) the computed difference \( \tilde{\phi}_k - f'(\bar{x}) \).

(d) Referring explicitly to the hexadecimal numbers in (c)(ii), explain why the error in \( \tilde{\phi}_k \) does not decrease as \( h_k \) becomes smaller. What source of error, in addition to truncation error, affects the computed values \( \{\tilde{\phi}_k\} \)? Specifically discuss how the phenomenon of cancellation is revealed in the computed results of (c)(ii).

Exercise 3.5.

(a) Let
\[
A = \begin{pmatrix} 0.550 & 0.423 \\ 0.473 & 0.364 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 0.8757 \\ 0.7533 \end{pmatrix}.
\]
Show that the exact solution \( x^* \) to \( Ax = b \) is \( x^* = (0.9, 0.9)^T \). Give the computed solution \( \tilde{x} \) obtained by executing the Matlab command \( A\backslash b \) (or the equivalent in octave or SciPy). Compute \( d = \tilde{x} - x^*, r^* = b - Ax^*, \) and \( \tilde{r} = b - A\tilde{x} \), and comment on the relative size of their norms.

(b) Let \( \tilde{x} \) be a potential solution of \( Ax = b \), with residual \( \tilde{r} = b - A\tilde{x} \), and define \( E \) as the rank-one matrix
\[
E = \frac{1}{\tilde{x}^T\tilde{x}} \tilde{r}\tilde{x}^T.
\]
Show mathematically that the exact matrix \( E \) satisfies \( (A + E)\tilde{x} = b \).

(c) Consider the vector \( \hat{x} \)
\[
\hat{x} = \begin{pmatrix} 40.9 \\ -51.1 \end{pmatrix}.
\]
Would you say that \( \hat{x} \) is close to either \( x^* \) or \( \tilde{x} \)? Explain.
(d) Given the vector $\tilde{x}$ from (c), compute $\tilde{r} = b - A\tilde{x}$, the matrix $E$ in (3.5), and $\|E\|_2$. Is $\|E\|_2$ “small” or “large”?

(e) Compute and print the solution $\bar{x}$ to $(A + E)\bar{x} = b$, obtained using the “\” command.

(f) Based on $\|\bar{x} - \tilde{x}\|$, is $\bar{x}$ “close to” $\tilde{x}$? Explain why or why not.

(g) Justify the statement: $\tilde{x}$ is close to being the exact solution of a system that is close to the original system.